

TEL AVIV UNIVERSITY

The Iby and Aladar Fleischman Faculty of Engineering
The Zandman-Slaner School of Graduate Studies

**Dithered Quantization via
Orthogonal Transformations and a
Cauchy-Schwarz like Inequality**

A thesis submitted toward the degree of
Master of Science in Electrical and Electronic Engineering

by

Ran Hadad

January, 2016

TEL AVIV UNIVERSITY

The Iby and Aladar Fleischman Faculty of Engineering
The Zandman-Slaner School of Graduate Studies

Dithered Quantization via Orthogonal Transformations and a Cauchy-Schwarz like Inequality

A thesis submitted toward the degree of
Master of Science in Electrical and Electronic Engineering

by

Ran Hadad

This research was carried out in
The School of Electrical Engineering
The Department of Electrical Engineering - Systems
Under the supervision of
Prof. Uri Erez

January, 2016

Acknowledgements

First, I would like to express my deep gratitude to my advisor, Prof. Uri Erez, for his abundant help and superb advising.

I would also like to thank my colleagues from the communication laboratories in the Department of Electrical Engineering at Tel Aviv University. Special thanks go to Or Ordentlich, Anatoly Khina and Elad Domanovitz, for the helpful discussions and great advice throughout this period.

Finally warm thanks to my parents Bruria and Zion, to my brothers Liad and Nadav, and especially to my wife Irina for her love and support.

Abstract

This thesis consists of two parts. In the first part an inequality is derived for the correlation of two univariate functions operating on symmetric bivariate normal random variables. The inequality though being a simple consequence of the Cauchy-Schwarz inequality, has some interesting applications beyond its use in the latter part of the thesis. We briefly explore some of these.

In the second part of the thesis, we develop a novel dithered quantization scheme based on the application of orthogonal transformations. Dithered quantization is a technique used to reduce or eliminate the statistical dependence between the signal and quantization error. This is most often achieved via adding pseudo-random noise prior to quantization. The present work develops a different dithering method, where dithering is accomplished by applying an orthogonal transformation to a vector of samples prior to quantization, and applying its inverse to the output of the quantizer. Focusing on uniform scalar quantization, it is shown that for any quantization rate, the proposed architecture approaches second-order independence, i.e., vanishing correlation, as the dimension of the vector of samples jointly processed grows.

Contents

1	An Inequality for the Correlation of Two Functions Operating on Symmetric Bivariate Normal Variables	1
1.1	Introduction	1
1.2	Statement of Result and Proof	2
1.3	Generalized Form of Inequality	4
1.4	Application: A Criterion for Identification of a Memoryless Non-Linearity .	5
2	Dithered Quantization via Orthogonal Transformations	7
2.1	Introduction	7
2.2	Notation	11
2.3	Background	12
2.3.1	Subtractive Dithered Quantization	12
2.3.2	Multiple Descriptions via Dithered Quantization	15
2.3.3	Orthogonal Matrix Preliminaries	19
2.4	Proposed Architecture and Basic Properties	22
2.4.1	Optimization Goal	25
2.5	Performance Analysis	26
2.5.1	Bound on Performance for the Haar and Randomized Hadamard Ensembles	30
2.5.2	Performance Analysis for Symmetric 1-bit Quantization	32
2.5.3	Conjectured Tighter Bound on the Error Correlation Coefficient . .	33
2.6	Numerical Design of Orthogonal Transformations	34
2.6.1	Numerical Optimization Procedure	35
3	Summary	41
A	Optimal Scalar Uniform Quantizer for a Gaussian Source	42
A.1	Undithered Quantizer	42

A.2 Subtractive Dithered Quantizer	44
B Output Correlation Function Expansion	46
C Properties of the Output Correlation Function	49
D Properties of the Error Correlation Coefficient Function	51
E Properties of the Randomized Hadamard Matrix Ensemble	53
F Hadamard Matrix of Order N is Optimal for $M = 2$ and $L_1 = L_2$	55

Chapter 1

An Inequality for the Correlation of Two Functions Operating on Symmetric Bivariate Normal Variables

1.1 Introduction

Statistical characterization of the output of non-linear systems operating on stochastic processes is in general difficult. Nonetheless, when the input process is Gaussian and the system is a memoryless non-linearity, several particularly simple and useful properties are known. Among these are Busgang's theorem [1] and its generalizations (e.g., [2, 3]), results concerning the maximal correlation coefficient [4], as well as results on the output distortion-to-signal power ratio [5]. In this chapter, we describe another simple result that arose in the course of our work on the main topic of this thesis (Chapter 2). The result is an inequality described in the next section. A generalization to a more general class of random variables is described in Section 1.3. An application is presented in Section 1.4. An application of the inequality is presented in Section 1.4.

The material in this chapter appears, in part, in [6].

1.2 Statement of Result and Proof

Lemma 1. Let Z_1 and Z_2 be zero-mean bivariate normal random variables with variance σ^2 and correlation coefficient $\rho > 0$. Then,

$$\mathbb{E}^2 \{g_1(Z_1)g_2(Z_2)\} \leq \mathbb{E} \{g_1(Z_1)g_1(Z_2)\} \mathbb{E} \{g_2(Z_1)g_2(Z_2)\},$$

for any g_1 and g_2 for which the expectations exist, with equality if and only if g_1 and g_2 are equal up to a multiplicative constant.

Note that for $\rho = 1$ the lemma reduces to the standard probabilistic Cauchy-Schwarz inequality.

Proof. Consider the inner product of real functions defined by

$$\langle f, g \rangle \triangleq \mathbb{E} \{f(Z)g(Z)\}$$

where $Z \sim \mathcal{N}(0, \sigma^2)$.

According to Mehler's formula [7] (see also [8]) the joint density function of (Z_1, Z_2) may be written as,

$$\begin{aligned} f_{Z_1, Z_2}(z_1, z_2) &= \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} e^{-\frac{1}{2}\frac{z_1^2 - 2\rho z_1 z_2 + z_2^2}{\sigma^2(1-\rho^2)}} \\ &= \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{z_1^2 + z_2^2}{\sigma^2}} \sum_{n=0}^{\infty} \frac{1}{n!} He_n\left(\frac{z_1}{\sigma}\right) He_n\left(\frac{z_2}{\sigma}\right) \rho^n, \end{aligned} \quad (1.1)$$

where $He_n(x)$ are the probabilists' Hermite polynomials defined as,

$$He_n(x) = (-1)^n e^{\frac{x^2}{2}} \frac{d^n}{dx^n} e^{-\frac{x^2}{2}}, \quad n \geq 0.$$

The Hermite polynomials constitute a complete orthogonal basis of polynomials with respect to the normal probability density function [9], so that

$$\langle \tilde{He}_n, \tilde{He}_m \rangle = \mathbb{E} \{ \tilde{He}_n(Z) \tilde{He}_m(Z) \} = n! \delta_{n,m}$$

where $\tilde{He}_n(x) = He_n(\frac{x}{\sigma})$, and $\delta_{n,m}$ is the Kronecker delta function.

Let $a_{g_i, n} = \langle g_i, \tilde{He}_n \rangle$. The function $g_i(x)$ may be represented by the series

$$g_i(x) = \sum_{n=0}^{\infty} \frac{1}{n!} a_{g_i, n} He_n\left(\frac{x}{\sigma}\right), \quad i = 1, 2. \quad (1.2)$$

The following expectations are obtained by applying (1.1):

$$\begin{aligned}\mathbb{E}\{g_1(Z_1)g_2(Z_2)\} &= \sum_{n=0}^{\infty} \frac{1}{n!} a_{g_1,n} a_{g_2,n} \rho^n, \\ \mathbb{E}\{g_i(Z_1)g_i(Z_2)\} &= \sum_{n=0}^{\infty} \frac{1}{n!} a_{g_i,n}^2 \rho^n, \quad i = 1, 2.\end{aligned}$$

Using these identities and assuming $\rho > 0$ we have,

$$\begin{aligned}\mathbb{E}^2\{g_1(Z_1)g_2(Z_2)\} &= \left(\sum_{n=0}^{\infty} \frac{a_{g_1,n} a_{g_2,n}}{n!} \rho^n \right)^2 \\ &= \left(\sum_{n=0}^{\infty} \frac{a_{g_1,n} \rho^{\frac{n}{2}}}{\sqrt{n!}} \frac{a_{g_2,n} \rho^{\frac{n}{2}}}{\sqrt{n!}} \right)^2 \\ &\leq \left(\sum_{n=0}^{\infty} \frac{a_{g_1,n}^2 \rho^n}{n!} \right) \left(\sum_{n=0}^{\infty} \frac{a_{g_2,n}^2 \rho^n}{n!} \right) \\ &= \mathbb{E}\{g_1(Z_1)g_1(Z_2)\} \mathbb{E}\{g_2(Z_1)g_2(Z_2)\},\end{aligned}\tag{1.3}$$

where the inequality follows by the Cauchy-Schwarz inequality for sequences, which holds with equality only when $a_{g_1,n} = c \cdot a_{g_2,n}$ for some constant c , $\rho > 0$, and for all $n \geq 0$. Since the case of equality holds only when the coefficients in the series (1.2) are equal up to a multiplicative constant, it follows that equality holds only when g_1 and g_2 are equal up to a multiplicative constant. $\square$

When both functions are even or odd, we may apply the lemma to Z_1 and $-Z_2$ to obtain,

Corollary 1. Let Z_1 and Z_2 be zero-mean bivariate normal random variables with variance σ^2 and correlation coefficient $\rho \neq 0$. Then, for g_1 and g_2 that are both even or odd functions,

$$\mathbb{E}^2\{g_1(Z_1)g_2(Z_2)\} \leq \mathbb{E}\{g_1(Z_1)g_1(Z_2)\} \mathbb{E}\{g_2(Z_1)g_2(Z_2)\},$$

for any such g_1 and g_2 for which the expectations exist, with equality if and only if g_1 and g_2 are equal up to a multiplicative constant.

Remark 1. It is interesting to contrast the lemma with the maximal correlation property of normal vectors. Specifically, consider the case where Z_1 and Z_2 are zero mean and both g_1 and g_2 are odd, so that $\mathbb{E}\{g_i(Z_j)\} = 0$, for all $i = 1, 2$ and $j = 1, 2$. Then the maximal

correlation property [4] yields the inequality

$$\mathbb{E}^2 \{g_1(Z_1)g_2(Z_2)\} \leq \rho^2 \mathbb{E} \{g_1^2(Z_i)\} \mathbb{E} \{g_2^2(Z_i)\}.$$

1.3 Generalized Form of Inequality

The lemma may be generalized to a broader class of random variables (Z_1, Z_2) , where Z_1 and Z_2 result from passing some random variable Z through two independent realizations of the same channel. This generalization is stated in the next lemma. We note however that unlike Lemma 1, the “only if” condition for equality does not necessarily apply.

Lemma 2. Let Z_1 and Z_2 be random variables such that for some random variable Z , the Markov relation $Z_1 \leftrightarrow Z \leftrightarrow Z_2$ is satisfied and further, given Z , the pair (Z_1, Z_2) is i.i.d. Then,

$$\mathbb{E}^2 \{g_1(Z_1)g_2(Z_2)\} \leq \mathbb{E} \{g_1(Z_1)g_1(Z_2)\} \mathbb{E} \{g_2(Z_1)g_2(Z_2)\},$$

for any g_1 and g_2 for which the expectations exist, with equality if (but not necessarily only if) g_1 and g_2 are equal up to a multiplicative constant.

Proof. Denote for $i = 1, 2$,

$$\begin{aligned} h_i(z) &= \mathbb{E} \{g_i(Z_i)|Z = z\} \\ &= \mathbb{E} \{g_i(Z_2)|Z = z\} \end{aligned}$$

since Z_1 and Z_2 are i.i.d. given Z . Then,

$$\begin{aligned} \mathbb{E} \{g_1(Z_1)g_2(Z_2)\} &= \mathbb{E} \{\mathbb{E} \{g_1(Z_1)g_2(Z_2)|Z\}\} \\ &= \mathbb{E} \{h_1(Z)h_2(Z)\}. \end{aligned}$$

Following the same steps, we also have,

$$\mathbb{E} \{g_i(Z_1)g_i(Z_2)\} = \mathbb{E} \{h_i^2(Z)\}, \quad i = 1, 2.$$

The claim now follows by applying the Cauchy-Schwarz inequality to obtain

$$\mathbb{E}^2 \{h_1(Z)h_2(Z)\} \leq \mathbb{E} \{h_1^2(Z)\} \mathbb{E} \{h_2^2(Z)\}.$$

□

Remark 2. We note the method that is used in [4] to prove the maximal correlation property of bivariate normal random variables utilizes series expansions involving (probabilists') Hermite polynomials via Mehler's formula, similar to the approach taken in Lemma 1. In contrast, the proof in Lemma 2 follows the approach taken in [10], where an alternative proof to the maximal correlation property is derived.

1.4 Application: A Criterion for Identification of a Memoryless Non-Linearity

Consider a memoryless non-linearity f operating on a discrete-time signal corrupted by additive white Gaussian noise (AWGN), as depicted in Figure 1.1. Thus, the input z_n

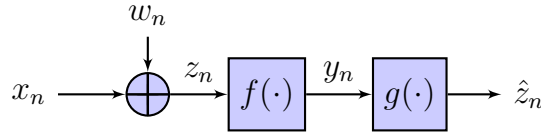


Figure 1.1: Memoryless non-linearity operating on a signal corrupted by AWGN.

consists of the sum of the signal x_n and AWGN w_n having variance σ_w^2 . The output is thus,

$$y_n = f(x_n + w_n).$$

We assume that we observe both the input sequence x_n as well as the output y_n . The function f on the other hand is unknown and we wish to estimate it.

Let us consider first the case where it is known that f is invertible. A possible means to identify f is as follows. Apply another function g to the output to obtain

$$\begin{aligned} \hat{z}_n &= g(y_n) \\ &= g(f(z_n)) \\ &= h(z_n), \end{aligned}$$

where $h = g \circ f$ is the composition of the functions f and g . Assume now that x_n is an AWGN process as well (i.e., a training sequence drawn according to such statistics) with variance σ_x^2 . We may further assume for simplicity that $\mathbb{E}\{h(z_n)\} = 0$. Define

$$K_1 = \frac{\mathbb{E}^2\{h(z_n)x_n\}}{\mathbb{E}\{h^2(z_n)\}\mathbb{E}\{x_n^2\}}.$$

Then, since for bivariate normal random variables nonlinear functions cannot increase (the absolute value of) the correlation coefficient [4], it follows that K_1 is maximized (only) when $g = c \cdot f^{-1}$ for some constant c , so that h is a linear function. As K_1 may be estimated by replacing expectations with time averages, we have obtained a simple criterion for identification of the non-linearity f (up to to a scale factor that may easily be subsequently estimated).

A limitation of the identification criterion described above, is that it does not apply to non-linearities that are not invertible.¹ We now outline how Lemma 1 may serve to overcome this limitation. We note, however, that a drawback of the system described next is that we need to assume that the signal-to-noise ratio σ_x^2/σ_w^2 at the input of the non-linearity is known, unlike for the system described above.

Let $\alpha = \sqrt{\frac{\sigma_x^2 + \sigma_w^2}{\sigma_x^2}}$ so that αx_n has the same variance as z_n . Consider now passing αx_n through a non-linearity g as depicted in Figure 1.2, to obtain $u_n = g(\alpha x_n)$. Define

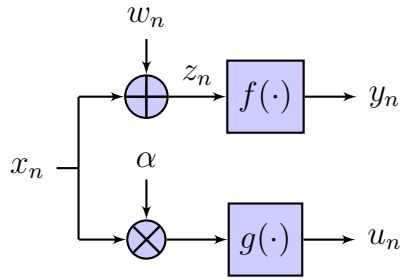


Figure 1.2: A system for identification of a non-linearity.

$$K_2 = \frac{\mathbb{E}\{f(z_n)g(\alpha x_n)\}}{\mathbb{E}\{g(z_n)g(\alpha x_n)\}}.$$

It follows from Lemma 1 that K_2 is maximized (only) when $g = c \cdot f$ for some constant c . Again, K_2 may be computed by replacing expectations with time averages. This is the case, as although we do not observe w_n , we can replace it with AWGN noise w'_n generated with the same variance, to compute

$$\mathbb{E}\{g(z_n)g(\alpha x_n)\} = \mathbb{E}\{g(x_n + w'_n)g(\alpha x_n)\}.$$

¹Nonetheless, it's a suitable criterion for memoryless nonlinear compensation, where the inverse of the non-linearity (if exists) is desired (see, e.g., [11]).

Chapter 2

Dithered Quantization via Orthogonal Transformations

2.1 Introduction

Dithered quantization is a well-known technique to reduce the statistical dependence between the signal and the quantization error, as such dependence is often undesirable. The original and most common approach to achieving this goal is by adding a (pseudo-) random noise prior to quantization, as first proposed by Roberts [12]. Indeed, in *subtractive dithering*, where the dither is further subtracted from the quantizer’s output, statistical independence is achieved in the limit of high resolution. In practical scenarios, when subtractive dithering is considered too complex for implementation, non-subtractive dithering may be employed, achieving a certain measure of “reduced” statistical dependence, see, e.g., [13].

The effectiveness of a dithering method in general depends on the specific application. For instance, in audio and image processing, where dithering is often employed, it is desirable to reduce perceptual artifacts. In this work, in contrast, we will consider the effect of dithering only on the mean-squared error (MSE), similarly to the approach of [15]. In this context, the role of dithering is to reduce second order correlations. Such a criterion is relevant in applications where multiple quantized signals are linearly combined; an example is the multiple description problem (see, e.g., Section 2.3.2).

Our interest will be in quantization of an independent, identically distributed (i.i.d.) Gaussian source. As a baseline for comparison, we will consider subtractive dithered

The material in this chapter appears, in part, in [14].

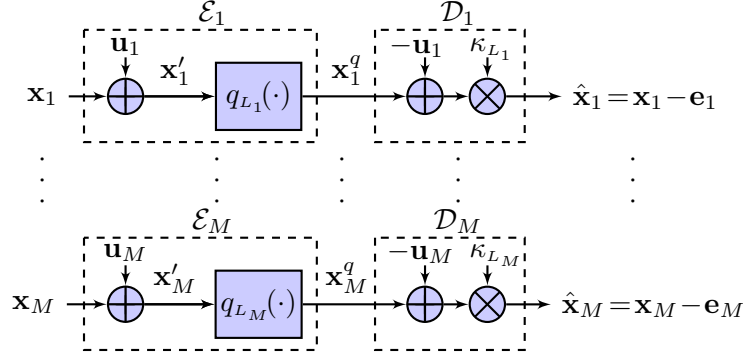


Figure 2.1: A schematic overview of dithering using subtractive dithering.

quantization (SDQ), schematically depicted in Figure 2.1.^{1,2} It is well-known that multi-dimensional SDQ, i.e., nested lattice quantization, can approach the rate-distortion optimal performance in the limit of high dimensional lattices (see, e.g., [16]). Our interest is in low-complexity schemes, and we therefore limit our attention to one-dimensional scalar quantization. Further, we also do not consider entropy coding and thus the bit rate is dictated solely by the number of quantization levels.

Under these constraints, it is known (see also Section 2.3.1) that the performance of SDQ is severely degraded at low bit rates, due to the loss incurred by using a “shifted” (non-symmetric around the origin) quantizer. For example, consider a single branch in Figure 2.1, e.g., SDQ performed on $\mathbf{x}_1$. For a Gaussian signal, the power of the quantization noise of a 1-bit quantizer with SDQ (without imposing zero correlation between input and error) is amplified by a factor of 1.221 w.r.t. its performance without dithering. Furthermore, the correlation coefficient between signal and (quantization) noise is approximately 0.66 which implies that the two signals are rather far from being independent. As will be shown, one may impose zero correlation between input and error by incorporating a multiplicative factor (κ_{L_1} in Figure 2.1) at the output. However, the power of the resulting uncorrelated error for SDQ is amplified to a factor of 1.398 w.r.t. the undithered error (see Section 2.3.1).

Another limitation of SDQ is that it requires using uniform quantizers and precludes source-optimized non-uniform (Lloyd-Max) quantization.³ For the sake of easy compari-

¹The role of scaling the output of the quantizer by κ_L is to impose zero correlation between the input and quantization error of a single branch, where L is the number of quantization levels. For high rates, $\kappa_L \approx 1$, yielding the more common depiction of SDQ.

²The mapping of the quantizer’s output to bits is of no importance for our purposes and hence the mapping/demapping operations are omitted in both encoder and decoder blocks.

³A partial remedy to this drawback is developed in [15].

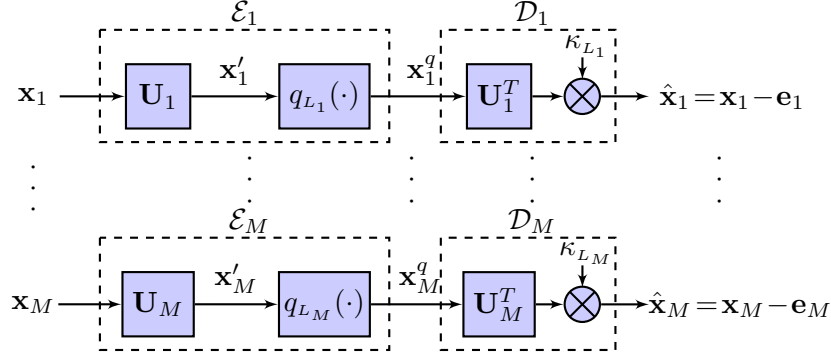


Figure 2.2: A schematic overview of dithering using orthogonal transformations.

son with SDQ, we will assume uniform quantization in the sequel.

Considering a single Gaussian source passing through multiple SDQ branches, correlation exists also between the quantization errors of different branches. Even after imposing zero correlation between input and error in each branch, with 1-bit quantization, the correlation coefficient between the errors in different branches amounts to 0.208.

Our goal will be to design an alternative dithering scheme that operates at least as well, in the sense of achieved distortion, as one-dimensional SDQ (with no entropy coding), at all bit rates, while offering significant improvement at low bit rates. The architecture considered is schematically depicted in Figure 2.2. Essentially, the dither addition and subtraction operations are replaced by multiplication by a judiciously chosen orthogonal transformation and its inverse. In other words, rather than using randomly shifted codebooks, we consider randomly rotated (and reflected) ones.

Unlike for subtractive dithering, which results in an approximately (at high resolution) additive quantization noise model, where the noise is independent of the source, the more modest goal of “second-order” independence will be pursued in this work, where the signal and quantization errors will be required to be mutually uncorrelated. In applications where performance is measured in terms of MSE, this weaker notion of second-order independence is sufficient.

Similarly to subtractive dithering, the proposed scheme requires knowledge of the dither at both transmission ends. Nonetheless, as discussed in the sequel, beyond achieving lower distortion at low bit rates, the scheme may also have implementation advantages in the sense that the resolution of the dither needed to attain good performance may be lower than for SDQ. On the other hand, the scheme involves the application of an orthogonal transformation to a block of samples and thus is ill-suited to delay-sensitive scenarios such as when the output of the quantizer is to be fed into a feedback loop.

Another difference when compared to SDQ is that the marginal distribution of the quantization error (for a large enough block length and under the Lindeberg condition [17], see Section 2.3.3) is approximately Gaussian, whereas for SDQ it is uniform at high bit rates.

The organization of this chapter is as follows. Section 2.3 provides background on SDQ and the multiple description problem, as well as recalls some basic definitions and properties of some ensembles of orthogonal transformations. Section 2.4 presents the proposed scheme. Section 2.5 develops a bound on achievable performance, with an exact analysis for the symmetric 1-bit case. Section 2.6 introduces an algorithm for numerically designing (non-random) orthogonal matrices. Technical proofs are deferred to the Appendix.

2.2 Notation

We use lowercase to denote scalars, e.g., $a \in \mathbb{R}$, boldface lowercase to denote column vectors, e.g., $\mathbf{a} \in \mathbb{R}^N$, and boldface uppercase for matrices, e.g., $\mathbf{A} \in \mathbb{R}^{N \times M}$. The transpose of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}^T$, and its n -th column by $\text{col}_n(\mathbf{A})$. The trace of a square matrix $\mathbf{A}$ is denoted by $\text{tr}(\mathbf{A})$. The ℓ_2 -norm of a vector $\mathbf{a}$ is denoted by $\|\mathbf{a}\| \triangleq \sqrt{\sum_i a_i^2}$. We denote by $\text{diag}(\mathbf{a})$, a square diagonal matrix with the entries of $\mathbf{a}$ appearing on the diagonal. We use square brackets to denote an element-wise operation of a scalar function on a vector or matrix, e.g., $f[\mathbf{A}]$. The n -dimensional identity matrix is denoted by $\mathbf{I}$ and the all-zero matrix by $\mathbf{0}$, where the dimension will be clear from the context.

Throughout the chapter we will consider the behavior of various functions in the vicinity of zero. When comparing two functions, $f(x) \in o(g(x))$ means $\lim_{x \rightarrow 0} \frac{f(x)}{g(x)} = 0$.

We will consider M Gaussian sources, generally spatially correlated, that are i.i.d. w.r.t. the time axis. Thus, each source is an i.i.d. Gaussian vector, $\mathbf{x}_k \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I})$, where $k = 1, \dots, M$. Considering the Gaussian source vectors over a block of N time instants results in a matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$ where

$$\mathbf{X} \triangleq \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_M^T \end{bmatrix},$$

such that $\text{col}_n(\mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ for all $n = 1, \dots, N$. Thus, $\mathbf{C}$ is the spatial covariance matrix with entries

$$c_{k,l} \triangleq (\mathbf{C})_{k,l}, \tag{2.1}$$

where $c_{k,k} = \sigma_k^2$ and $|c_{k,l}| \leq \sigma_k \sigma_l$ for all $k \neq l$.

2.3 Background

2.3.1 Subtractive Dithered Quantization

Consider M branches (realizations) of SDQ elements operating in parallel as depicted in Figure 2.1. The k -th reconstruction ($k=1, \dots, M$), denoted by $\hat{\mathbf{x}}_k$, is described by

$$\begin{aligned}\hat{\mathbf{x}}_k &= \kappa_{L_k} (\mathbf{x}_k^q - \mathbf{u}_k) \\ &= \kappa_{L_k} (q_{L_k} [\mathbf{x}'_k] - \mathbf{u}_k),\end{aligned}$$

where $\mathbf{x}'_k = \mathbf{x}_k + \mathbf{u}_k$, and where $q_{L_k}[\cdot]$ is an element-wise operation of a uniform scalar quantizer with L_k levels and step size Δ_k . A more formal definition of the quantizer is given in Section 2.4. The quantization error is given by

$$\begin{aligned}\mathbf{e}_k &= \mathbf{x}_k - \hat{\mathbf{x}}_k \\ &= \mathbf{x}_k + \kappa_{L_k} (\mathbf{u}_k - q_{L_k} [\mathbf{x}_k + \mathbf{u}_k]).\end{aligned}\tag{2.2}$$

Since the input/output vectors (in scalar SDQ) are i.i.d., we may restrict attention to an arbitrary entry. We denote such an entry by $x_k, u_k, x_k^q, \hat{x}_k$, etc. It is assumed that the step size of $q_{L_k}(\cdot)$ is optimized to minimize (given the choice of dither vector) the average MSE distortion measure

$$\begin{aligned}D_k &= \frac{1}{N} \mathbb{E} \{ \|\mathbf{x}'_k - \mathbf{x}_k^q\|^2 \} \\ &= \mathbb{E} \{ (x'_k - x_k^q)^2 \}.\end{aligned}$$

The most common choice of dither is taking $\mathbf{u}_k$ to be an i.i.d. vector with entries uniformly distributed over the interval $[-\frac{\Delta_k}{2}, \frac{\Delta_k}{2}]$. For this choice of dither, as demonstrated by Schuchman [18], the quantization error is independent of the input, *conditioned* on the event that the quantizer does not overload. For a quantizer optimized to minimize the quantization MSE, this condition in general (and specifically for Gaussian sources) is approximately satisfied at high resolution.

On the other hand, the assumption that the quantization error is independent of the input is far from valid at low bit rates. Nonetheless, we can always impose the condition that the error be *uncorrelated* with the input, i.e.,

$$\mathbb{E}\{x_k e_k\} = 0,\tag{2.3}$$

by choosing the constant κ_{L_k} accordingly. By substituting $e_k = x_k - \hat{x}_k$, and since

$\mathbb{E}\{x_k^2\} = \sigma_k^2$, condition (2.3) is equivalent to

$$\mathbb{E}\{x_k \hat{x}_k\} = \sigma_k^2. \quad (2.4)$$

The constant κ_{L_k} depends on the bit rate, and $\kappa_{L_k} \approx 1$ for high rates. It can be shown (see (A.5) in Appendix A) that the optimization of the step size for minimal distortion when using a uniform dither leads to the condition,

$$\mathbb{E}\{(x'_k - x_k^q)(x_k^q - u_k)\} = 0, \quad (2.5)$$

so that the distortion may be written as

$$\begin{aligned} D_k &= \mathbb{E}\{(x'_k - x_k^q)x_k\} \\ &= \sigma_k^2 - \mathbb{E}\{x_k x_k^q\}. \end{aligned} \quad (2.6)$$

Substituting $\hat{x}_k = \kappa_{L_k}(x_k^q - u_k)$ into (2.4), we obtain

$$\begin{aligned} \kappa_{L_k} &= \frac{\sigma_k^2}{\mathbb{E}\{x_k x_k^q\}} \\ &= \frac{\sigma_k^2}{\sigma_k^2 - D_k}, \end{aligned} \quad (2.7)$$

where the last equality follows from (2.6). We denote the resulting uncorrelated distortion by

$$\begin{aligned} D_k^{\text{uc}} &= \mathbb{E}\{e_k^2\} \\ &= \mathbb{E}\{\hat{x}_k^2\} - \sigma_k^2 \\ &= \frac{\sigma_k^2 D_k}{\sigma_k^2 - D_k}, \end{aligned} \quad (2.8)$$

where the second equality follows from (2.4). The last equality is the well-known relation (see, e.g. [19]) between optimal correlated and uncorrelated distortion and follows from (2.5), (2.6) and (2.7). Note that the dependence of D_k and D_k^{uc} on the number of quantization levels L_k is implicit.

The best tradeoff between distortion and rate one could hope for is given by the information-theoretic distortion-rate function. For a zero-mean Gaussian source with variance σ^2 , the distortion-rate function [20] is given by

$$\underline{D}(R) = \sigma^2 2^{-2R},$$

with corresponding uncorrelated distortion

$$\begin{aligned}\underline{D}^{\text{uc}}(R) &= \frac{\sigma^2 \underline{D}(R)}{\sigma^2 - \underline{D}(R)} \\ &= \frac{\sigma^2}{2^{2R} - 1}.\end{aligned}$$

Thus, $\underline{D}(R)$ is a lower bound on the distortion achieved by any scheme. Similarly, $\underline{D}^{\text{uc}}(R)$ is a lower bound for systems where the condition of zero correlation (between input and error) is imposed.

It is instructive to compare the performance of SDQ with that of undithered (scalar) quantization. The distortion achieved with the latter can be found in [21], and is also computed in Appendix A for various bit rates. It is further shown in Section 2.4 that the relation between the correlated and uncorrelated distortion satisfies (2.8) in the case of undithered quantization as well.

Let $D^{\text{SDQ,uc}}$ denote the uncorrelated distortion with SDQ. Similarly, let $D^{\text{UD,uc}}$ denote the uncorrelated distortion without dithering. In Figure 2.3, the ratios $D^{\text{SDQ,uc}}/\underline{D}^{\text{uc}}$, and $D^{\text{UD,uc}}/\underline{D}^{\text{uc}}$ are plotted as a function of the rate $R = \log_2(L)$, for (scalar uniform) quantization with $L = 2^R$ levels. Evidently, for low bit rates, subtractive dithering incurs considerable performance degradation. The loss is maximal for 1-bit quantization, and amounts to a factor of 1.398 in excess distortion (1.455 dB).

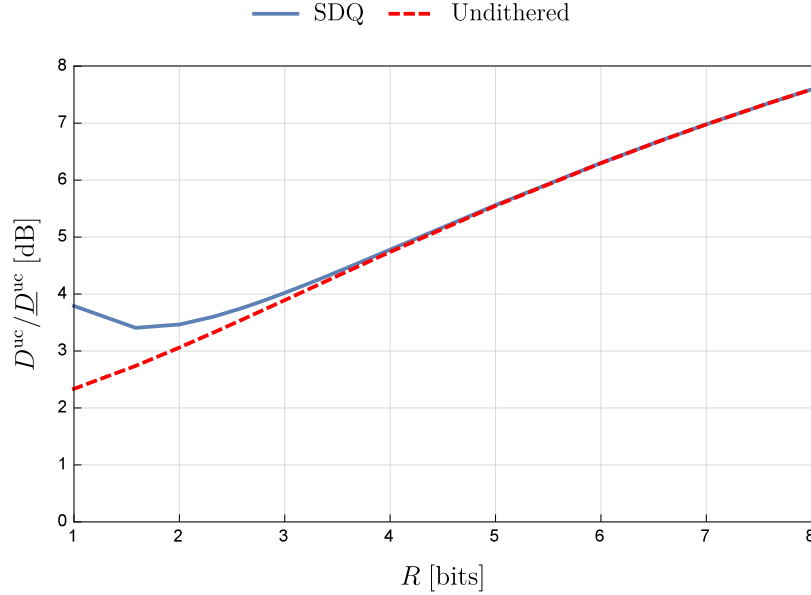


Figure 2.3: The ratio $D^{\text{uc}}/\underline{D}^{\text{uc}}$, as a function of the rate, $R = \log_2(L)$, for SDQ and undithered scalar quantization of a Gaussian source.

2.3.2 Multiple Descriptions via Dithered Quantization

Dithering is important in the implementation of multiple description (MD) coding. In addition, the MD problem will serve as a convenient means to measure how close we approach the goal of second-order independence of the quantization errors of different descriptions (branches), when using orthogonal transformations, in comparison to what is achieved with SDQ.

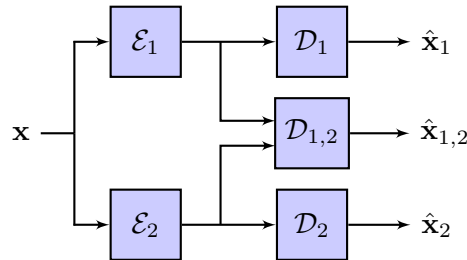


Figure 2.4: The multiple description problem with 2 encoders and 3 decoders.

We briefly recall the MD problem (see e.g., [22] for a survey). Generally, there are $M \geq 2$ different encoders which all encode an identical source resulting in M descriptions, with decoders that may receive any one or more of the descriptions and whose

goal is to reconstruct the source with minimal distortion. The simplest (and most often considered) case of the MD problem is that of $M = 2$ encoders and three possible decoders, as depicted in Figure 2.4. The distortion $D_{1,2}$ achieved by the joint decoder is the central distortion, while the distortions D_1 and D_2 , achieved by decoders receiving individual descriptions are the side distortions. For this case, the MD region is the set of achievable $(R_1, R_2, D_1, D_2, D_{1,2})$, and was explicitly characterized in [23] and [24] for an i.i.d. Gaussian source. There are different tradeoffs between the central and side distortions, and these cannot all be made small simultaneously. We consider only the case where the side distortions are minimal given the chosen rates, referred to as the case of no excess marginal rate/distortion. When considering further the case $R_1 = R_2 = R$ and $D_1 = D_2 = \underline{D} = \sigma^2 2^{-2R}$, which is referred to as the *balanced* case, the central distortion satisfies [23], [24]

$$D_{1,2} \geq \frac{\sigma^2 \underline{D}}{2\sigma^2 - \underline{D}}.$$

Equivalently, the uncorrelated central distortion satisfies

$$\begin{aligned} D_{1,2}^{\text{uc}} &= \frac{\sigma^2 D_{1,2}}{\sigma^2 - D_{1,2}} \\ &\geq \frac{\sigma^2 \underline{D}}{2(\sigma^2 - \underline{D})} \\ &= \frac{\underline{D}^{\text{uc}}}{2}, \end{aligned}$$

where $\underline{D}^{\text{uc}} = \sigma^2 \underline{D} / (\sigma^2 - \underline{D})$ is the uncorrelated side distortion. This lower bound on the central distortion may be achieved if the quantization errors are uncorrelated.

Remark 3. It is well-known that MD scalar quantization for the balanced case can enjoy a distortion reduction of up to a factor of 1/4 w.r.t. the uncorrelated side distortion, using techniques such as index assignment [25] that result in negatively correlated quantization errors. This additional gain (over the factor of 1/2) is possible, since for scalar quantization, there is a gap in the side distortions from the distortion-rate function lower bound, and therefore this is not strictly a case of no excess marginal rate. It is explained in Remark 4 in Section 2.4 below how the full rate/distortion tradeoff may be approached with the proposed dithering scheme.

When SDQ is used with scalar quantization, the second order independence of the quantization errors is satisfied at high rates only. Consequently, there are two sources of performance loss at low to moderate bit rates. One is the excess distortion of the side

distortions w.r.t. undithered scalar quantization, as can be seen in Figure 2.3. The second loss is due to the non-zero correlation between the errors of different branches. We may demonstrate these two sources of loss for the case of a general number of branches M . For simplicity, consider the balanced case, where $D_k = D$ for all $k = 1, \dots, M$, and denote the error correlation coefficient between different branches when SDQ with L quantization levels is used by

$$r^{\text{SDQ}}(L) = \frac{\mathbb{E}\{e_k e_l\}}{D^{\text{SDQ,uc}}(L)}.$$

Table 2.1 gives some numerically computed values of $r^{\text{SDQ}}(L)$. Clearly, the quantization errors are quite far from being uncorrelated for these rates. As we now show, the impact on performance is significant.

L	2	4	8	16	32	64	128	256
$r^{\text{SDQ}}(L)$	0.2081	0.2401	0.2135	0.1813	0.1523	0.1319	0.1147	0.1006

Table 2.1: The error correlation coefficients $r^{\text{SDQ}}(L)$ with $L = 2^R$ levels, for $R = 1, \dots, 8$.

Consider a decoder receiving the M descriptions. Ideally, if the errors are uncorrelated (and also are uncorrelated with the input) and there is no excess distortion due to dithering, simple averaging of the descriptions (which is optimal since we are considering the balanced case) achieves central uncorrelated distortion,

$$D_{\text{central}}^{\text{ideal,uc}}(L, M) = \frac{D^{\text{UD,uc}}(L)}{M},$$

for all bit rates. Applying optimal weighted least squares estimation which again amounts to simple averaging, when correlation between errors is non-zero, yields central distortion

$$\begin{aligned} D_{\text{central}}^{\text{SDQ,uc}}(L, M) &= \mathbb{E}\left\{\left(x - \frac{1}{M}\sum_{k=1}^M \hat{x}_k\right)^2\right\} \\ &= \mathbb{E}\left\{\left(\frac{1}{M}\sum_{k=1}^M e_k\right)^2\right\} \\ &= \frac{1}{M^2}\left(\sum_{k=1}^M \mathbb{E}\{e_k^2\} + \sum_{k=1}^M \sum_{l \neq k}^M \mathbb{E}\{e_k e_l\}\right) \\ &= \frac{D^{\text{SDQ,uc}}(L)}{M} (1 + (M-1)r^{\text{SDQ}}(L)), \end{aligned}$$

where the last equality follows from considering the balanced case. Moreover, for large

M we have that

$$D_{\text{central}}^{\text{SDQ,uc}}(L, M) \xrightarrow{M \rightarrow \infty} D^{\text{SDQ,uc}}(L) r^{\text{SDQ}}(L),$$

rather than approaching zero as in the zero correlation case.

We turn our attention back to the case of $M = 2$ descriptions. In Figure 2.5, the ratio between $D_{\text{central}}^{\text{SDQ,uc}}$ and $D_{\text{central}}^{\text{ideal,uc}}$ is plotted as a function of the bit rate R , illustrating the degradation in performance.

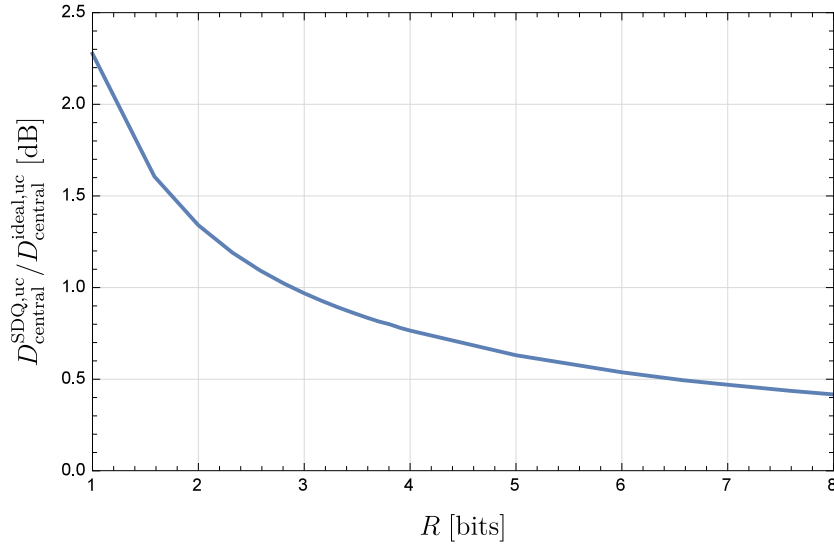


Figure 2.5: The ratio between $D_{\text{central}}^{\text{SDQ,uc}}$ and $D_{\text{central}}^{\text{ideal,uc}}$ as a function of the bit rate $R = \log_2(L)$ for $M=2$ descriptions, with uniform scalar quantization of a Gaussian source.

As we show in the sequel, the ideal (corresponding to zero correlation) central distortion can be approached for all bit rates, assuming a large enough block length N , by replacing SDQ with the proposed dithering scheme.

2.3.3 Orthogonal Matrix Preliminaries

The set of all $N \times N$ real orthogonal matrices is called the *orthogonal group*, $O(N)$.⁴ It is a sub-group of the general linear group $GL(N; \mathbb{R})$, i.e., the group of all $N \times N$ invertible matrices with real entries. Since the determinant of an orthogonal matrix is ± 1 , $O(N)$ has two subgroups. The set of $N \times N$ orthogonal matrices with determinant 1 is called the *special orthogonal group* $SO(N)$. Similarly, the set of orthogonal matrices with determinant -1 is called $SO^-(N)$. Geometrically, elements of $O(N)$ are either rotations, or combinations of rotations and reflections, while the elements of $SO(N)$ are just rotations. The sets $O(N)$ and $SO(N)$ are *compact Lie groups*. That is, they are groups (with matrix multiplication as the operation), and they are manifolds. See [26] for further details.

Random Haar-Distributed Orthogonal Matrix Ensemble

There is a unique translation-invariant probability measure, called the *Haar measure* on $O(N)$. A random $\mathbf{U} \in O(N)$ is uniformly distributed according to the Haar measure, if for any fixed $\mathbf{M} \in O(N)$, we have

$$\mathbf{M}\mathbf{U} \stackrel{d}{=} \mathbf{U}\mathbf{M} \stackrel{d}{=} \mathbf{U},$$

where $\stackrel{d}{=}$ denotes equality in distribution. It is well-known that if $\mathbf{U} \in O(N)$ is uniformly distributed according to the Haar measure, then by the translation-invariance property, all of the entries of $\mathbf{U}$ are identically distributed.

A simple way to generate matrices uniformly distributed w.r.t. the Haar measure on $O(N)$ is the Gaussian approach [27]. We begin with an $N \times N$ matrix $\mathbf{G}$, with i.i.d. standard Gaussian entries. Then, QR -decomposition is performed, such that $\mathbf{G} = \mathbf{Q}\mathbf{R}$, and the diagonal of $\mathbf{R}$ contains only positive entries, resulting in $\mathbf{Q}$ being a Haar-distributed random orthogonal matrix.

The probability density function (PDF) of each entry of an $N \times N$ orthogonal Haar-distributed matrix is given by [28]:

$$f_u(u) = \begin{cases} \nu_N (1 - u^2)^{\frac{N-3}{2}} & , |u| \leq 1 \\ 0 & , \text{else} \end{cases}, \quad (2.9)$$

⁴All orthogonal matrices in this thesis are assumed to have normalized column/row vectors.

where

$$\nu_N = \frac{\Gamma(\frac{N}{2})}{\sqrt{\pi} \Gamma(\frac{N-1}{2})}, \quad (2.10)$$

for all $N \geq 2$, and where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ is the Gamma function. By Stirling's formula $\frac{\Gamma(t+\alpha)}{\Gamma(t+\beta)} \sim t^{\alpha-\beta}$, as $t \rightarrow \infty$. Therefore, asymptotically as $N \rightarrow \infty$,

$$\nu_N \sim \sqrt{\frac{N}{2\pi}}. \quad (2.11)$$

The moments of each entry are given by,

$$\begin{aligned} \mathbb{E}\{u^{2n}\} &= \frac{(2n)!}{2^n n! \prod_{i=0}^{n-1} (N+2i)}, \\ \mathbb{E}\{u^{2n-1}\} &= 0, \end{aligned}$$

for $n \geq 1$, and in particular

$$\begin{aligned} \mathbb{E}\{u^2\} &= \frac{1}{N}, \\ \mathbb{E}\{u^4\} &= \frac{3}{N(N+2)}. \end{aligned} \quad (2.12)$$

Randomized Hadamard Matrix Ensemble

To reduce computational complexity, it may be more useful to utilize an ensemble based on the Hadamard matrices. We further restrict attention to dimensions that are powers of 2 (i.e., we assume that $N = 2^n$), so that we may utilize the fast Walsh-Hadamard transform. We propose to employ (random) orthogonal matrices of the form,

$$\mathbf{U} = \mathbf{H}_N \mathbf{B}, \quad (2.13)$$

where $\mathbf{H}_N$ is the normalized Hadamard matrix of size $N \times N$, and $\mathbf{B}$ is a diagonal orthogonal matrix defined as,

$$\mathbf{B} = \text{diag}(\mathbf{b}),$$

where $\mathbf{b}$ in turn is an i.i.d. vector with entries distributed according to

$$b_i = \begin{cases} 1 & w.p. \ 1/2 \\ -1 & w.p. \ 1/2 \end{cases}, \quad i = 1, \dots, N.$$

Right-multiplying $\mathbf{H}_N$ by $\mathbf{B}$ is equivalent to multiplying the columns of $\mathbf{H}_N$ by ± 1 randomly. It is for this reason that we refer to the ensemble as *randomized* Hadamard.

Since we assume that $N = 2^n$, we may define $\mathbf{H}_N$ by the recursion

$$\begin{aligned} \mathbf{H}_1 &= 1, \\ \mathbf{H}_N &= \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{H}_{N/2} & \mathbf{H}_{N/2} \\ \mathbf{H}_{N/2} & -\mathbf{H}_{N/2} \end{bmatrix}. \end{aligned} \quad (2.14)$$

This ensemble has the advantage of greatly reduced computational complexity (if one leverages the fast Walsh-Hadamard transform), while nearly attaining the same performance as that of the Haar ensemble (as shown in Section 2.5.2).

Lindeberg Condition and Robustness to the Input Distribution

Consider a single quantization branch in the proposed scheme. Although the output of the quantizer is not Gaussian, since it is i.i.d., the resulting entries of $\hat{\mathbf{x}}_k$ converge in distribution to Gaussian random variables (as the dimension N grows), and so do also the entries of $\mathbf{e}_k = \mathbf{x}_k - \hat{\mathbf{x}}_k$, if the orthogonal transformation $\mathbf{U}_k$ satisfies the Lindeberg condition [17].

For a source with finite variance, Lindeberg's condition essentially requires that the entries of each row of the orthogonal matrix are proportional to $N^{-0.5}$, so that there is no single entry that dominates the norm of its row. The orthogonal matrices should therefore be such that this condition is satisfied. It is readily verified that matrices drawn from the two random orthogonal ensembles considered in this thesis indeed satisfy Lindeberg's condition for a large enough dimension, with high probability for the Haar ensemble and always for the randomized Hadamard ensemble.

It is worthwhile noting that if the Lindeberg condition holds, the distortion will essentially remain unchanged if the source is non-Gaussian (but i.i.d.). It follows that optimizing the quantizers for a Gaussian source, ensures that the system is robust to any (i.i.d.) source distribution (c.f. [29]).

2.4 Proposed Architecture and Basic Properties

A schematic overview of the system with M branches (realizations) operating in parallel is depicted in Figure 2.2. Henceforth, we will assume without loss of generality that each of the M source vectors $\mathbf{x}_k$ ($k = 1, \dots, M$), is not only zero-mean Gaussian but also that its covariance matrix is the identity matrix, i.e., the diagonal entries of the spatial covariance matrix $\mathbf{C}$ are $c_{k,k} = 1$, and the off-diagonal entries satisfy $|c_{k,l}| \leq 1$.⁵

Each encoder $\mathcal{E}_k$ left-multiplies $\mathbf{x}_k$ by an orthogonal matrix $\mathbf{U}_k \in \text{O}(N)$, resulting in

$$\mathbf{x}'_k = \mathbf{U}_k \mathbf{x}_k, \quad (2.15)$$

so that $\mathbf{x}'_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, since the distribution of $\mathbf{x}_k$ is invariant to orthogonal transformations. For the time being, the orthogonal matrices may be taken from any ensemble. Specific choices are considered in Section 2.6. Next, scalar quantization of $\mathbf{x}'_k$ is performed, yielding

$$\mathbf{x}_k^q = q_{L_k} [\mathbf{x}'_k].$$

Since the input/output vectors of the scalar quantizers are i.i.d., we may restrict attention to arbitrary entries denoted by x'_k and x_k^q . It is assumed that $q_{L_k}(\cdot)$ is a uniform midrise quantizer with an even number of L_k levels, with step size optimized to minimize the average MSE distortion measure

$$\begin{aligned} D_k &= \frac{1}{N} \mathbb{E} \{ \|\mathbf{x}'_k - \mathbf{x}_k^q\|^2 \} \\ &= \mathbb{E} \{ (x'_k - x_k^q)^2 \}. \end{aligned} \quad (2.16)$$

Note that we could equivalently define the quantizer to be as specified for SDQ in Section 2.3.1, taking the dither to be identically zero. In this respect, there is a slight abuse of notation, as the dependence of q_{L_k} (via the optimal step size) on the dither (or lack of it) is not explicit.

Let us define the output alphabet of such a quantizer by

$$\mathcal{Q}_{L_k} = \left\{ \frac{\Delta_{0,L_k}}{2} (2i + 1 - L_k) \mid i = 0, \dots, L_k - 1 \right\},$$

where $\Delta_{0,L_k} \in \mathbb{R}_+$ is the optimal step size (as given in Appendix A), satisfying the

⁵We assume that the k -th quantizer is matched to its input variance σ_k^2 , such that all quantities (output, error, etc.) are scaled by σ_k accordingly. Since we are only interested in the normalized (Pearson) correlation coefficient between the errors, scaling of the sources has no effect.

optimality condition

$$\mathbb{E}\{(x'_k - x_k^q) x_k^q\} = 0. \quad (2.17)$$

The uniform midrise quantizer is defined by

$$q_{L_k}(x) = \underset{\lambda \in \mathcal{Q}_{L_k}}{\operatorname{argmin}} |x - \lambda|,$$

and is illustrated in Figure 2.6.

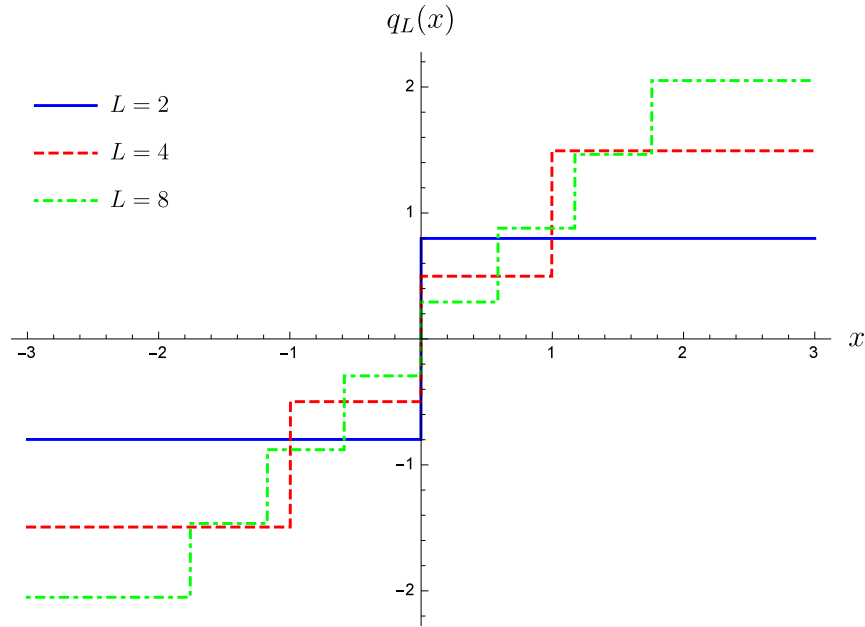


Figure 2.6: Optimal uniform quantizers with $L = 2, 4, 8$ levels, for a Gaussian random variable with zero mean and unit variance.

The decoder $\mathcal{D}_k$ computes

$$\hat{\mathbf{x}}_k = \kappa_{L_k} \mathbf{U}_k^T \mathbf{x}_k^q, \quad (2.18)$$

where $\kappa_{L_k} \in \mathbb{R}_+$. The constants $\{\kappa_{L_k}\}_{k=1}^M$ are chosen so that the error vector

$$\mathbf{e}_k = \mathbf{x}_k - \hat{\mathbf{x}}_k \quad (2.19)$$

$$= \mathbf{U}_k^T (\mathbf{x}'_k - \kappa_{L_k} \mathbf{x}_k^q), \quad (2.20)$$

is uncorrelated with its corresponding source vector, i.e.,

$$\mathbb{E}\{\mathbf{x}_k^T \mathbf{e}_k\} = 0, \quad (2.21)$$

similarly to the approach of [15], for achieving an optimal deterministic quantizer with the constraint that the input and error be uncorrelated.

The correlation may be written using (2.15) and (2.20) as

$$\begin{aligned}\mathbb{E}\{\mathbf{x}_k^T \mathbf{e}_k\} &= \mathbb{E}\{\mathbf{x}_k'^T (\mathbf{x}_k' - \kappa_{L_k} \mathbf{x}_k^q)\} \\ &= N \mathbb{E}\{x_k' (x_k' - \kappa_{L_k} x_k^q)\}.\end{aligned}$$

Imposing that it vanishes, yields

$$\kappa_{L_k} = \frac{1}{\mathbb{E}\{x_k' x_k^q\}} \quad (2.22)$$

$$\begin{aligned}&= \frac{1}{\frac{\Delta_{0,L_k}}{\sqrt{2\pi}} \left(1 + 2 \sum_{i=1}^{\frac{L_k}{2}-1} e^{-\frac{\Delta_{0,L_k}^2}{2} i^2}\right)} \\ &= \frac{1}{1 - D_k},\end{aligned} \quad (2.23)$$

where the second and third equalities follow from (A.3) and (A.4) in Appendix A, respectively. Thus, a “second order additive” (uncorrelated) noise model is obtained for each branch. The constant κ_{L_k} is given in Table 2.2 for various bit rates. Equivalently, from (2.23), we may write the distortion as

$$D_k = 1 - \frac{1}{\kappa_{L_k}}.$$

We denote the uncorrelated distortion of each branch as,

$$\begin{aligned}D_k^{\text{uc}} &= \frac{1}{N} \mathbb{E}\{\|\mathbf{e}_k\|^2\} \\ &= \mathbb{E}\{(x_k' - \kappa_{L_k} x_k^q)^2\} \\ &= \kappa_{L_k} - 1 \\ &= \frac{D_k}{1 - D_k},\end{aligned} \quad (2.24)$$

where the second equality follows from (2.20) using the property that the ℓ_2 -norm is invariant to orthogonal transformations and the fact that the input/output of the scalar quantizer are i.i.d.; the third equality results from (2.17) and (2.22); and to obtain the last equality, (2.23) is used, yielding the well-known relation between optimal D_k^{uc} and D_k .

L	2	4	8	16	32	64	128	256
κ_L	$\frac{\pi}{2}$	1.13488	1.0389	1.01168	1.00351	1.00104	1.0003	1.00009

Table 2.2: Values of κ_L for optimal uniform scalar quantization with $L = 2^R$ levels, $R = 1, \dots, 8$.

2.4.1 Optimization Goal

The proposed scheme is now characterized up to specifying the orthogonal transformations to be used. We now formalize the criterion for choosing the matrices $\{\mathbf{U}_k\}_{k=1}^M$ so as to ensure that the errors of the different branches have vanishing correlation. Let us define the error correlation coefficient.

Definition 1 (Error correlation coefficient). Let

$$r_{k,l}(N) \triangleq \frac{\frac{1}{N} \mathbb{E}\{\mathbf{e}_k^T \mathbf{e}_l\}}{\sqrt{D_k^{\text{uc}}} \sqrt{D_l^{\text{uc}}}}, \quad (2.25)$$

for $k, l = 1, \dots, M$, and $k \neq l$.

Our goal is therefore to find (a family of) orthogonal matrices such that $r_{k,l}(N)$ vanishes as $N \rightarrow \infty$.

Remark 4. While our design goal is to ensure zero correlation between quantization errors, it may be used as a component for approaching the multiple description region, where non-zero correlation is desired. Specifically in [30] an architecture is described for attaining the MD region using entropy coded dithered quantization (ECDQ) in conjunction with successive quantization and quantization splitting. As explained in Section V of [30], the key underlying property that is leveraged is second-order independence. Since the proposed system offers second-order independence, it may replace the ECDQ in the scheme of [30].

2.5 Performance Analysis

We wish to choose the orthogonal ensembles so as to minimize (2.25), which may be written as an expectation over the *conditional* error correlation coefficient defined by

$$r_{k,l}(N; \mathbf{U}_k, \mathbf{U}_l) \triangleq \frac{\frac{1}{N} \mathbb{E}\{\mathbf{e}_k^T \mathbf{e}_l | \mathbf{U}_k, \mathbf{U}_l\}}{\sqrt{D_k^{\text{uc}}} \sqrt{D_l^{\text{uc}}}}, \quad (2.26)$$

so that $r_{k,l}(N) = \mathbb{E}\{r_{k,l}(N; \mathbf{U}_k, \mathbf{U}_l)\}$.

We first investigate the conditional correlation between the error vectors for a given set of (non-random) orthogonal matrices $\{\mathbf{U}_k\}_{k=1}^M$. To that end, we need some additional definitions. Let

$$\mathbf{P}_{k,l} \triangleq \mathbf{U}_k \mathbf{U}_l^T,$$

and denote $\rho_{i,j}^{k,l} \triangleq (\mathbf{P}_{k,l})_{i,j}$.

Definition 2 (Output correlation function). The output correlation function $\gamma_{L_k, L_l} : [-1, 1] \rightarrow \mathbb{R}$ is given by

$$\begin{aligned} \gamma_{L_k, L_l}(\rho) &\triangleq \kappa_{L_k} \kappa_{L_l} \mathbb{E}\{q_{L_k}(x'_k) q_{L_l}(x'_l) | \rho\}, \\ (x'_k, x'_l) | \rho &\sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \end{aligned} \quad (2.27)$$

where $k, l = 1, \dots, M$, and $k \neq l$.

Definition 3 (Error correlation coefficient function). The error correlation coefficient function $\varphi_{L_k, L_l} : [-1, 1] \rightarrow [-1, 1]$ is given by

$$\begin{aligned} \varphi_{L_k, L_l}(\rho) &\triangleq \frac{\mathbb{E}\{(x'_k - \kappa_{L_k} q_{L_k}(x'_k))(x'_l - \kappa_{L_l} q_{L_l}(x'_l)) | \rho\}}{\sqrt{D_k^{\text{uc}}} \sqrt{D_l^{\text{uc}}}}, \\ (x'_k, x'_l) | \rho &\sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \end{aligned} \quad (2.28)$$

where $k, l = 1, \dots, M$, and $k \neq l$.

The output correlation and error correlation coefficient functions play a key role in the analysis. The following two lemmas characterize the output correlation function and are proved in Appendix B and C, respectively. These are followed by a corollary concerning the error correlation coefficient function that is proved in Appendix D.

Lemma 3 (Output correlation function expansion). The output correlation function satisfies:

$$\begin{aligned}\gamma_{L_k, L_l}(\rho) &= \sum_{n=0}^{\infty} \frac{a_{L_k, 2n+1} a_{L_l, 2n+1}}{(2n+1)!} \rho^{2n+1} \\ &= \rho + \frac{1}{6} a_{L_k, 3} a_{L_l, 3} \rho^3 + o(\rho^3) \quad \forall |\rho| \leq 1,\end{aligned}\tag{2.29}$$

using Mehler's formula [8], where

$$a_{L_k, 2n+1} \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \kappa_{L_k} q_{L_k}(z) He_{2n+1}(z) e^{-\frac{z^2}{2}} dz,\tag{2.30}$$

$He_{2n+1}(z)$ being the probabilists' Hermite polynomials. An explicit expression for the Mehler coefficient, $a_{L_k, 2n+1}$ (for general values of k and n), is given by (B.6) in Appendix B. Table 2.3 provides the values of $a_{L, 3}$ according to (B.7), for various bit rates.

L	2	4	8	16	32	64	128	256
$a_{L, 3}$	-1	-0.4555	-0.2036	-0.08474	-0.03297	-0.01212	-0.004242	-0.001428

Table 2.3: The coefficient $a_{L, 3}$ given by (B.7) for $L = 2^R$, $R = 1, \dots, 8$.

Lemma 4 (Output correlation function properties). The output correlation function satisfies the following properties:

1. For the symmetric case where $L_k = L_l$,

$$\gamma_{L_k, L_k}(1) = -\gamma_{L_k, L_k}(-1) = \kappa_{L_k}.$$

2. $\gamma_{L_k, L_l}(\rho)$ is an odd monotonically increasing function of ρ , such that $\gamma_{L_k, L_l}(|\rho|) = |\gamma_{L_k, L_l}(\rho)|$.

3. For the case $L_k = L_l = 2$, we have $\gamma_{2, 2}(\rho) = \arcsin(\rho)$.

Corollary 2 (Error correlation coefficient function properties). The error correlation coefficient function satisfies the following properties:

1. $\varphi_{L_k, L_l}(\rho)$ may be rewritten as

$$\varphi_{L_k, L_l}(\rho) = \frac{\gamma_{L_k, L_l}(\rho) - \rho}{\sqrt{D_k^{\text{uc}}} \sqrt{D_l^{\text{uc}}}}.\tag{2.31}$$

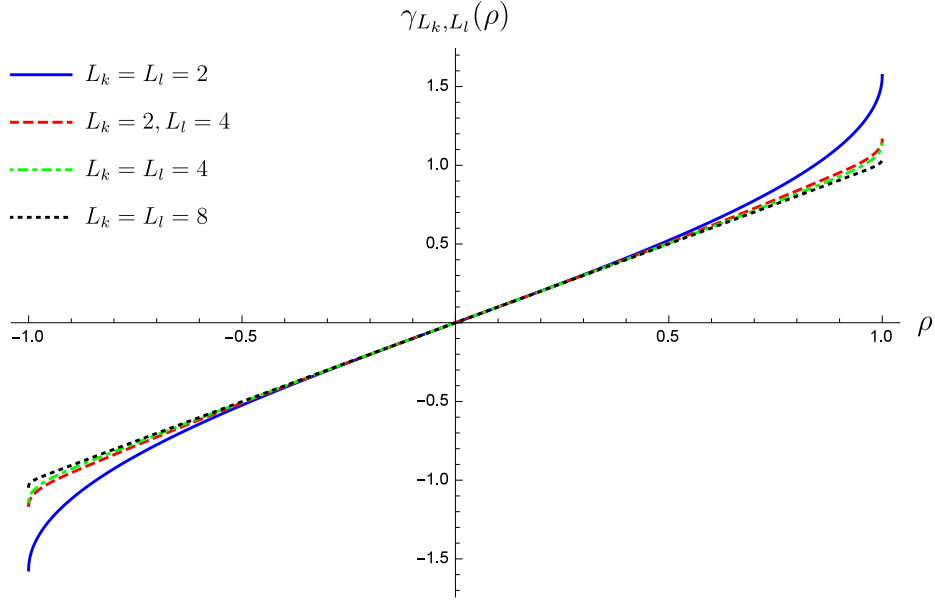


Figure 2.7: Numerical evaluation (by integration of (C.1) in Appendix C) of $\gamma_{L_k, L_l}(\rho)$ as defined in (2.27).

2. The error correlation coefficient function satisfies:

$$\varphi_{L_k, L_l}(\rho) = \sum_{n=1}^{\infty} \frac{1}{(2n+1)!} \frac{a_{L_k, 2n+1} a_{L_l, 2n+1}}{\sqrt{D_k^{\text{uc}}} \sqrt{D_l^{\text{uc}}}} \rho^{2n+1}. \quad (2.32)$$

3. For the symmetric case where $L_k = L_l$,

$$\varphi_{L_k, L_k}(1) = -\varphi_{L_k, L_k}(-1) = 1.$$

4. $\varphi_{L_k, L_l}(\rho)$ is an odd function of ρ , and for $L_k = L_l$ is also a monotonically increasing function of ρ .

5. For all L_k, L_l ,

$$|\varphi_{L_k, L_l}(\rho)| \leq |\rho|^3. \quad (2.33)$$

Figure 2.7 and Figure 2.8 depict the output correlation function and the error correlation coefficient function, respectively, for a few pairs of bit rates.

Remark 5. Numerical analysis of the error correlation coefficient function indicates that $|\varphi_{L_k, L_l}(\rho)| \leq |\varphi_{2,2}(\rho)|$, and also $\varphi_{L_k, L_l}(|\rho|) \geq 0$. That is, the error correlation function for 1-bit quantization serves as an upper bound for all other values of (L_k, L_l) . Further,

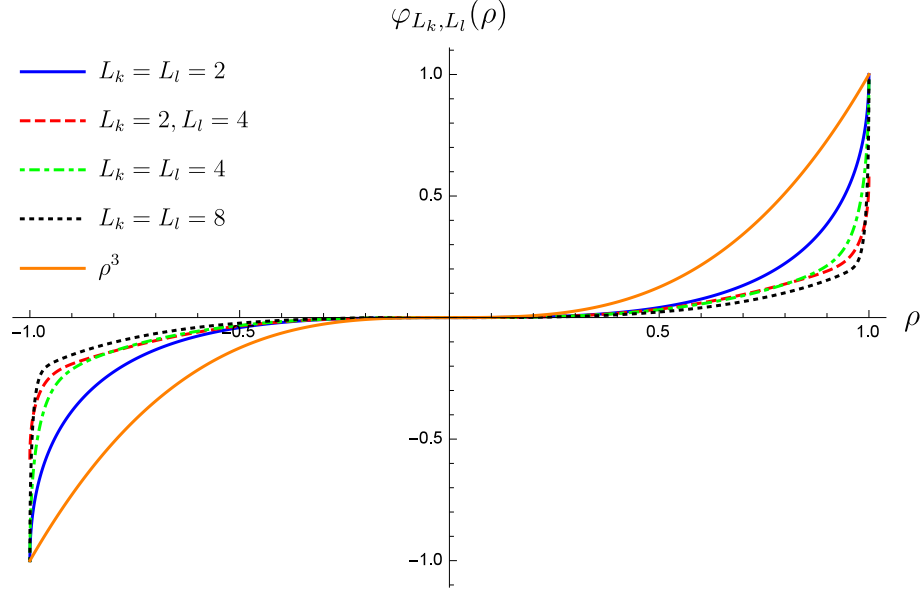


Figure 2.8: Numerical evaluation of $\varphi_{L_k, L_l}(\rho)$ as defined in (2.28). For reference we plot the function ρ^3 that is used as a bound in Corollary 2.

$|\rho| |\varphi_{L_k, L_l}(\rho)| = \rho \varphi_{L_k, L_l}(\rho)$ as a product of odd functions. Finally, $\varphi_{L_k, L_l}(\rho)$ is a monotonically increasing function of ρ . We conjecture that these properties hold and they will guide us in the numerical optimization procedures developed in Section 2.6.

We are now ready to compute the conditional error correlation. The conditional correlation between the errors in branch k and l , given $\mathbf{U}_k$ and $\mathbf{U}_l$, is given by

$$\begin{aligned} \mathbb{E} \{ \mathbf{e}_k^T \mathbf{e}_l | \mathbf{U}_k, \mathbf{U}_l \} &= \mathbb{E} \{ (\mathbf{x}_k - \hat{\mathbf{x}}_k)^T (\mathbf{x}_l - \hat{\mathbf{x}}_l) | \mathbf{U}_k, \mathbf{U}_l \} \\ &= \text{tr}(\mathbb{E} \{ (\mathbf{x}_k - \hat{\mathbf{x}}_k)(\mathbf{x}_l - \hat{\mathbf{x}}_l)^T | \mathbf{U}_k, \mathbf{U}_l \}) \\ &= \text{tr}(\mathbb{E} \{ \mathbf{U}_k^T (\mathbf{x}'_k - \kappa_{L_k} \mathbf{x}_k^q) (\mathbf{x}'_l - \kappa_{L_l} \mathbf{x}_l^q)^T \mathbf{U}_l | \mathbf{U}_k, \mathbf{U}_l \}) \end{aligned} \quad (2.34)$$

$$= \text{tr}(\mathbf{U}_l \mathbf{U}_k^T \mathbb{E} \{ (\mathbf{x}'_k - \kappa_{L_k} \mathbf{x}_k^q) (\mathbf{x}'_l - \kappa_{L_l} \mathbf{x}_l^q)^T | \mathbf{U}_k, \mathbf{U}_l \}) \quad (2.35)$$

$$\begin{aligned} &= \text{tr}(\mathbf{P}_{k,l}^T \mathbb{E} \{ (\mathbf{x}'_k - \kappa_{L_k} \mathbf{x}_k^q) (\mathbf{x}'_l - \kappa_{L_l} \mathbf{x}_l^q)^T | \mathbf{P}_{k,l} \}) \\ &= \sqrt{D_k^{\text{uc}}} \sqrt{D_l^{\text{uc}}} \text{tr}(\mathbf{P}_{k,l}^T \varphi_{L_k, L_l} [c_{k,l} \mathbf{P}_{k,l}]), \end{aligned} \quad (2.36)$$

where (2.34) follows from (2.20), (2.35) follows from the cyclic property of the trace, and in (2.36) the function $\varphi_{L_k, L_l}(\cdot)$ is used element-wise according to (2.28), since $\mathbf{x}_k^q = q_{L_k} [\mathbf{x}'_k]$, $\mathbf{x}_l^q = q_{L_l} [\mathbf{x}'_l]$, and $\mathbb{E} \{ \mathbf{x}'_k \mathbf{x}_l'^T \} = c_{k,l} \mathbf{P}_{k,l}$, where $c_{k,l}$ is the spatial correlation between the sources as defined in (2.1).

We may now express the conditional error correlation coefficient as defined in (2.26) using $\varphi_{L_k, L_l}(\cdot)$ as,

$$\begin{aligned} r_{k,l}(N; \mathbf{U}_k, \mathbf{U}_l) &= r_{k,l}(N; \mathbf{P}_{k,l}) \\ &= \frac{1}{N} \text{tr} \left(\mathbf{P}_{k,l}^T \varphi_{L_k, L_l} [c_{k,l} \mathbf{P}_{k,l}] \right) \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \rho_{i,j}^{k,l} \varphi_{L_k, L_l} (c_{k,l} \rho_{i,j}^{k,l}). \end{aligned} \quad (2.37)$$

Corollary 3. The error correlation coefficient satisfies

$$|r_{k,l}(N)| \leq \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbb{E} \left\{ (\rho_{i,j}^{k,l})^4 \right\}. \quad (2.38)$$

Proof. We obtain the following upper bound on the absolute value of the error correlation coefficient by writing it as an expectation over (2.37).

$$\begin{aligned} |r_{k,l}(N)| &= |\mathbb{E}\{r_{k,l}(N; \mathbf{U}_k, \mathbf{U}_l)\}| \\ &= \left| \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbb{E} \left\{ \rho_{i,j}^{k,l} \varphi_{L_k, L_l} (c_{k,l} \rho_{i,j}^{k,l}) \right\} \right| \\ &\leq \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbb{E} \left\{ |\rho_{i,j}^{k,l}| \cdot |\varphi_{L_k, L_l} (c_{k,l} \rho_{i,j}^{k,l})| \right\} \end{aligned} \quad (2.39)$$

$$\leq \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbb{E} \left\{ (\rho_{i,j}^{k,l})^4 \right\}, \quad (2.40)$$

where (2.39) follows from the triangle inequality, and (2.40) follows from property (2.33) in Corollary 2, and since $|c_{k,l}| \leq 1$ when considering unit variance sources.

2.5.1 Bound on Performance for the Haar and Randomized Hadamard Ensembles

We now consider the performance for the two orthogonal ensembles defined in Section 2.3.3. The next proposition gives an explicit upper bound on the attained error correlation coefficient.

Proposition 1. Consider a random orthogonal dithered quantization system as described by (2.15), (2.18), and (2.19). Then the error correlation coefficient for the Haar distributed ensemble satisfies:

$$|r_{k,l}^{\text{Haar}}(N)| \leq \frac{3}{N+2} \quad \forall k \neq l.$$

Similarly for the randomized Hadamard ensemble ($N = 2^n$):

$$|r_{k,l}^{\text{rHad}}(N)| \leq \frac{3N-2}{N^2} \quad \forall k \neq l.$$

In particular, the error correlation coefficient vanishes as $N \rightarrow \infty$ for both ensembles.

Proof. For the Haar distributed ensemble, by the translation invariance property of the Haar measure, the product of any two independent Haar-distributed random orthogonal matrices $\mathbf{P}_{k,l} = \mathbf{U}_k \mathbf{U}_l^T$ is also Haar distributed. It follows that the entries of $\mathbf{P}_{k,l}$ remain identically distributed. Using (2.12) and (2.38) we have,

$$\begin{aligned} |r_{k,l}^{\text{Haar}}(N)| &\leq \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbb{E} \left\{ (\rho_{i,j}^{k,l})^4 \right\} \\ &= N \mathbb{E} \left\{ \rho^4 \right\} \\ &= \frac{3}{N+2}. \end{aligned}$$

We turn now to the randomized Hadamard ensemble. It is shown in Appendix E that for this ensemble too, all entries of $\mathbf{P}_{k,l}$ have the same distribution, and this distribution is given by

$$\Pr \left\{ \rho = 2 \frac{n}{N} - 1 \right\} = \binom{N}{n} 2^{-N}, \quad n = 0, \dots, N. \quad (2.41)$$

Straightforward calculation reveals that the fourth moment of each entry is given by

$$\mathbb{E} \{ \rho^4 \} = \frac{3N-2}{N^3}.$$

Using these two properties, we obtain

$$\begin{aligned} |r_{k,l}^{\text{rHad}}(N)| &\leq \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbb{E} \left\{ (\rho_{i,j}^{k,l})^4 \right\} \\ &= N \mathbb{E} \left\{ \rho^4 \right\} \\ &= \frac{3N-2}{N^2}. \end{aligned}$$

Remark 6. We note that Proposition 1 is not limited to uniform quantization. It holds for the system described by (2.15), (2.18), and (2.19) as long as the scalar quantizers satisfy the constraint that input and error are uncorrelated. This follows since the bound (2.33) only requires that the latter property is satisfied.

2.5.2 Performance Analysis for Symmetric 1-bit Quantization

For the symmetric 1-bit case ($L_k = L_l = 2$), we derive an exact expression for the error correlation coefficient, for the two orthogonal ensembles. Since the matrices of both ensembles have identically distributed entries, the error correlation coefficient may be written as

$$r_{k,l}(N) = N \mathbb{E}\{\rho \varphi_{2,2}(c_{k,l}\rho)\},$$

where by property 3 of Lemma 4

$$\varphi_{2,2}(c_{k,l}\rho) = \frac{\arcsin(c_{k,l}\rho) - c_{k,l}\rho}{\pi/2 - 1}.$$

Since $\rho \varphi_{2,2}(c_{k,l}\rho)$ is a monotonically increasing function of $c_{k,l}$, it is maximized by $c_{k,l} = 1$. We therefore define the worst-case correlation coefficient for 1-bit quantization as

$$r_{1\text{-bit}}(N) \triangleq r_{k,l}(N) \Big|_{\substack{L_k=L_l=2, \\ c_{k,l}=1}}.$$

Random Haar-Distributed Orthogonal Matrix Ensemble

Using the distribution of ρ as given in (2.9), by standard integration methods, we obtain

$$\begin{aligned} r_{1\text{-bit}}^{\text{Haar}}(N) &= N \mathbb{E}\{\rho \varphi_{2,2}(\rho)\} \\ &= \frac{1}{\pi/2 - 1} \left(N \int_{-1}^1 \nu_N (1 - \rho^2)^{\frac{N-3}{2}} \rho \arcsin(\rho) d\rho - 1 \right) \\ &= \frac{1}{\pi/2 - 1} \left(\nu_N N \int_{-\pi/2}^{\pi/2} \cos(t)^{N-2} \sin(t) t dt - 1 \right) \\ &= \frac{1}{\pi/2 - 1} \left(\nu_N \frac{N}{N-1} \int_{-\pi/2}^{\pi/2} \cos(t)^{N-1} dt - 1 \right) \\ &= \frac{1}{\pi/2 - 1} \left(2\pi \nu_N^2 \frac{N}{(N-1)^2} - 1 \right), \end{aligned} \tag{2.42}$$

where ν_N is defined in (2.10). Note that by (2.11), $r_{1\text{-bit}}^{\text{Haar}}(N)$ vanishes as $N \rightarrow \infty$ as expected.

Randomized Hadamard Matrix Ensemble

Using the distribution of ρ as given by (2.41), we derive the following expression for $r_{1\text{-bit}}^{\text{rHad}}(N)$:

$$\begin{aligned}
 r_{1\text{-bit}}^{\text{rHad}}(N) &= N\mathbb{E}\{\rho\varphi_{2,2}(\rho)\} \\
 &= \frac{N}{2^N} \sum_{n=0}^N \binom{N}{n} \left(2\frac{n}{N} - 1\right) \varphi_{2,2}\left(2\frac{n}{N} - 1\right) \\
 &= \frac{N}{2^{N-1}} \sum_{n=0}^N \binom{N}{n} \frac{n}{N} \varphi_{2,2}\left(2\frac{n}{N} - 1\right) \\
 &= \frac{N}{2^{N-1}} \sum_{n=1}^N \binom{N-1}{n-1} \varphi_{2,2}\left(2\frac{n}{N} - 1\right), \tag{2.43}
 \end{aligned}$$

where in the third equality we used the symmetry of the distribution of ρ , and the fact that $\varphi_{2,2}(\rho)$ is an odd function.

A comparison between the error correlation coefficient (for 1-bit quantization and $c_{k,l} = 1$) using the Haar and randomized Hadamard ensembles is given in Table 2.4.

N	2	4	8	16	32	64	128	256
$r_{1\text{-bit}}^{\text{Haar}}(N)$	0.478	0.231	0.113	0.0555	0.0275	0.0137	0.00685	0.00343
$r_{1\text{-bit}}^{\text{rHad}}(N)$	1	0.541	0.167	0.0618	0.0289	0.0141	0.00694	0.00345

Table 2.4: The exact values of $r_{1\text{-bit}}(N)$ using the Haar and randomized Hadamard ensembles for $N = 2^n$, $n = 1, \dots, 8$.

2.5.3 Conjectured Tighter Bound on the Error Correlation Coefficient

We note that the bound $|\varphi_{L_k, L_l}(\rho)| \leq |\rho|^3$ (as appears in (2.33)), while being simple, is however quite loose, as can be inferred from Figure 2.8. As a consequence, the performance guarantee provided in Proposition 1 is also quite loose. We observed through numerical analysis that the true performance is governed by replacing in inequality (2.39), the bound $|\varphi_{L_k, L_l}(\rho)| \leq |\rho|^3$ with the conjectured bound $|\varphi_{L_k, L_l}(\rho)| \leq |\varphi_{2,2}(\rho)|$ (see Remark 5 above).

Once this is done, one arrives at the tighter (conjectured) bound $|r_{k,l}(N)| \leq r_{1\text{-bit}}(N)$, where $r_{1\text{-bit}}(N)$ is given by (2.42) for the Haar ensemble, and by (2.43) for the randomized Hadamard ensemble. The conjectured bound appears (based on numerical evidence) to hold also for optimal non-uniform scalar quantization.

2.6 Numerical Design of Orthogonal Transformations

We would like to find a set of non-random orthogonal matrices $\{\mathbf{U}_1, \dots, \mathbf{U}_M\}$ that achieve error correlation coefficients as close as possible to zero. As an aid to design, we consider a simple case depicted in Figure 2.9. Here, a single Gaussian source $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ gives rise to M branches as described in Section 2.4.⁶ Since in the setting of Figure 2.9 the error correlation coefficients are non-negative (see Remark 5), the best one can hope for is that they are all zero. To obtain a simple (scalar) objective function let us consider the average conditional error correlation coefficient over all different branch pairs

$$\begin{aligned} \bar{r}(N, M; \{\mathbf{P}_{k,l}\}) &\triangleq \frac{\sum_{k=1}^M \sum_{l \neq k} r_{k,l}(N; \mathbf{P}_{k,l})}{M(M-1)} \\ &= \frac{\sum_{k=1}^M \sum_{l \neq k} \frac{1}{N} \text{tr}(\mathbf{P}_{k,l}^T \varphi_{L_k, L_l}[\mathbf{P}_{k,l}])}{M(M-1)}, \end{aligned} \quad (2.44)$$

where $\mathbf{P}_{k,l} = \mathbf{U}_k \mathbf{U}_l^T$ and the second equality follows by substituting (2.37) with $c_{k,l} = 1$.

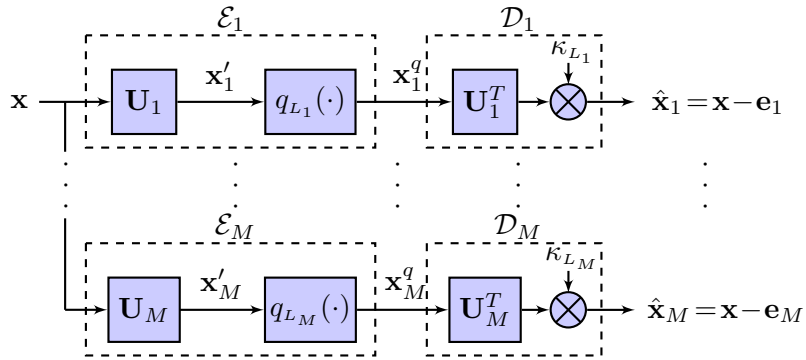


Figure 2.9: A scalar quantization system using orthogonal transformations with M branches and a single Gaussian source.

⁶The use of a single source is equivalent to the case of $c_{k,l} = 1$ for all k, l in Section 2.4. As we've observed via numerical analysis, these are the worst conditions for a choice of a spatial covariance matrix, and therefore the optimization is carried out for such a covariance matrix.

2.6.1 Numerical Optimization Procedure

Our objective is to find a family of orthogonal matrices for a given number of branches M and a block of length N , for which the average error correlation coefficient $\bar{r}(N, M; \{\mathbf{P}_{k,l}\})$ as given in (2.44) is minimized. We pose this as the following minimization problem,

$$\min_{\substack{\mathbf{U}_1, \dots, \mathbf{U}_M \\ \mathbf{U}_k \mathbf{U}_k^T = \mathbf{I} \ \forall k}} \bar{r}(N, M; \{\mathbf{P}_{k,l}\}) = \min_{\substack{\mathbf{U}_1, \dots, \mathbf{U}_M \\ \mathbf{U}_k \mathbf{U}_k^T = \mathbf{I} \ \forall k}} \frac{\sum_{k=1}^M \sum_{l \neq k} \frac{1}{N} \text{tr}(\mathbf{P}_{k,l}^T \varphi_{L_k, L_l}[\mathbf{P}_{k,l}])}{M(M-1)} \quad (2.45)$$

where $\mathbf{P}_{k,l} = \mathbf{U}_k \mathbf{U}_l^T$.

We choose to optimize for the balanced case, where the quantization rates of all branches are equal. We observed that this choice is the most challenging case, and thus results in a system that is robust to different choices of quantization rates in the various branches. Further, recall that, as stated in Remark 5, we conjecture that $|\varphi_{L_k, L_l}(\rho)| \leq |\varphi_{2,2}(\rho)|$. Therefore, we will optimize the orthogonal matrices, setting $L_k = 2$ for all $k = 1, \dots, M$. Thus, our objective is to find matrices $\{\mathbf{U}_1, \dots, \mathbf{U}_M\}$ minimizing

$$\begin{aligned} \bar{r}_{1\text{-bit}}(N, M; \{\mathbf{P}_{k,l}\}) &\triangleq \frac{\sum_{k=1}^M \sum_{l \neq k} \frac{1}{N} \text{tr}(\mathbf{P}_{k,l}^T \varphi_{2,2}[\mathbf{P}_{k,l}])}{M(M-1)} \\ &= \frac{\sum_{k=1}^M \sum_{l \neq k} \left(\frac{1}{N} \text{tr}(\mathbf{P}_{k,l}^T \arcsin[\mathbf{P}_{k,l}]) - 1 \right)}{M(M-1)(\pi/2 - 1)}. \end{aligned}$$

To that end, we will find a local minimum of the following objective function,

$$h(\mathbf{U}_1, \dots, \mathbf{U}_M) = \sum_{k=1}^M \sum_{l \neq k} \text{tr} \left((\mathbf{U}_k \mathbf{U}_l^T)^T \arcsin[\mathbf{U}_k \mathbf{U}_l^T] \right). \quad (2.46)$$

Finding orthogonal matrices (locally) minimizing (2.46) requires performing numerical search over the Lie group $\text{SO}(N)$ defined in Section 2.3.3. Without the orthogonality constraint, the standard gradient descent procedure could be used. However, as there would be no guarantee that the output matrices are orthogonal, this method is inapplicable.

In order to use a gradient descent procedure that starts with orthogonal matrices and maintains the orthogonality constraints, we use the method of *geodesic flow* on Lie groups. This method, introduced in [31], locates a local minimum as explained in [32]. It was further applied and summarized succinctly in [33]. For our purposes, a very concise description will do, and we refer the reader to [32] for a detailed account.

The essence of the procedure is as follows. Given an objective function f of a matrix $\mathbf{U}$, and a constant step size μ , we recursively update according to

$$\begin{aligned}\mathbf{U}^{(i+1)} &= \exp(-\mu \mathbf{G}^{(i)}) \mathbf{U}^{(i)} \\ \mathbf{G}^{(i)} &= \mathbf{\Gamma}^{(i)} \mathbf{U}^{(i)T} - \mathbf{U}^{(i)} \mathbf{\Gamma}^{(i)T},\end{aligned}\tag{2.47}$$

where $\mathbf{\Gamma}^{(i)} = \frac{\partial f}{\partial \mathbf{U}}(\mathbf{U}^{(i)})$ is the usual Euclidean gradient at $\mathbf{U}^{(i)}$, and $\exp(\mathbf{A})$ is defined by the power series,

$$\exp(\mathbf{A}) = \sum_{n=0}^{\infty} \frac{1}{n!} \mathbf{A}^n,$$

where $\mathbf{A}^0 = \mathbf{I}$. This results in $\exp(-\mu \mathbf{G}^{(i)})$ being an element of $\text{SO}(N)$. Therefore, the search is restricted to the space of orthogonal matrices when starting with an orthogonal matrix $\mathbf{U}^{(0)}$.

Algorithm Description

The algorithm consists of the following steps.

1. Initial orthogonal matrices: $\mathbf{U}_k^{(0)}$, $k = 1, \dots, M$, are Haar-distributed random orthogonal matrices generated independently.
2. Steps for a single matrix $\mathbf{U}_k^{(i)}$ in the i th iteration (in each iteration perform for $\mathbf{U}_k^{(i)}$, $k = 1, \dots, M$, in ascending order):
 - (a) Gradient computation:

$$\begin{aligned}\mathbf{\Gamma}_k^{(i)} &= \frac{\partial h}{\partial \mathbf{U}_k}(\mathbf{U}_1^{(i+1)}, \dots, \mathbf{U}_{k-1}^{(i+1)}, \mathbf{U}_k^{(i)}, \dots, \mathbf{U}_M^{(i)}) \\ \mathbf{G}_k^{(i)} &= \mathbf{\Gamma}_k^{(i)} \mathbf{U}_k^{(i)T} - \mathbf{U}_k^{(i)} \mathbf{\Gamma}_k^{(i)T},\end{aligned}$$

where for the objective function $h(\mathbf{U}_1, \dots, \mathbf{U}_M)$ as given in (2.46),

$$\frac{\partial h}{\partial \mathbf{U}_k} = 2 \sum_{l \neq k} \beta[\mathbf{U}_k \mathbf{U}_l^T] \mathbf{U}_l$$

and $\beta(x) = \frac{d}{dx}(x \arcsin(x)) = \arcsin(x) + \frac{x}{\sqrt{1-x^2}}$ is used element-wise.

- (b) Update:

$$\mathbf{U}_k^{(i+1)} = \exp(-\mu \mathbf{G}_k^{(i)}) \mathbf{U}_k^{(i)}.$$

3. Stopping criterion: Stop when relative change in objective function is less than a chosen $\epsilon > 0$, i.e.,

$$\left| \frac{h(\mathbf{U}_1^{(i)}, \dots, \mathbf{U}_M^{(i)}) - h(\mathbf{U}_1^{(i+1)}, \dots, \mathbf{U}_M^{(i+1)})}{h(\mathbf{U}_1^{(i)}, \dots, \mathbf{U}_M^{(i)})} \right| < \epsilon,$$

else $i := i + 1$, and go to step 2.

Results

We were able to obtain very satisfying results for $M = 2, 3, 4$ branches with dimension ranging up to $N = 16$. For larger block lengths, it appears that a local minimum is obtained, which depends on the initial random matrices drawn. For the set $\{\mathbf{U}_1, \dots, \mathbf{U}_M\}$ that we obtain from the algorithm, we have that indeed $\bar{r}_{1\text{-bit}}(N, M; \{\mathbf{P}_{k,l}\}) \leq r_{1\text{-bit}}(N)$ for both considered random ensembles.

In Appendix F we show for $M = 2$ and quantizers of equal rate, taking $\mathbf{P}_{1,2}$ as a Hadamard matrix minimizes $\bar{r}(N, 2; \mathbf{P}_{1,2})$, for dimensions N where such a matrix exists. Therefore, for this case ($M = 2$), we measure the effectiveness of our optimization algorithms by comparing the results with that attainable using a Hadamard matrix. That is, we take the solution of $\mathbf{U}_1, \mathbf{U}_2$ satisfying $\mathbf{P}_{1,2} = \mathbf{U}_1 \mathbf{U}_2^T = \mathbf{H}_N$ as our benchmark. The corresponding expression for $\bar{r}(N, 2; \mathbf{H}_N)$ (for quantizers of equal rate) is given by (F.3). Specifically, for $L_1 = L_2 = 2$, substituting (2.31) in (F.3) we have

$$\begin{aligned} \bar{r}_{1\text{-bit}}(N, 2; \mathbf{H}_N) &= \frac{\sqrt{N} \arcsin(\frac{1}{\sqrt{N}}) - 1}{\pi/2 - 1} \\ &= \frac{\frac{1}{6} \frac{1}{N} + o(\frac{1}{N})}{\pi/2 - 1}, \end{aligned}$$

as given in Table 2.5.

N	2	4	8	16	32	64	128	256
$\bar{r}_{1\text{-bit}}(N, 2; \mathbf{H}_N)$	0.194	0.0827	0.0387	0.0188	0.0093	0.0046	0.0023	0.0011

Table 2.5: The error correlation coefficient for $M = 2$ branches with $L_1 = L_2 = 2$, when $\mathbf{P}_{1,2} = \mathbf{H}_N$, for $N = 2^n$, $n = 1, \dots, 8$.

In this case ($M = 2$) the algorithm indeed locates $\mathbf{U}_1, \mathbf{U}_2$ with $\mathbf{P}_{1,2} = \mathbf{H}_N$ for $N = 2, 4, 8, 12, 16$, but for larger N (for which a Hadamard matrix exists) it usually locates a matrix $\mathbf{P}_{1,2}$ with no apparent structure and slightly worse performance.

Interestingly, for small values of M and N (where the Hadamard structure is inapplicable), the located matrices $\mathbf{P}_{k,l}$ nonetheless exhibit clear structure, where also the alphabet cardinality of the entries is small. This hints at the possibility of global optimality. We give some examples:

- For $M = 2$ and $N = 6$, the choice of $\mathbf{U}_1$ and $\mathbf{U}_2$ with

$$\mathbf{P}_{1,2} = \mathbf{U}_1 \mathbf{U}_2^T = \frac{1}{\sqrt{5}} \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & -1 & 1 & -1 & 1 \\ 1 & -1 & 0 & -1 & 1 & 1 \\ 1 & 1 & -1 & 0 & 1 & -1 \\ 1 & -1 & 1 & 1 & 0 & -1 \\ 1 & 1 & 1 & -1 & -1 & 0 \end{bmatrix}$$

is located, which results in $\bar{r}_{1\text{-bit}}(6, 2; \mathbf{P}_{1,2}) = 0.0644$. Similar solutions with the entries of $\mathbf{P}_{1,2}$ belonging to the alphabet $\{0, \pm \frac{1}{\sqrt{N-1}}\}$ are located for $N = 10, 14$, with $\bar{r}_{1\text{-bit}}(10, 2; \mathbf{P}_{1,2}) = 0.0342$, and $\bar{r}_{1\text{-bit}}(14, 2; \mathbf{P}_{1,2}) = 0.0233$ respectively. One method to extract the matrices $\mathbf{U}_1, \mathbf{U}_2$ is via the singular value decomposition of $\mathbf{P}_{1,2}$. Alternatively, we may take $\mathbf{U}_1 = \mathbf{I}$ and $\mathbf{U}_2 = \mathbf{P}_{1,2}$.

- For $M = 3$ and $N = 4$, an interesting solution is found,

$$\mathbf{U}_1 = \mathbf{I}, \mathbf{U}_2 = \mathbf{H}_4, \mathbf{U}_3 = \frac{1}{2} \begin{bmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix},$$

such that all the entries of $\mathbf{P}_{1,2}$, $\mathbf{P}_{1,3}$, and $\mathbf{P}_{2,3}$ are $\pm \frac{1}{2}$, with $\bar{r}_{1\text{-bit}}(4, 3; \{\mathbf{P}_{k,l}\}) = \bar{r}_{1\text{-bit}}(4, 2; \mathbf{H}_4) = 0.0827$.

- For $M = 3$ and $N = 8$, the solution $\mathbf{U}_1 = \mathbf{I}$,

$$\mathbf{U}_2 = \frac{1}{\sqrt{28}} \begin{bmatrix} \sqrt{7} & -\sqrt{3} & \sqrt{3} & \sqrt{3} & \sqrt{3} & -\sqrt{3} & \sqrt{3} & -\sqrt{3} \\ \sqrt{3} & \sqrt{7} & -\sqrt{3} & -\sqrt{3} & \sqrt{3} & \sqrt{3} & \sqrt{3} & -\sqrt{3} \\ \sqrt{3} & -\sqrt{3} & -\sqrt{7} & -\sqrt{3} & \sqrt{3} & -\sqrt{3} & -\sqrt{3} & \sqrt{3} \\ \sqrt{3} & -\sqrt{3} & \sqrt{3} & -\sqrt{7} & -\sqrt{3} & \sqrt{3} & -\sqrt{3} & -\sqrt{3} \\ \sqrt{3} & \sqrt{3} & -\sqrt{3} & \sqrt{3} & -\sqrt{7} & -\sqrt{3} & -\sqrt{3} & -\sqrt{3} \\ \sqrt{3} & -\sqrt{3} & -\sqrt{3} & \sqrt{3} & -\sqrt{3} & \sqrt{7} & \sqrt{3} & \sqrt{3} \\ \sqrt{3} & \sqrt{3} & \sqrt{3} & \sqrt{3} & \sqrt{3} & \sqrt{3} & -\sqrt{7} & \sqrt{3} \\ \sqrt{3} & \sqrt{3} & \sqrt{3} & -\sqrt{3} & -\sqrt{3} & -\sqrt{3} & \sqrt{3} & \sqrt{7} \end{bmatrix},$$

$$\mathbf{U}_3 = \frac{1}{\sqrt{28}} \begin{bmatrix} \sqrt{7} & \sqrt{3} & -\sqrt{3} & -\sqrt{3} & -\sqrt{3} & \sqrt{3} & -\sqrt{3} & \sqrt{3} \\ \sqrt{3} & -\sqrt{7} & -\sqrt{3} & -\sqrt{3} & \sqrt{3} & \sqrt{3} & \sqrt{3} & -\sqrt{3} \\ \sqrt{3} & -\sqrt{3} & \sqrt{7} & -\sqrt{3} & \sqrt{3} & -\sqrt{3} & -\sqrt{3} & \sqrt{3} \\ \sqrt{3} & -\sqrt{3} & \sqrt{3} & \sqrt{7} & -\sqrt{3} & \sqrt{3} & -\sqrt{3} & -\sqrt{3} \\ \sqrt{3} & \sqrt{3} & -\sqrt{3} & \sqrt{3} & \sqrt{7} & -\sqrt{3} & -\sqrt{3} & -\sqrt{3} \\ \sqrt{3} & -\sqrt{3} & -\sqrt{3} & \sqrt{3} & -\sqrt{3} & -\sqrt{7} & \sqrt{3} & \sqrt{3} \\ \sqrt{3} & \sqrt{3} & \sqrt{3} & \sqrt{3} & \sqrt{3} & \sqrt{3} & \sqrt{7} & \sqrt{3} \\ \sqrt{3} & \sqrt{3} & \sqrt{3} & -\sqrt{3} & -\sqrt{3} & -\sqrt{3} & \sqrt{3} & -\sqrt{7} \end{bmatrix}$$

is located with $\bar{r}_{1\text{-bit}}(8, 3; \{\mathbf{P}_{k,l}\}) = 0.0453$.

- For $M = 3$ and $N = 12$, with the constraint that $\mathbf{U}_1 = \mathbf{I}$ and $\mathbf{U}_2 = \mathbf{H}_{12}$ (in standard form and normalized),

$$\mathbf{U}_3 = \frac{1}{\sqrt{12}} \begin{bmatrix} 1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix}$$

is located with $\bar{r}_{1\text{-bit}}(12, 3; \{\mathbf{P}_{k,l}\}) = 0.0283$.

- For $M = 4$ and $N = 4$, the solution

$$\mathbf{U}_1 = \mathbf{I}, \mathbf{U}_2 = \frac{1}{\sqrt{3}} \begin{bmatrix} 0 & 1 & -1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & -1 \\ 1 & -1 & -1 & 0 \end{bmatrix}, \mathbf{U}_3 = \frac{1}{\sqrt{6}} \begin{bmatrix} \sqrt{3} & -1 & 1 & -1 \\ 1 & \sqrt{3} & 1 & 1 \\ 1 & 1 & -\sqrt{3} & -1 \\ 1 & -1 & -1 & \sqrt{3} \end{bmatrix}, \mathbf{U}_4 = \frac{1}{\sqrt{6}} \begin{bmatrix} \sqrt{3} & 1 & -1 & 1 \\ 1 & -\sqrt{3} & 1 & 1 \\ 1 & 1 & \sqrt{3} & -1 \\ 1 & -1 & -1 & -\sqrt{3} \end{bmatrix}$$

is located with $\bar{r}_{1\text{-bit}}(4, 4; \{\mathbf{P}_{k,l}\}) = 0.1208$.

- For $M = 4$ and $N = 8$, with the constraint that $\mathbf{U}_1 = \mathbf{I}$ and $\mathbf{U}_2 = \mathbf{H}_8$,

$$\mathbf{U}_3 = \frac{1}{\sqrt{792}} \begin{bmatrix} \sqrt{29} & -\sqrt{29} & \sqrt{29} & \sqrt{29} & 13 & -13 & -13 & -13 \\ \sqrt{29} & -\sqrt{29} & -\sqrt{29} & -\sqrt{29} & 13 & 13 & 13 & -13 \\ \sqrt{29} & \sqrt{29} & \sqrt{29} & -\sqrt{29} & 13 & 13 & -13 & 13 \\ \sqrt{29} & \sqrt{29} & -\sqrt{29} & \sqrt{29} & 13 & -13 & 13 & 13 \\ 13 & -13 & -13 & -13 & -\sqrt{29} & -\sqrt{29} & -\sqrt{29} & \sqrt{29} \\ 13 & 13 & -13 & 13 & -\sqrt{29} & \sqrt{29} & -\sqrt{29} & -\sqrt{29} \\ 13 & 13 & 13 & -13 & -\sqrt{29} & -\sqrt{29} & \sqrt{29} & -\sqrt{29} \\ 13 & -13 & 13 & 13 & -\sqrt{29} & \sqrt{29} & \sqrt{29} & \sqrt{29} \end{bmatrix},$$

$$\mathbf{U}_4 = \frac{1}{\sqrt{792}} \begin{bmatrix} \sqrt{29} & 13 & \sqrt{29} & 13 & -13 & -\sqrt{29} & 13 & \sqrt{29} \\ 13 & -\sqrt{29} & 13 & -\sqrt{29} & \sqrt{29} & -13 & -\sqrt{29} & 13 \\ \sqrt{29} & -13 & -\sqrt{29} & 13 & -13 & -\sqrt{29} & -13 & -\sqrt{29} \\ 13 & \sqrt{29} & 13 & \sqrt{29} & \sqrt{29} & 13 & -\sqrt{29} & -13 \\ 13 & \sqrt{29} & -13 & -\sqrt{29} & \sqrt{29} & -13 & \sqrt{29} & -13 \\ \sqrt{29} & 13 & -\sqrt{29} & -13 & -13 & \sqrt{29} & -13 & \sqrt{29} \\ 13 & -\sqrt{29} & -13 & \sqrt{29} & \sqrt{29} & 13 & \sqrt{29} & 13 \\ \sqrt{29} & -13 & \sqrt{29} & -13 & -13 & \sqrt{29} & 13 & -\sqrt{29} \end{bmatrix}$$

are located with $\bar{r}_{1\text{-bit}}(8, 4; \{\mathbf{P}_{k,l}\}) = 0.0533$.

In general, although there is no guarantee of a global optimum, the performance with the output of this algorithm is clearly better than the performance with random matrices, and usually exhibits some structure.

Remark 7. Given any set $\{\mathbf{U}_k\}_{k=1}^M$, right-multiplying all matrices by an orthogonal fixed matrix $\mathbf{A}$ resulting in a set $\{\mathbf{U}_k\mathbf{A}\}_{k=1}^M$, yields an equivalent solution since the matrices $\mathbf{P}_{k,l}$ are invariant to this operation. Further, permuting the rows of any of the resulting matrices, or multiplying any row by -1 , also yields an equivalent solution performance-wise.

Remark 8. A permutation matrix (including the identity matrix) which has a single 1 entry in any row or column should be avoided since it does not satisfy the Lindeberg condition as explained in Section 2.3.3. Using such a matrix has two downsides. First, the marginal distribution of the error vector of each branch will not be (even approximately) Gaussian. Second, it results in the loss of robustness [29] when the source is not Gaussian. This problem may be resolved by right multiplying by an orthogonal matrix as explained in the previous remark, which should be chosen such that all matrices satisfy the Lindeberg condition.

Remark 9. The algorithm may be improved, utilizing more advanced techniques such as incorporating an adaptive step size for the geodesic flow algorithm as proposed by Abrudan et al. in [34].

Chapter 3

Summary

In this thesis we have developed a novel dithering scheme based on the application of orthogonal transformations before and after quantization. For fixed-rate low resolution quantization, the proposed scheme offers better performance than that attained with subtractive dithering. Further, the gain over the latter grows when considering multiple quantization branches (descriptions).

A numerical optimization procedure was further proposed for finding a set of orthogonal matrices with better performance than the average of either random ensemble. Furthermore, for the case of two quantization branches, it was demonstrated that the Hadamard matrix (for dimensions where such a matrix exists) yields optimal performance.

Implementation of the scheme using the randomized Hadamard ensemble has many advantages. The alphabet has a cardinality of two, and the fast Walsh-Hadamard transform may be leveraged for low computational complexity and storage. Further, the Lindeberg condition is always met for a large enough dimension, so that robustness (w.r.t. source distribution) is guaranteed. In practice, the random sequences of ± 1 entries that appear in the ensemble, may be replaced by adequate pseudo-random sequences found offline via numerical search.

An interesting topic for further investigation is to consider replacing the orthogonal matrix operations with all-pass filtering, in order to avoid block operations in a similar spirit to [35].

We also derived a simple inequality on the correlation of two functions operating on symmetric bivariate normal random variables, and described its possible application to the identification of a memoryless non-linearity.

Appendix A

Optimal Scalar Uniform Quantizer for a Gaussian Source

We optimize (according to the MSE distortion measure) a scalar uniform quantizer for a Gaussian random variable with zero mean and unit variance, i.e., $x \sim \mathcal{N}(0, 1)$. We do so for both undithered quantization as well as for SDQ with a uniform dither.

The quantizer is assumed to be midrise with uniform step size and with L levels, where L is an even integer,

$$q_L(x) = \frac{\Delta}{2} \cdot \text{sign}(x) \left(1 + 2 \sum_{i=1}^{\frac{L}{2}-1} 1_{\{i \cdot \Delta \leq |x|\}} \right), \quad (\text{A.1})$$

where $1_{\{\cdot\}}$ is the indicator function, and $\Delta \in \mathbb{R}_+$ is the step size. The same optimality conditions may similarly be established for midtread quantizers with an odd number of quantization levels.

A.1 Undithered Quantizer

As shown by Max [21], the optimal step size satisfies the condition,

$$\mathbb{E}\{(x - q_L(x))q_L(x)\} = \mathbb{E}\{xq_L(x)\} - \mathbb{E}\{q_L^2(x)\} = 0. \quad (\text{A.2})$$

The optimal step size $\Delta_{0,L} > 0$ may be found numerically by applying this condition. The term $\mathbb{E}\{xq_L(x)\}$, appearing in the condition above, is repeatedly used throughout

this work. It may be expressed in closed form as follows.

$$\begin{aligned}
\mathbb{E}\{xq_L(x)\} &= \int_{-\infty}^{\infty} x q_L(x) f_x(x) dx \\
&= 2 \int_0^{\infty} x q_L(x) f_x(x) dx \\
&= 2 \int_0^{\infty} x \cdot \frac{\Delta}{2} \cdot f_x(x) dx + 2 \sum_{i=1}^{\frac{L}{2}-1} \int_{i\Delta}^{\infty} x \cdot \Delta \cdot f_x(x) dx \\
&= \frac{\Delta}{2} \sqrt{\frac{2}{\pi}} \left(1 + 2 \sum_{i=1}^{\frac{L}{2}-1} e^{-\frac{\Delta^2}{2} i^2} \right), \tag{A.3}
\end{aligned}$$

where in the second equality, the symmetry of the integrand was used, and in the last equality we used

$$2 \int_a^{\infty} x \cdot f_x(x) dx = \sqrt{\frac{2}{\pi}} e^{-\frac{a^2}{2}} \quad \forall a \geq 0.$$

Substituting the optimal step size, the achieved distortion is

$$\begin{aligned}
D_{0,L} &= \mathbb{E}\{(x - q_L(x))^2\} \\
&= \mathbb{E}\{(x - q_L(x))x\} \\
&= 1 - \frac{\Delta_{0,L}}{\sqrt{2\pi}} \left(1 + 2 \sum_{i=1}^{\frac{L}{2}-1} e^{-\frac{\Delta_{0,L}^2}{2} i^2} \right), \tag{A.4}
\end{aligned}$$

where in the second equality the optimality condition (A.2) was used, and the last equality follows from (A.3). Table A.1 gives the optimal step sizes and achieved distortion for integer bit rates up to $R = 8$.

L	2	4	8	16	32	64	128	256
$\Delta_{0,L}$	$2\sqrt{\frac{2}{\pi}}$	0.9957	0.5860	0.3352	0.1881	0.1041	0.05687	0.03076
$D_{0,L}$	0.3634	0.1188	0.0374	0.0115	0.00349	0.00104	0.0003	0.000087

Table A.1: Optimal step sizes and distortions for a Gaussian source with zero mean and unit variance for $L = 2^R$, $R = 1, \dots, 8$.

A.2 Subtractive Dithered Quantizer

For SDQ, the MSE is given by

$$\begin{aligned}\mathbb{E}\{(x + u - q_L(x + u))^2\} &= \mathbb{E}\{(x + \Delta \cdot \tilde{u} - q_L(x + \Delta \cdot \tilde{u}))^2\} \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \left[2 \left(\sum_{i=1}^{\frac{L}{2}-1} \int_{(i-1-\tilde{u})\Delta}^{(i-\tilde{u})\Delta} \left(x + \Delta \cdot \tilde{u} - \frac{2i-1}{2}\Delta \right)^2 f_x(x) dx \right. \right. \\ &\quad \left. \left. + 2 \int_{(\frac{L}{2}-1-\tilde{u})\Delta}^{\infty} \left(x + \Delta \cdot \tilde{u} - \frac{L-1}{2}\Delta \right)^2 f_x(x) dx \right) \right] d\tilde{u},\end{aligned}$$

where $u = \Delta \cdot \tilde{u}$, and $\tilde{u}$ is uniformly distributed over the interval $[-\frac{1}{2}, \frac{1}{2}]$.

The derivative of the MSE w.r.t. Δ can be written, using the Leibniz integral rule

$$\frac{d}{dt} \left(\int_{a(t)}^{b(t)} g(x, t) dx \right) = \int_{a(t)}^{b(t)} \frac{\partial g(x, t)}{\partial t} dx + g(b(t), t) \cdot b'(t) - g(a(t), t) \cdot a'(t),$$

for the inner integrals w.r.t. x , as

$$\begin{aligned}& 2\mathbb{E} \left\{ (x + \Delta \cdot \tilde{u} - q_L(x + \Delta \cdot \tilde{u}))(\tilde{u} - \frac{1}{\Delta} q_L(x + \Delta \cdot \tilde{u})) \right\} \\ & + 2 \left(\frac{\Delta}{2} \right)^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\sum_{i=1}^{\frac{L}{2}-1} \left[(i - \tilde{u}) f_x((i - \tilde{u})\Delta) - (i - 1 - \tilde{u}) f_x((i - 1 - \tilde{u})\Delta) \right] \right. \\ & \quad \left. - \left(\frac{L}{2} - 1 - \tilde{u} \right) f_x \left(\left(\frac{L}{2} - 1 - \tilde{u} \right) \Delta \right) \right) d\tilde{u} \\ & = \frac{2}{\Delta} \mathbb{E} \left\{ (x + u - q_L(x + u))(u - q_L(x + u)) \right\} + 2 \left(\frac{\Delta}{2} \right)^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \tilde{u} f_x(-\Delta \cdot \tilde{u}) d\tilde{u},\end{aligned}$$

where in the explicit integral w.r.t $\tilde{u}$, the terms in the sum cancel each other, so that only the term appearing in last equality remains, and equals zero as an integral of an odd function.

Setting the derivative equal to zero results in the condition

$$\mathbb{E} \left\{ (x + u - q_L(x + u))(q_L(x + u) - u) \right\} = 0. \quad (\text{A.5})$$

The optimal step size $\Delta_{0,L}^{\text{SDQ}} > 0$ may be found using numerical search, such that condition (A.5) is satisfied. Table A.2 gives the optimal step sizes and the achieved distortion for integer bit rates up to $R = 8$.

L	2	4	8	16	32	64	128	256
$\Delta_{0,L}^{\text{SDQ}}$	1.5316	1.0121	0.5912	0.3363	0.1884	0.1042	0.0569	0.0308
$D_{0,L}^{\text{SDQ}}$	0.4438	0.1290	0.0385	0.0117	0.0035	0.00104	0.0003	0.000087

Table A.2: Optimal step sizes and distortions for SDQ of a Gaussian source with zero mean and unit variance, for $L = 2^R$, $R = 1, \dots, 8$.

Appendix B

Output Correlation Function Expansion

We compute the Taylor expansion of $\gamma_{L_k, L_l}(\rho)$ around the point $\rho = 0$. To that end, we will use Mehler's formula [8] for the bivariate Gaussian probability density function. Specifically, for a Gaussian vector

$$(z_1, z_2) \sim \mathcal{N}(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}), \quad (\text{B.1})$$

Mehler's formula states that

$$\begin{aligned} f_{z_1, z_2}(z_1, z_2) &= \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2} \frac{z_1^2 - 2\rho z_1 z_2 + z_2^2}{1-\rho^2}} \\ &= \frac{1}{2\pi} e^{-\frac{1}{2}(z_1^2 + z_2^2)} \sum_{n=0}^{\infty} \frac{1}{n!} He_n(z_1) He_n(z_2) \rho^n, \end{aligned} \quad (\text{B.2})$$

where $He_n(z)$ are the probabilists' Hermite polynomials defined as,

$$He_n(z) = (-1)^n e^{\frac{z^2}{2}} \frac{d^n}{dz^n} e^{-\frac{z^2}{2}}, \quad n \geq 0, \quad (\text{B.3})$$

with the recursion relation,

$$\begin{aligned} He_{n+1}(z) &= z He_n(z) - n He_{n-1}(z), \\ He_0(z) &= 1, \\ He_1(z) &= z, \end{aligned} \quad (\text{B.4})$$

resulting in $He_{2m}(z)$ and $He_{2m+1}(z)$ being even and odd functions of z , respectively, for $m \geq 0$.

Using (B.2), the Taylor expansion of $\gamma_{L_k, L_l}(\rho)$ is given by

$$\begin{aligned}\gamma_{L_k, L_l}(\rho) &= \kappa_{L_k} \kappa_{L_l} \mathbb{E}\{q_{L_k}(x'_k) q_{L_l}(x'_l) | \rho\} \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} a_{L_k, n} a_{L_l, n} \rho^n,\end{aligned}$$

where

$$a_{L_k, n} \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \kappa_{L_k} q_{L_k}(z) He_n(z) e^{-\frac{z^2}{2}} dz. \quad (\text{B.5})$$

Since $He_{2m}(z)$ and $q_{L_k}(z)$ are even and odd functions of z , respectively, $a_{L_k, 2m} = 0$ for all $m \geq 0$, as an integral of an odd function of z . For $n = 2m + 1$, using

$$\int_x^{\infty} He_n(z) e^{-\frac{z^2}{2}} dz = He_{n-1}(x) e^{-\frac{x^2}{2}},$$

and the explicit expression (A.1) for $q_{L_k}(z)$, we have

$$\begin{aligned}a_{L_k, 2m+1} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \kappa_{L_k} q_{L_k}(z) He_{2m+1}(z) e^{-\frac{z^2}{2}} dz \\ &= \kappa_{L_k} \frac{\Delta_{0, L_k}}{\sqrt{2\pi}} \left(\int_0^{\infty} He_{2m+1}(z) e^{-\frac{z^2}{2}} dz + 2 \sum_{i=1}^{\frac{L_k}{2}-1} \int_{i\Delta_{0, L_k}}^{\infty} He_{2m+1}(z) e^{-\frac{z^2}{2}} dz \right) \\ &= \kappa_{L_k} \frac{\Delta_{0, L_k}}{\sqrt{2\pi}} \left(He_{2m}(0) + 2 \sum_{i=1}^{\frac{L_k}{2}-1} He_{2m}(i\Delta_{0, L_k}) e^{-\frac{\Delta_{0, L_k}^2}{2} i^2} \right) \\ &= \frac{He_{2m}(0) + 2 \sum_{i=1}^{\frac{L_k}{2}-1} He_{2m}(i\Delta_{0, L_k}) e^{-\frac{\Delta_{0, L_k}^2}{2} i^2}}{1 + 2 \sum_{i=1}^{\frac{L_k}{2}-1} e^{-\frac{\Delta_{0, L_k}^2}{2} i^2}}, \quad (\text{B.6})\end{aligned}$$

so that $a_{L_k, 1} = 1$, for all $k = 1, \dots, M$.

It follows that for $|\rho| \leq 1$,

$$\begin{aligned}\gamma_{L_k, L_l}(\rho) &= \sum_{n=0}^{\infty} \frac{a_{L_k, 2n+1} a_{L_l, 2n+1}}{(2n+1)!} \rho^{2n+1} \\ &= \rho + \frac{1}{6} a_{L_k, 3} a_{L_l, 3} \rho^3 + o(\rho^3),\end{aligned}$$

where

$$a_{L,3} = \frac{2 \sum_{i=1}^{\frac{L}{2}-1} (i \Delta_{0,L})^2 e^{-\frac{\Delta_{0,L}^2}{2} i^2}}{2 \sum_{i=1}^{\frac{L}{2}-1} e^{-\frac{\Delta_{0,L}^2}{2} i^2} + 1} - 1, \quad (\text{B.7})$$

where we have used (B.6) with $He_2(z) = z^2 - 1$. Table 2.3 in Section 2.4 gives the values of $a_{L,3}$ for integer bit rates up to $R = 8$.

Appendix C

Properties of the Output Correlation Function

We prove the properties of $\gamma_{L_k, L_l}(\rho)$ as stated in Lemma 4:

1. For $\rho = \pm 1$, and for the symmetric case where $L_k = L_l$, we have $x'_l = \pm x'_k$, and hence

$$\begin{aligned}\gamma_{L_k, L_k}(1) &= -\gamma_{L_k, L_k}(-1) \\ &= \kappa_{L_k}^2 \mathbb{E}\{q_{L_k}^2(x'_k)\} \\ &= \kappa_{L_k},\end{aligned}$$

where the last equality follows from (2.22) and (A.2).

2. $\gamma_{L_k, L_l}(\rho)$ is an odd function, since its power series contains only odd powers of ρ . We find the derivative of $\gamma_{L_k, L_l}(\rho)$ to establish that is a monotonically increasing function of ρ . Indeed, using Price's theorem [2] yields

$$\dot{\gamma}_{L_k, L_l}(\rho) = \kappa_{L_k} \kappa_{L_l} \mathbb{E}\{\dot{q}_{L_k}(x'_k) \dot{q}_{L_l}(x'_l) | \rho\},$$

where $\dot{f}(x)$ denotes the derivative of $f(x)$ w.r.t. x . The derivative of a midrise quantizer as given in (A.1), may be expressed via a sum of Dirac delta functions

$$\dot{q}_{L_k}(z) = \Delta_{0, L_k} \left(\delta(z) + \sum_{i=1}^{\frac{L_k}{2}-1} [\delta(z - i\Delta_{0, L_k}) + \delta(z + i\Delta_{0, L_k})] \right).$$

Performing the expectation for (z_1, z_2) as specified in (B.1) results in a sum consist-

ing of samples of their joint PDF on a two-dimensional grid. Namely,

$$\begin{aligned} \dot{\gamma}_{L_k, L_l}(\rho) = \kappa_{L_k} \kappa_{L_l} \frac{\Delta_{0, L_k} \Delta_{0, L_l}}{2\pi} \frac{1}{\sqrt{1 - \rho^2}} & \left(1 + 2 \sum_{i=1}^{\frac{L_k}{2}-1} e^{-\frac{1}{2} \frac{\Delta_{0, L_k}^2}{1 - \rho^2} i^2} + 2 \sum_{j=1}^{\frac{L_l}{2}-1} e^{-\frac{1}{2} \frac{\Delta_{0, L_l}^2}{1 - \rho^2} j^2} \right. \\ & \left. + 4 \sum_{i=1}^{\frac{L_k}{2}-1} \sum_{j=1}^{\frac{L_l}{2}-1} e^{-\frac{1}{2} \frac{\Delta_{0, L_k}^2 i^2 + \Delta_{0, L_l}^2 j^2}{1 - \rho^2}} \cosh \left(\frac{\rho i \Delta_{0, L_k} j \Delta_{0, L_l}}{1 - \rho^2} \right) \right). \quad (\text{C.1}) \end{aligned}$$

Therefore, $\dot{\gamma}_{L_k, L_l}(\rho) > 0$ for all $|\rho| < 1$, since it is composed of sums and products of positive functions, so that $\gamma_{L_k, L_l}(\rho)$ is a strictly increasing function of ρ . Further, since $\gamma_{L_k, L_l}(\rho)$ is an odd function with $\gamma_{L_k, L_l}(0) = 0$, and since $\dot{\gamma}_{L_k, L_l}(\rho) > 0$, it follows that

$$\gamma_{L_k, L_l}(|\rho|) \geq 0, \quad (\text{C.2})$$

and

$$\gamma_{L_k, L_l}(|\rho|) = |\gamma_{L_k, L_l}(\rho)|. \quad (\text{C.3})$$

3. For the case of $L_k = L_l = 2$, from (C.1) we have that

$$\dot{\gamma}_{2,2}(\rho) = \frac{1}{\sqrt{1 - \rho^2}} \quad \forall |\rho| < 1,$$

where we used $\kappa_2 \cdot \Delta_{0,2} = \sqrt{2\pi}$, resulting in

$$\gamma_{2,2}(\rho) = \arcsin(\rho) \quad \forall |\rho| \leq 1. \quad (\text{C.4})$$

Appendix D

Properties of the Error Correlation Coefficient Function

We prove the properties of $\varphi_{L_k, L_l}(\rho)$ as stated in Corollary 2:

1. $\varphi_{L_k, L_l}(\rho)$ may be rewritten as

$$\begin{aligned}\varphi_{L_k, L_l}(\rho) &= \frac{\mathbb{E}\{(x'_k - \kappa_{L_k} q_{L_k}(x'_k))(x'_l - \kappa_{L_l} q_{L_l}(x'_l)) | \rho\}}{\sqrt{D_k^{\text{uc}}} \sqrt{D_l^{\text{uc}}}} \\ &= \frac{\kappa_{L_k} \kappa_{L_l} \mathbb{E}\{q_{L_k}(x'_k) q_{L_l}(x'_l) | \rho\} - \rho}{\sqrt{D_k^{\text{uc}}} \sqrt{D_l^{\text{uc}}}} \\ &= \frac{\gamma_{L_k, L_l}(\rho) - \rho}{\sqrt{D_k^{\text{uc}}} \sqrt{D_l^{\text{uc}}}},\end{aligned}$$

where the second equality follows since

$$\mathbb{E}\{x'_k \kappa_{L_l} q_{L_l}(x'_l) | \rho\} = \rho,$$

for all $k \neq l$. This may be derived by writing $x'_k = v + \rho x'_l$, where $v \sim \mathcal{N}(0, 1 - \rho^2)$, independent of x'_l , and using relation (2.22).

2. Substituting (2.29) into (2.31) results in the expansion

$$\varphi_{L_k, L_l}(\rho) = \sum_{n=1}^{\infty} \frac{1}{(2n+1)!} \frac{a_{L_k, 2n+1} a_{L_l, 2n+1}}{\sqrt{D_k^{\text{uc}}} \sqrt{D_l^{\text{uc}}}} \rho^{2n+1}.$$

3. For $\rho = \pm 1$, and the symmetric case where $L_k = L_l$, we have

$$\begin{aligned}\varphi_{L_k, L_k}(1) &= -\varphi_{L_k, L_k}(-1) \\ &= \frac{\gamma_{L_k, L_k}(1) - 1}{D_k^{\text{uc}}} \\ &= 1,\end{aligned}$$

where we used $D_k^{\text{uc}} = \kappa_{L_k} - 1$, and $\gamma_{L_k, L_k}(1) = \kappa_{L_k}$.

4. $\varphi_{L_k, L_l}(\rho)$ is an odd function as the difference of two odd functions, $\gamma_{L_k, L_l}(\rho)$ and ρ (up to a constant). Moreover, $\varphi_{L_k, L_k}(\rho)$ (symmetric case where $L_k = L_l$) is also a monotonically increasing function of ρ , since its expansion (2.32) consists only of positive coefficients.

5. Using (2.32) for the symmetric case $L_k = L_l$ yields,

$$\begin{aligned}|\varphi_{L_k, L_k}(\rho)| &= \left| \frac{1}{D_k^{\text{uc}}} \sum_{n=1}^{\infty} \frac{a_{L_k, 2n+1}^2}{(2n+1)!} \rho^{2n+1} \right| \\ &\leq \frac{1}{D_k^{\text{uc}}} \sum_{n=1}^{\infty} \frac{a_{L_k, 2n+1}^2}{(2n+1)!} |\rho|^{2n+1}\end{aligned}\tag{D.1}$$

$$\leq \left(\frac{1}{D_k^{\text{uc}}} \sum_{n=1}^{\infty} \frac{a_{L_k, 2n+1}^2}{(2n+1)!} \right) |\rho|^3\tag{D.2}$$

$$\begin{aligned}&= \varphi_{L_k, L_k}(1) |\rho|^3 \\ &= |\rho|^3,\end{aligned}\tag{D.3}$$

where (D.1) follows from the triangle inequality, (D.2) follows from the fact that $|\rho|^{2n+1} \leq |\rho|^3$ for all $|\rho| \leq 1$ and $n \geq 1$, and in the last equality we used the fact that $\varphi_{L_k, L_k}(1) = 1$ for all L_k . For the general case, it follows from Corollary 1 in Chapter 1 that

$$|\varphi_{L_k, L_l}(\rho)| \leq \sqrt{|\varphi_{L_k, L_k}(\rho)| |\varphi_{L_l, L_l}(\rho)|},$$

since $\varphi_{L_k, L_l}(\rho) = \mathbb{E} \{g_k(z_1)g_l(z_2)\}$, where $g_m(z) = (z - \kappa_{L_m} q_{L_m}(z)) / \sqrt{D_m^{\text{uc}}}$, $m = k, l$, are both odd functions, and z_1, z_2 are zero mean, unit variance Gaussian random variables with correlation ρ . Therefore, using (D.3), we have

$$\begin{aligned}|\varphi_{L_k, L_l}(\rho)| &\leq \sqrt{|\varphi_{L_k, L_k}(\rho)| |\varphi_{L_l, L_l}(\rho)|} \\ &\leq |\rho|^3.\end{aligned}\tag{D.4}$$

Appendix E

Properties of the Randomized Hadamard Matrix Ensemble

We find the distribution of the entries of $\mathbf{P}_{k,l} = \mathbf{U}_k \mathbf{U}_l^T$, where $\{\mathbf{U}_k\}_{k=1}^M$ are independent randomized Hadamard matrices for $N = 2^n$, as defined in Section 2.3.3. We have

$$\begin{aligned}\rho_{i,j}^{k,l} &= (\mathbf{P}_{k,l})_{i,j} \\ &= (\mathbf{H}_N \mathbf{B}_k \mathbf{B}_l^T \mathbf{H}_N^T)_{i,j} \\ &= \sum_{m=1}^N (\mathbf{H}_N)_{i,m} b_{k,m} (\mathbf{H}_N)_{j,m} b_{l,m}.\end{aligned}$$

Using the fact that the entries of the Hadamard matrix equal $\pm \frac{1}{\sqrt{N}}$, any entry of $\mathbf{P}_{k,l}$ is a sum of independent random variables taking the values $\pm \frac{1}{N}$ with equal probability. That is, we may express each entry, which with abuse of notation we now denote by ρ , as

$$\begin{aligned}\rho &= \frac{1}{N} \sum_{i=1}^N b'_i \\ &= \frac{2}{N} (s - \mathbb{E}\{s\}),\end{aligned}$$

where b'_i are i.i.d. and equal ± 1 with equal probability, and $s \sim B(N, 1/2)$ (binomially distributed) with $\mathbb{E}\{s\} = \frac{N}{2}$. Thus, the entries of $\mathbf{P}_{k,l}$ are centralized binomial random variables up to scaling, and are distributed according to,

$$\Pr \left\{ \rho = 2 \frac{n}{N} - 1 \right\} = \binom{N}{n} 2^{-N}, \quad n = 0, \dots, N,$$

where $\binom{N}{n} = \frac{N!}{(N-n)!n!}$ is the binomial coefficient. The second and fourth moments are given by

$$\begin{aligned}\mathbb{E}\{\rho^2\} &= \frac{1}{N}, \\ \mathbb{E}\{\rho^4\} &= \frac{3N-2}{N^3}.\end{aligned}\tag{E.1}$$

Appendix F

Hadamard Matrix of Order N is Optimal for $M = 2$ and $L_1 = L_2$

We consider the case of $M = 2$ branches with quantizers of equal rate $L_1 = L_2$. We show that any $\mathbf{U}_1$ and $\mathbf{U}_2$ with $\mathbf{P}_{1,2} = \mathbf{U}_1 \mathbf{U}_2^T = \mathbf{H}_N$, where N is a dimension for which a Hadamard matrix is known to exist, is a local minimum of the error correlation coefficient (2.44), which for the considered scenario can be written as

$$\begin{aligned} \bar{r}_{\text{symm}}(N, 2; \mathbf{P}_{1,2}) &= \frac{1}{N} \text{tr} (\mathbf{P}_{1,2}^T \varphi_{L_1, L_1} [\mathbf{P}_{1,2}]) \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \rho_{i,j}^{1,2} \varphi_{L_1, L_1} (\rho_{i,j}^{1,2}), \end{aligned} \quad (\text{F.1})$$

with the constraint that $\mathbf{P}_{1,2}$ is orthogonal, where $(\mathbf{P}_{1,2})_{i,j} = \rho_{i,j}^{1,2}$. The error correlation coefficient function for the symmetric case may be written using (2.32) as

$$\varphi_{L_1, L_1} [\mathbf{P}_{1,2}] = \sum_{n=1}^{\infty} \frac{a_{L_1, 2n+1}^2}{D_1^{\text{uc}}(2n+1)!} [\mathbf{P}_{1,2}]^{2n+1}, \quad (\text{F.2})$$

where $[\mathbf{P}_{1,2}]^n$ is an element-wise n th power of $\mathbf{P}_{1,2}$. Substituting (F.2) into (F.1) we obtain

$$\bar{r}_{\text{symm}}(N, 2; \mathbf{P}_{1,2}) = \frac{1}{N} \sum_{n=1}^{\infty} \frac{a_{L_1, 2n+1}^2}{D_1^{\text{uc}}(2n+1)!} \text{tr} (\mathbf{P}_{1,2}^T [\mathbf{P}_{1,2}]^{2n+1}).$$

We show that subject to the orthogonality constraint, the Hadamard matrix minimizes each element in the sum separately. In fact, we show that the Hadamard matrix minimizes each element in the sum subject to a weaker constraint, requiring only that all rows have unit norm. It follows that the Hadamard matrix also minimizes the sum since all the summands are positive.

To that end, write the constrained minimization using Lagrange multipliers as the minimization of

$$J = \text{tr}(\mathbf{P}_{1,2}^T [\mathbf{P}_{1,2}]^{2n+1}) - \text{tr}(\boldsymbol{\lambda} (\mathbf{P}_{1,2} \mathbf{P}_{1,2}^T - \mathbf{I})) \quad n \geq 1,$$

such that $\boldsymbol{\lambda} = \text{diag}([\lambda_1 \cdots \lambda_N])$ (optimization subject to normalized rows). Differentiating with respect to $\mathbf{P}_{1,2}$, we obtain

$$\frac{\partial J}{\partial \mathbf{P}_{1,2}} = (2n+2) [\mathbf{P}_{1,2}]^{2n+1} - 2\boldsymbol{\lambda} \mathbf{P}_{1,2}.$$

Setting the derivative to zero, we get

$$(n+1) (\rho_{i,j}^{1,2})^{2n+1} - \lambda_i \rho_{i,j}^{1,2} = 0,$$

with real solutions $\rho_{i,j}^{1,2} = 0, \pm \left(\frac{\lambda_i}{n+1}\right)^{\frac{1}{2n}}$ for $\lambda_i \geq 0$. Choosing the i th row to have N_i non-zero entries results in $\lambda_i = \frac{n+1}{N_i^n}$ after substituting in the constraint. Thus, the possible solutions are $\rho_{i,j}^{1,2} = 0, \pm \frac{1}{\sqrt{N_i}}$. The objective function's value at a local minimum is given by

$$\begin{aligned} \text{tr}(\mathbf{P}_{1,2}^T [\mathbf{P}_{1,2}]^{2n+1}) &= \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N (\rho_{i,j}^{1,2})^{2n+2} \\ &= \frac{1}{N} \sum_{i=1}^N \frac{1}{N_i^n}, \end{aligned}$$

and is minimized taking $N_i = N$ for all $i = 1, \dots, N$, thus yielding a global minimum. Choosing $\mathbf{P}_{1,2} = \mathbf{H}_N$ (assuming the dimension N is valid) satisfies this condition, since the entries of $\mathbf{H}_N$ equal $\pm \frac{1}{\sqrt{N}}$, and also the constraint is satisfied. The error correlation coefficient for $\mathbf{P}_{1,2} = \mathbf{H}_N$ with $\rho_{i,j}^{1,2} = \pm \frac{1}{\sqrt{N}}$ is given by

$$\begin{aligned} \bar{r}_{\text{symm}}(N, 2; \mathbf{H}_N) &= \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \rho_{i,j}^{1,2} \varphi_{L_1, L_1}(\rho_{i,j}^{1,2}) \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N |\rho_{i,j}^{1,2}| \varphi_{L_1, L_1}(|\rho_{i,j}^{1,2}|) \\ &= \sqrt{N} \varphi_{L_1, L_1} \left(\frac{1}{\sqrt{N}} \right), \end{aligned} \tag{F.3}$$

where the second equality follows since $\rho \varphi_{L_1, L_1}(\rho)$ is an even and positive function of ρ .

Bibliography

- [1] J. J. Bussgang, “Cross-correlation function of amplitude-distorted Gaussian signals,” Research Lab. of Electronics, M.I.T., Tech. Rep., 1952.
- [2] R. Price, “A useful theorem for nonlinear devices having Gaussian inputs,” *Information Theory, IRE Transactions on*, vol. 4, no. 2, pp. 69–72, 1958.
- [3] W. McGee, “Circularly complex Gaussian noise—a Price theorem and a Mehler expansion (corresp.),” *Information Theory, IEEE Transactions on*, vol. 15, no. 2, pp. 317–319, 1969.
- [4] H. O. Lancaster, “Some properties of the bivariate normal distribution considered in the form of a contingency table,” *Biometrika*, pp. 289–292, 1957.
- [5] H. Rowe, “Memoryless nonlinearities with Gaussian inputs: Elementary results,” *Bell System Technical Journal*, vol. 61, no. 7, pp. 1519–1526, 1982.
- [6] R. Hadad, U. Erez, and Y. Yu, “An inequality for the correlation of two functions operating on symmetric bivariate normal variables,” *Information Theory, IEEE Transactions on*, submitted July 2015.
- [7] F. G. Mehler, “Ueber die Entwicklung einer Function von beliebig vielen Variablen nach Laplaceschen Functionen höherer Ordnung.” *Journal für die reine und angewandte Mathematik*, vol. 66, pp. 161–176, 1866.
- [8] W. Kibble, “An extension of a theorem of Mehler’s on Hermite polynomials,” in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 41, no. 01, 1945, pp. 12–15.
- [9] G. Szegő, *Orthogonal Polynomials*. American Mathematical Soc., 1939, vol. 23.
- [10] Y. Yu, “On the maximal correlation coefficient,” *Statistics & Probability Letters*, vol. 78, no. 9, pp. 1072–1075, 2008.

- [11] J. Tsimbinos and K. V. Lever, “Nonlinear system compensation based on orthogonal polynomial inverses,” *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, vol. 48, no. 4, pp. 406–417, 2001.
- [12] L. Roberts, “Picture coding using pseudo-random noise,” *Information Theory, IRE Transactions on*, vol. 8, no. 2, pp. 145–154, 1962.
- [13] R. M. Gray and D. L. Neuhoff, “Quantization,” *Information Theory, IEEE Transactions on*, vol. 44, no. 6, pp. 2325–2383, 1998.
- [14] R. Hadad and U. Erez, “Dithered quantization via orthogonal transformations,” *Signal Processing, IEEE Transactions on*, submitted December 2015.
- [15] E. Akyol and K. Rose, “On constrained randomized quantization,” *Signal Processing, IEEE Transactions on*, vol. 61, no. 13, pp. 3291–3302.
- [16] R. Zamir, S. Shamai, and U. Erez, “Nested linear/lattice codes for structured multiterminal binning,” *Information Theory, IEEE Transactions on*, vol. 48, no. 6, pp. 1250–1276, 2002.
- [17] J. W. Lindeberg, “Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung,” *Mathematische Zeitschrift*, vol. 15, no. 1, pp. 211–225, 1922.
- [18] L. Schuchman, “Dither signals and their effect on quantization noise,” *Communication Technology, IEEE Transactions on*, vol. 12, no. 4, pp. 162–165, 1964.
- [19] J. M. Cioffi, G. P. Dudevoir, V. M. Eyuboglu, and D. G. Forney Jr, “MMSE decision-feedback equalizers and coding. I. Equalization results,” *Communications, IEEE Transactions on*, vol. 43, no. 10, pp. 2582–2594, 1995.
- [20] T. M. Cover and J. A. Thomas, “Elements of information theory 2nd edition,” 2006.
- [21] J. Max, “Quantizing for minimum distortion,” *Information Theory, IRE Transactions on*, vol. 6, no. 1, pp. 7–12, 1960.
- [22] V. K. Goyal, “Multiple description coding: Compression meets the network,” *Signal Processing Magazine, IEEE*, vol. 18, no. 5, pp. 74–93, 2001.
- [23] L. Ozarow, “On a source-coding problem with two channels and three receivers,” *Bell System Technical Journal, The*, vol. 59, no. 10, pp. 1909–1921, 1980.

- [24] A. E. Gamal and T. M. Cover, “Achievable rates for multiple descriptions,” *Information Theory, IEEE Transactions on*, vol. 28, no. 6, pp. 851–857, 1982.
- [25] V. A. Vaishampayan and J.-C. Batllo, “Asymptotic analysis of multiple description quantizers,” *Information Theory, IEEE Transactions on*, vol. 44, no. 1, pp. 278–284, 1998.
- [26] N. Jacobson, *Lie algebras*. Courier Corporation, 2013.
- [27] G. W. Stewart, “The efficient generation of random orthogonal matrices with an application to condition estimators,” *SIAM Journal on Numerical Analysis*, vol. 17, no. 3, pp. 403–409, 1980.
- [28] A. Stam, “Limit theorems for uniform distributions on spheres in high-dimensional Euclidean spaces,” *Journal of Applied Probability*, pp. 221–228, 1982.
- [29] A. C. Hung and T. H. Meng, “Multidimensional rotations for robust quantization of image data,” *Image Processing, IEEE Transactions on*, vol. 7, no. 1, pp. 1–12, 1998.
- [30] J. Chen, C. Tian, T. Berger, and S. S. Hemami, “Multiple description quantization via Gram–Schmidt orthogonalization,” *Information Theory, IEEE Transactions on*, vol. 52, no. 12, pp. 5197–5217, 2006.
- [31] Y. Nishimori, “Learning algorithm for independent component analysis by geodesic flows on orthogonal group,” in *International Joint Conference on Neural Networks, Washington DC*, 1999.
- [32] M. Plumbley, “Lie group methods for optimization with orthogonality constraints,” in *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation*, 2004.
- [33] D. A. Karpuk and C. Hollanti, “Rotating Non-Uniform and High-Dimensional Constellations Using Geodesic Flow on Lie Groups,” *CoRR*, vol. abs/1309.5247, 2013.
- [34] T. E. Abruđan, J. Eriksson, and V. Koivunen, “Steepest Descent Algorithms for Optimization Under Unitary Matrix Constraint.” *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1134–1147, 2008.
- [35] K. Popat and K. Zeger, “Robust quantization of memoryless sources using dispersive FIR filters.” *IEEE Transactions on Communications*, vol. 40, no. 11, pp. 1670–1674, 1992.