Dithered Quantization via Orthogonal Transformations

Ran Hadad and Uri Erez, *Member, IEEE*

***Abstract*—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., asymptotically vanishing correlation, as the dimension of the vector of samples processed jointly grows.**

***Index Terms*—Multidimensional signal processing, quantization, dither, diversity, multiple descriptions.**

I. Introduction

DITHERED quantization is a well-known technique to reduce the statistical dependence between a signal and its 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 [1]. Indeed, in *subtractive dithering* (depicted in Fig. 1), 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., [2].

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 [3]. In this context, the role of dithering is to reduce second order correlations. Such a criterion is relevant in applications where multiple quantized signals, each encoded separately (distributed encoders), are linearly combined.

An important such application is the multiple description coding problem (see, e.g., [4]), and its use in diversity-based transmission over fading channels [5]. In fact, the multiple description problem corresponds to the extreme case where all inputs are identical in Fig. 1. The source may be coarsely reconstructed given only one available output branch, or from any combination of available branch outputs with improved performance.

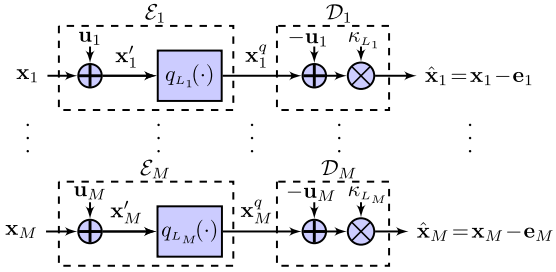

Fig. 1. A schematic overview of dithering using subtractive dithering.

Our interest will be in quantization of an independent, identically distributed (i.i.d.) Gaussian source. Indeed, transform coding (see, e.g., [6]) may be seamlessly combined with the proposed scheme to justify this assumption. We note that the purpose of transform coding is to reduce the correlation between different samples, whereas our goal is to reduce correlation between the samples and their corresponding quantization errors.[1]

As a baseline for comparison, we will consider subtractive dithered quantization (SDQ), schematically depicted in Fig. 1.[2,3]

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., [7]). 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 II-B) 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 Fig. 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

Manuscript received December 9, 2015; revised June 26, 2016; accepted July 24, 2016. Date of publication August 11, 2016; date of current version September 21, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ruixin Niu. The work of U. Erez was supported in part by the Israel Science Foundation under Grant 1956-15.

The authors are with the Department of Electrical Engineering-Systems, Tel-Aviv University, Tel-Aviv 69978, Israel (e-mail: ranhadad@post.tau.ac.il; uri@eng.tau.ac.il).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2016.2599482

[1] Indeed, we will require more than this when we consider multiple branches.

[2] The 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.

[3] The 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.

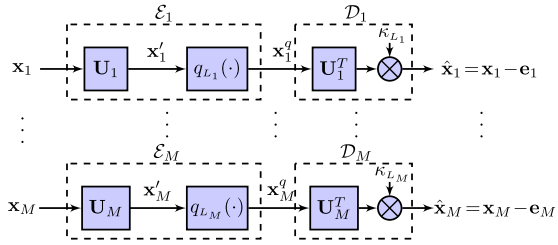

Fig. 2. A schematic overview of dithering using orthogonal transformations.

factor (κ_{L_1} in Fig. 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 II).

Another limitation of SDQ is that it requires using uniform quantizers and precludes source-optimized non-uniform quantization.[4] For the sake of easy comparison 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 Fig. 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.

We begin by analyzing the performance of the proposed scheme when drawing the transformations from the random (Haar) ensemble of orthogonal matrices and show that asymptotic second-order independence is achieved in the limit of large dimension. It is further shown that the latter goal may also be achieved, albeit with slightly slower convergence, by drawing the transformations from a much more structured (Hadamard based) ensemble of orthogonal matrices, allowing greatly reduced implementation complexity. The goal of finding by numerical means explicit (non-random) matrices with close to optimal performance is also pursued. It is demonstrated that, as is to be expected, for moderate dimensions the obtained matrices achieve significantly better performance (i.e., smaller correlation) than that of both random ensembles.

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 [8], see Section II-C3) is approximately Gaussian, whereas for SDQ it is uniform at high bit rates.

The organization of the paper is as follows. Section II establishes notation, provides background on SDQ, as well as recalls some basic definitions and properties of some ensembles of orthogonal transformations. Section III presents the proposed scheme. Section IV develops a bound on achievable performance, with an exact analysis for the symmetric 1-bit case. Section V introduces an algorithm for numerically designing (non-random) orthogonal matrices. Conclusions are given in Section VI. Technical proofs are deferred to the Appendix.

II. Notation and Background

A. 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 $\mathrm{col}_n(\mathbf{A})$. The trace of a square matrix $\mathbf{A}$ is denoted by $\mathrm{tr}(\mathbf{A})$. The ℓ_2-norm of a vector $\mathbf{a}$ is denoted by $\|\mathbf{a}\| \triangleq \sqrt{\sum_i a_i^2}$. We denote by $\mathrm{diag}(\mathbf{a})$, 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.

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 $k = 1, \ldots, M$. 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 $\mathrm{col}_n(\mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ for all $n = 1, \ldots, N$. Thus, $\mathbf{C}$ is the spatial covariance matrix with entries

$$c_{k,l} \triangleq (\mathbf{C})_{k,l}, \tag{1}$$

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

B. Subtractive Dithered Quantization

Consider M branches (realizations) of SDQ elements operating in parallel as depicted in Fig. 1. The k-th reconstruction

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

$(k=1,\ldots,M)$, denoted by $\hat{\mathbf{x}}_k$, is described by

$$\begin{aligned}\hat{\mathbf{x}}_k &= \kappa_{L_k}\left(\mathbf{x}_k^q - \mathbf{u}_k\right)\\ &= \kappa_{L_k}\left(q_{L_k}\left[\mathbf{x}_k'\right] - \mathbf{u}_k\right),\end{aligned} \tag{2}$$

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 III. The quantization error is given by

$$\begin{aligned}\mathbf{e}_k &= \mathbf{x}_k - \hat{\mathbf{x}}_k\\ &= \mathbf{x}_k + \kappa_{L_k}\left(\mathbf{u}_k - q_{L_k}\left[\mathbf{x}_k + \mathbf{u}_k\right]\right).\end{aligned} \tag{3}$$

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

$$\begin{aligned}D_k &= \frac{1}{N}\mathbb{E}\left\{\|\mathbf{x}_k' - \mathbf{x}_k^q\|^2\right\}\\ &= \mathbb{E}\left\{\left(x_k' - x_k^q\right)^2\right\}.\end{aligned} \tag{4}$$

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 $\left[-\frac{\Delta_k}{2}, \frac{\Delta_k}{2}\right]$. For this choice of dither, as demonstrated by Schuchman [9], 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.,

$$\mathbb{E}\{x_k e_k\} = 0, \tag{5}$$

by choosing the constant κ_{L_k} accordingly. The minimal distortion (MSE) attainable subject to this constraint is referred to as the optimal uncorrelated distortion. For a quantizer with optimal step size (see Appendix A) this constant is given by

$$\kappa_{L_k} = \frac{\sigma_k^2}{\sigma_k^2 - D_k}, \tag{6}$$

with corresponding uncorrelated distortion

$$\begin{aligned}D_k^{\mathrm{uc}} &= \mathbb{E}\left\{e_k^2\right\}\\ &= \frac{\sigma_k^2 D_k}{\sigma_k^2 - D_k}.\end{aligned} \tag{7}$$

The last equality is the well-known relation (see, e.g. [10]) between optimal correlated and uncorrelated distortion. The constant κ_{L_k} depends on the bit rate, and $\kappa_{L_k} \approx 1$ for high rates. Note that the dependence of D_k and D_k^{uc} on the number of quantization levels L_k is implicit throughout this paper.

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 [11], and is also computed in Appendix A for various bit rates. It is further shown in Section III that the relation between the correlated and uncorrelated distortion satisfies (7) in the case of undithered quantization as well.

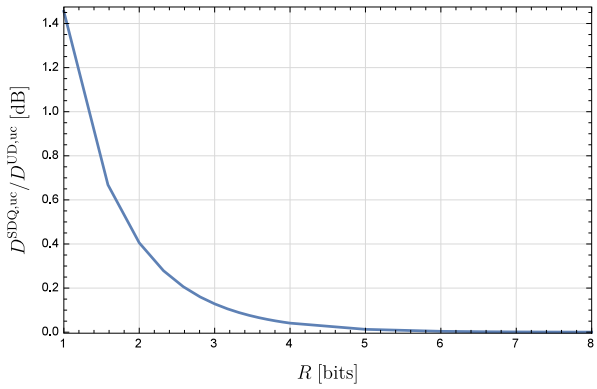

Fig. 3. The ratio $D^{\mathrm{SDQ,uc}}/D^{\mathrm{UD,uc}}$, as a function of the rate, $R = \log_2(L)$, for SDQ and undithered scalar quantization of a Gaussian source.

TABLE I
THE ERROR CORRELATION COEFFICIENTS $r^{\mathrm{SDQ}}(L)$ WITH $L = 2^R$ LEVELS, FOR $R = 1,\ldots,8$

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

Let $D^{\mathrm{SDQ,uc}}$ denote the uncorrelated distortion with SDQ. Similarly, let $D^{\mathrm{UD,uc}}$ denote the uncorrelated distortion without dithering. In Fig. 3, the ratio $D^{\mathrm{SDQ,uc}}/D^{\mathrm{UD,uc}}$ is 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).

When considering multiple SDQ branches, 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 a single SDQ branch w.r.t. undithered scalar quantization, as can be seen in Fig. 3. The second loss is due to the non-zero correlation between the errors of different branches.

To demonstrate these two sources of loss, we consider a central decoder receiving the output of M branches where the input $\mathbf{x}_k = \mathbf{x}$ is common to all. This may be seen as an instance of the multiple description problem (see, e.g., [4]). When considering further the case $D_k^{\mathrm{SDQ,uc}} = D^{\mathrm{SDQ,uc}}$ (or equivalently $L_k = L$) for all $k = 1,\ldots,M$, which is referred to as the *balanced* case, the error correlation coefficient between different SDQ branches is given by

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

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

Ideally, if the errors were uncorrelated (and also uncorrelated with the input), and there were no excess distortions due

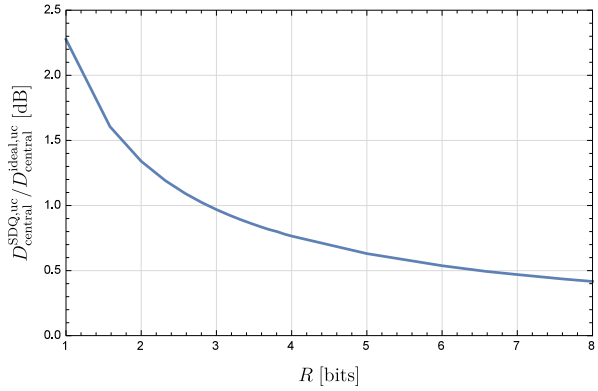

Fig. 4. The ratio between $D_{\text{central}}^{\text{SDQ,uc}}$ and $D_{\text{central}}^{\text{ideal,uc}}$ as a function of the bit rate $R = \log_2(L)$ for $M = 2$ branches, with uniform scalar quantization of a Gaussian source.

to dithering, simple averaging of the outputs (which is optimal since we are considering the balanced case) would achieve 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

$$D_{\text{central}}^{\text{SDQ,uc}}(L, M) = \frac{D^{\text{SDQ,uc}}(L)}{M}\left(1 + (M-1)\, r^{\text{SDQ}}(L)\right),$$

where the last equality follows from considering the balanced case. Moreover, for large M we have that

$$D_{\text{central}}^{\text{SDQ,uc}}(L, M) \underset{M\to\infty}{\longrightarrow} D^{\text{SDQ,uc}}(L) r^{\text{SDQ}}(L),$$

rather than approaching zero as in the case of zero correlation.

We turn our attention back to the case of $M = 2$ branches. In Fig. 4, 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.

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.

C. Orthogonal Matrix Preliminaries

The set of all $N \times N$ real orthogonal matrices is called the *orthogonal group*, $\mathrm{O}(N)$.[5] The set of $N \times N$ orthogonal matrices with determinant 1 is called the *special orthogonal group* $\mathrm{SO}(N)$. The sets $\mathrm{O}(N)$ and $\mathrm{SO}(N)$ are *compact Lie groups*. See [12] for further details.

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

$$\mathbf{MU} \overset{d}{=} \mathbf{UM} \overset{d}{=} \mathbf{U},$$

[5]All orthogonal matrices in this paper are assumed to have normalized column/row vectors.

where $\overset{d}{=}$ denotes equality in distribution (see [13] for efficient generation). It is well-known that if $\mathbf{U} \in \mathrm{O}(N)$ is uniformly distributed according to the Haar measure, then by the translation-invariance property, all of the entries of $\mathbf{U}$ are identically distributed.

The probability density function (PDF) of each entry of an $N \times N$ orthogonal Haar-distributed matrix is given by [14]:

$$f_u(u) = \begin{cases} \nu_N \left(1 - u^2\right)^{\frac{N-3}{2}}, & |u| \le 1 \\ 0, & else \end{cases}, \tag{8}$$

where

$$\nu_N = \frac{\Gamma(\frac{N}{2})}{\sqrt{\pi}\,\Gamma(\frac{N-1}{2})}, \tag{9}$$

for all $N \ge 2$, 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 $t \to \infty$. Therefore, asymptotically as $N \to \infty$,

$$\nu_N \sim \sqrt{\frac{N}{2\pi}}. \tag{10}$$

The second and fourth moments of each entry are given by,

$$\mathbb{E}\{u^2\} = \frac{1}{N},$$
$$\mathbb{E}\{u^4\} = \frac{3}{N\,(N+2)}. \tag{11}$$

2) Randomized Hadamard Matrix Ensemble: To reduce computational complexity, it may be more useful to utilize an ensemble based on Hadamard matrices. We further restrict attention to dimensions that are a 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,

$$\mathbf{U} = \mathbf{H}_N \mathbf{B}, \tag{12}$$

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

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

where $\mathbf{b}$ in turn is an i.i.d. vector with entries equal to ± 1 with probability $\frac{1}{2}$. We note that such an ensemble has been used in the literature on sparse coding, see, e.g., [15] and references therein.

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. 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 IV-B).

3) 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 $\hat{\mathbf{x}}_k$ are approximately Gaussian, and so are 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 [8].

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 section indeed satisfy Lindeberg's condition with high probability.

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. [16]).

III. Proposed Architecture and Basic Properties

A schematic overview of the system with M branches (realizations) operating in parallel is depicted in Fig. 2. Henceforth, we will assume without loss of generality that each of the M source vectors $\mathbf{x}_k$ $(k = 1, \ldots, M)$, 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 $|c_{k,l}| \leq 1$.[6]

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

$$\mathbf{x}'_k = \mathbf{U}_k \mathbf{x}_k, \tag{13}$$

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

$$\mathbf{x}^q_k = q_{L_k}\left[\mathbf{x}'_k\right].$$

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^q_k. 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

$$\begin{aligned} D_k &= \frac{1}{N}\mathbb{E}\left\{\|\mathbf{x}'_k - \mathbf{x}^q_k\|^2\right\} \\ &= \mathbb{E}\left\{(x'_k - x^q_k)^2\right\}. \end{aligned} \tag{14}$$

Note that we could equivalently define the quantizer to be as specified for SDQ in Section II-B, 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

$$\mathcal{Q}_{L_k} = \left\{ \tfrac{\Delta_{0,L_k}}{2}\left(2i + 1 - L_k\right) \middle| \, i = 0, \ldots, L_k - 1 \right\},$$

[6]We 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.

TABLE II
Values of κ_L for Optimal Uniform Scalar Quantization with $L = 2^R$ Levels, $R = 1, \ldots, 8$

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

where $\Delta_{0,L_k} \in \mathbb{R}_+$ is the optimal step size (as given in Appendix A). The uniform midrise quantizer is defined by

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

The decoder $\mathcal{D}_k$ computes

$$\hat{\mathbf{x}}_k = \kappa_{L_k} \mathbf{U}_k^T \mathbf{x}^q_k, \tag{15}$$

where $\kappa_{L_k} \in \mathbb{R}_+$. The constant κ_{L_k} is chosen so that the error vector

$$\begin{aligned} \mathbf{e}_k &= \mathbf{x}_k - \hat{\mathbf{x}}_k & (16)\\ &= \mathbf{U}_k^T\left(\mathbf{x}'_k - \kappa_{L_k}\mathbf{x}^q_k\right), & (17) \end{aligned}$$

is uncorrelated with its corresponding source vector $\mathbf{x}_k$. The quantities $\mathbf{x}_k^T \mathbf{e}_k$ and $\|\mathbf{e}_k\|^2$ are invariant to a transformation $\mathbf{U}_k \in \mathrm{O}(N)$ and completely determined by $\mathbf{x}_k$ (as seen from (13) and (17)). As shown in [3], for an optimal deterministic quantizer, the constraint that the input and error be uncorrelated is satisfied by setting

$$\begin{aligned} \kappa_{L_k} &= \frac{1}{\mathbb{E}\{{x^q_k}^2\}} & (18)\\ &= \frac{1}{1 - D_k}, & (19) \end{aligned}$$

and the resulting uncorrelated distortion is given by

$$\begin{aligned} D_k^{\mathrm{uc}} &\triangleq \frac{1}{N}\mathbb{E}\left\{\|\mathbf{e}_k\|^2\right\} \\ &= \frac{D_k}{1 - D_k}. \end{aligned} \tag{20}$$

Thus, a "second order additive" (uncorrelated) noise model is obtained for each branch. By applying (47) to (18), we also have that

$$\kappa_{L_k} = \frac{1}{\mathbb{E}\{x'_k x^q_k\}}, \tag{21}$$

Table II gives the values of κ_{L_k} for various bit rates.

A. 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

$$r_{k,l}(N) \triangleq \frac{\frac{1}{N}\mathbb{E}\{\mathbf{e}_k^T\,\mathbf{e}_l\}}{\sqrt{D_k^{\mathrm{uc}}}\sqrt{D_l^{\mathrm{uc}}}}, \tag{22}$$

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

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

IV. Performance Analysis

We wish to choose the orthogonal ensembles so as to minimize (22), which may be written as an expectation over the *conditional* error correlation coefficient defined by

$$r_{k,l}(N;\mathbf{U}_k,\mathbf{U}_l) \triangleq \frac{\frac{1}{N}\mathbb{E}\{\mathbf{e}_k^T\,\mathbf{e}_l|\mathbf{U}_k,\mathbf{U}_l\}}{\sqrt{D_k^{\rm uc}}\sqrt{D_l^{\rm uc}}}, \tag{23}$$

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

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

$$\gamma_{L_k,L_l}(\rho) \triangleq \kappa_{L_k}\kappa_{L_l}\mathbb{E}\{q_{L_k}(x_k')q_{L_l}(x_l')|\rho\},$$
$$(x_k',\ x_l')\,|\rho \sim \mathcal{N}\left(\mathbf{0},\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right), \tag{24}$$

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

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

$$\varphi_{L_k,L_l}(\rho) \triangleq \frac{\mathbb{E}\left\{\left(x_k' - \kappa_{L_k}q_{L_k}(x_k')\right)\left(x_l' - \kappa_{L_l}q_{L_l}(x_l')\right)|\rho\right\}}{\sqrt{D_k^{\rm uc}}\sqrt{D_l^{\rm uc}}},$$
$$(x_k',\ x_l')\,|\rho \sim \mathcal{N}\left(\mathbf{0},\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right), \tag{25}$$

where $k,l = 1,\ldots,M$, 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 1 (Output correlation function expansion): The output correlation function satisfies:

$$\gamma_{L_k,L_l}(\rho) = \rho + \sum_{n=1}^{\infty}\frac{a_{L_k,2n+1}a_{L_l,2n+1}}{(2n+1)!}\rho^{2n+1}, \tag{26}$$

for all $|\rho| \leq 1$ using Mehler's formula [17], where

$$a_{L_k,2n+1} \triangleq \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\kappa_{L_k}q_{L_k}(z)He_{2n+1}(z)e^{-\frac{z^2}{2}}dz, \tag{27}$$

$He_{2n+1}(z)$ being the probabilists' Hermite polynomials.

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

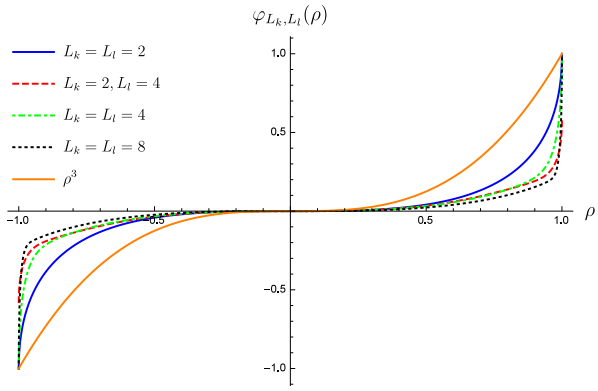

Fig. 5. Numerical evaluation of $\varphi_{L_k,L_l}(\rho)$ as defined in (25). For reference we plot the function ρ^3 that is used as a bound in Corollary 1.

Corollary 1 (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
$$\varphi_{L_k,L_l}(\rho) = \frac{\gamma_{L_k,L_l}(\rho) - \rho}{\sqrt{D_k^{\rm uc}}\sqrt{D_l^{\rm uc}}}. \tag{28}$$
2) The error correlation coefficient function satisfies:
$$\varphi_{L_k,L_l}(\rho) = \sum_{n=1}^{\infty}\frac{1}{(2n+1)!}\frac{a_{L_k,2n+1}a_{L_l,2n+1}}{\sqrt{D_k^{\rm uc}}\sqrt{D_l^{\rm uc}}}\rho^{2n+1}. \tag{29}$$
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,
$$\left|\varphi_{L_k,L_l}(\rho)\right| \leq |\rho|^3. \tag{30}$$

Fig. 5 depicts the error correlation coefficient function for a few pairs of bit rates, as well as ρ^3 for reference.

Remark 1: Numerical analysis of the error correlation coefficient function indicates that $|\varphi_{L_k,L_l}(\rho)| \leq |\varphi_{2,2}(\rho)|$, and also $\varphi_{L_k,L_l}(|\rho|) \geq 0$, as is evident from Fig. 5. 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, $|\rho||\varphi_{L_k,L_l}(\rho)| = \rho\varphi_{L_k,L_l}(\rho)$ as a product of odd functions. Finally, $\varphi_{L_k,L_l}(\rho)$ is a monotonically increasing function of ρ also when $L_k \neq L_l$. We conjecture that these properties hold and they will guide us in the numerical optimization procedure developed in Section V. We note however, that the bound on performance given in Proposition 1 derived below does not rely on this conjecture.

It is easy to verify that the conjecture in Remark 1 holds in the vicinity of the origin by inspecting the first summand in (29). That is from (29), the error correlation coefficient function satisfies[7]

$$\varphi_{L_k,L_l}(\rho) = \tilde{a}_{L_k,3}\tilde{a}_{L_l,3}\,\rho^3 + o(\rho^3)$$

[7]The notation $f(x) \in o\big(g(x)\big)$ means $\lim_{x\to 0}\frac{f(x)}{g(x)} = 0$.

TABLE III
THE COEFFICIENT $\tilde{a}_{L,3}$ GIVEN BY (31) FOR $L = 2^R$, $R = 1, \ldots, 8$

L	2	4	8	16	32	64	128	256
$\tilde{a}_{L,3}$	0.5403	0.5065	0.4217	0.3207	0.2274	0.1533	0.1000	0.0625

where

$$\tilde{a}_{L,3} = -\frac{a_{L,3}}{\sqrt{6 \cdot D^{\mathrm{uc}}(L)}}. \tag{31}$$

To evaluate the latter coefficient we compute $a_{L,3}$ using (27), and $D^{\mathrm{uc}}(L)$ using (20) and Table VI in Appendix A. In Table III, we provide values of $\tilde{a}_{L,3}$ for various bit rates. As can be seen, $\tilde{a}_{2,3}$ dominates $\tilde{a}_{L,3}$ for all other values of L in agreement with the conjecture.

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

$$\begin{aligned}
&\mathbb{E}\left\{\mathbf{e}_k^T\,\mathbf{e}_l|\mathbf{U}_k,\mathbf{U}_l\right\}\\
&= \mathbb{E}\left\{(\mathbf{x}_k-\hat{\mathbf{x}}_k)^T(\mathbf{x}_l-\hat{\mathbf{x}}_l)|\mathbf{U}_k,\mathbf{U}_l\right\}\\
&= \mathrm{tr}\big(\mathbb{E}\left\{(\mathbf{x}_k-\hat{\mathbf{x}}_k)(\mathbf{x}_l-\hat{\mathbf{x}}_l)^T|\mathbf{U}_k,\mathbf{U}_l\right\}\big)\\
&= \mathrm{tr}\big(\mathbb{E}\big\{\mathbf{U}_k^T(\mathbf{x}'_k-\kappa_{L_k}\mathbf{x}_k^q)(\mathbf{x}'_l-\kappa_{L_l}\mathbf{x}_l^q)^T\mathbf{U}_l|\mathbf{U}_k,\mathbf{U}_l\big\}\big) \quad (32)\\
&= \mathrm{tr}\big(\mathbf{U}_l\mathbf{U}_k^T\mathbb{E}\big\{(\mathbf{x}'_k-\kappa_{L_k}\mathbf{x}_k^q)(\mathbf{x}'_l-\kappa_{L_l}\mathbf{x}_l^q)^T|\mathbf{U}_k,\mathbf{U}_l\big\}\big)\\
&= \mathrm{tr}\big(\mathbf{P}_{k,l}^T\mathbb{E}\big\{\big(\mathbf{x}'_k-\kappa_{L_k}\mathbf{x}_k^q\big)\big(\mathbf{x}'_l-\kappa_{L_l}\mathbf{x}_l^q\big)^T|\mathbf{P}_{k,l}\big\}\big) \quad (33)\\
&= \sqrt{D_k^{\mathrm{uc}}}\sqrt{D_l^{\mathrm{uc}}}\,\mathrm{tr}\big(\mathbf{P}_{k,l}^T\varphi_{L_k,L_l}\left[c_{k,l}\mathbf{P}_{k,l}\right]\big), \quad (34)
\end{aligned}$$

where (32) follows from (17), (33) follows from the cyclic property of the trace, and in (34) the function $\varphi_{L_k,L_l}(\cdot)$ is used element-wise according to (25), since $\mathbf{x}_k^q = q_{L_k}[\mathbf{x}'_k]$, $\mathbf{x}_l^q = q_{L_l}[\mathbf{x}'_l]$, and $\mathbb{E}\{\mathbf{x}'_k\mathbf{x}'^T_l\} = c_{k,l}\mathbf{P}_{k,l}$, where $c_{k,l}$ is the spatial correlation between the sources as defined in (1).

We may now express the conditional error correlation coefficient as defined in (23) using $\varphi_{L_k,L_l}(\cdot)$ as,

$$\begin{aligned}
r_{k,l}(N;\mathbf{U}_k,\mathbf{U}_l) &= r_{k,l}(N;\mathbf{P}_{k,l})\\
&= \frac{1}{N}\,\mathrm{tr}\left(\mathbf{P}_{k,l}^T\varphi_{L_k,L_l}\left[c_{k,l}\mathbf{P}_{k,l}\right]\right) \quad (35)\\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{i,j}^{k,l}\varphi_{L_k,L_l}\big(c_{k,l}\rho_{i,j}^{k,l}\big).
\end{aligned}$$

Corollary 2: The error correlation coefficient satisfies

$$|r_{k,l}(N)| \le \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left\{\left(\rho_{i,j}^{k,l}\right)^4\right\}. \tag{36}$$

Proof: We obtain the following upper bound on the absolute value of the error correlation coefficient by writing it as an expectation over (35).

$$\begin{aligned}
|r_{k,l}(N)| &= |\mathbb{E}\{r_{k,l}(N;\mathbf{U}_k,\mathbf{U}_l)\}|\\
&= \left|\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left\{\rho_{i,j}^{k,l}\varphi_{L_k,L_l}\big(c_{k,l}\rho_{i,j}^{k,l}\big)\right\}\right|\\
&\le \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left\{\left(\rho_{i,j}^{k,l}\right)^4\right\}, \quad (37)
\end{aligned}$$

where the triangle inequality, property (30) in Corollary 1, and that $|c_{k,l}| \le 1$ were used in (37).

A. Bound on Performance for the Haar and Randomized Hadamard Ensembles

We now consider the performance for the two orthogonal ensembles defined in Section II-C. 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 (13), (15), and (16). Then the error correlation coefficient for the Haar distributed ensemble satisfies:

$$\left|r_{k,l}^{\mathrm{Haar}}(N)\right| \le \frac{3}{N+2} \quad \forall\, k \ne l.$$

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

$$\left|r_{k,l}^{\mathrm{rHad}}(N)\right| \le \frac{3N-2}{N^2} \quad \forall\, k \ne l.$$

In particular, the error correlation coefficient vanishes as $N \to \infty$ 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 (11) and (36) we have,

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

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

$$\Pr\left\{\rho = 2\frac{n}{N} - 1\right\} = \binom{N}{n}2^{-N}\ ,\ n = 0,\ldots,N. \tag{38}$$

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

$$\left|r_{k,l}^{\text{rHad}}(N)\right| \leq N\mathbb{E}\left\{\rho^4\right\} = \frac{3N-2}{N^2}.$$

Remark 2: We note that Proposition 1 is not limited to uniform quantization. It holds for the system described by (13), (15), and (16) as long as the scalar quantizers satisfy the constraint that the input and error are uncorrelated. This follows since the bound (30) only requires that the latter property is satisfied.

B. 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

$$r_{k,l}(N; L_k = 2, L_l = 2) = N\,\mathbb{E}\{\rho\,\varphi_{2,2}(c_{k,l}\rho)\}, \tag{39}$$

where by property 1 of Corollary 1, and property 2 of Lemma 2

$$\varphi_{2,2}(c_{k,l}\rho) = \frac{\arcsin(c_{k,l}\rho) - c_{k,l}\rho}{\pi/2 - 1}. \tag{40}$$

Note that $r_{k,l}(N)$ is a function of the spatial correlation $c_{k,l}$. In order to arrive at a simple objective function, we make use of the fact that (39) is maximized by setting $c_{k,l} = 1$. This follows by inspection of (40), which reveals that the function $\rho\,\varphi_{2,2}(c_{k,l}\rho)$ is a monotonically increasing function of $c_{k,l}$, for all values $|\rho| \leq 1$. We therefore define the worst-case correlation coefficient for 1-bit quantization as

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

1) Random Haar-Distributed Orthogonal Matrix Ensemble: Using the distribution of ρ as given in (8), by standard integration methods, we obtain

$$r_{\text{1-bit}}^{\text{Haar}}(N) = N\,\mathbb{E}\{\rho\,\varphi_{2,2}(\rho)\} = \frac{1}{\pi/2-1}\left(2\pi\nu_N^2\frac{N}{(N-1)^2} - 1\right), \tag{41}$$

where ν_N is defined in (9). Note that by (10), $r_{\text{1-bit}}^{\text{Haar}}(N)$ vanishes as $N \to \infty$ as expected.

2) Randomized Hadamard Matrix Ensemble: Using the distribution of ρ as given by (38), we compute the expression

$$r_{\text{1-bit}}^{\text{rHad}}(N) = N\mathbb{E}\left\{\rho\varphi_{2,2}\left(\rho\right)\right\} = \frac{N}{2^N}\sum_{n=0}^{N}\binom{N}{n}\left(2\frac{n}{N}-1\right)\varphi_{2,2}\left(2\frac{n}{N}-1\right). \tag{42}$$

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 IV.

TABLE IV
THE EXACT VALUES OF $r_{\text{1-bit}}(N)$ USING THE HAAR AND RANDOMIZED HADAMARD ENSEMBLES FOR $N = 2^n, n = 1, \ldots, 8$

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

C. 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 (30)), while being simple, is however quite loose, as can be inferred from Fig. 5. 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 (37), 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 1 above). Once this is done, one arrives at the tighter (conjectured) bound $|r_{k,l}(N)| \leq r_{\text{1-bit}}(N)$, where $r_{\text{1-bit}}(N)$ is given by (41) for the Haar ensemble, and by (42) for the randomized Hadamard ensemble. The conjectured bound appears (based on numerical evidence) to hold also for optimal non-uniform scalar quantization.

V. NUMERICAL DESIGN OF ORTHOGONAL TRANSFORMATIONS

We would like to find a set of non-random orthogonal matrices $\{\mathbf{U}_1, \ldots, \mathbf{U}_M\}$ that achieve error correlation coefficients as close as possible to zero. As an aid to design, we consider a single Gaussian source $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ giving rise to M branches as described in Section III for the case $c_{k,l} = 1$ for all k, l. 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. Since in the considered setting the error correlation coefficients are non-negative (see Remark 1), 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

$$\bar{r}(N, M; \{\mathbf{P}_{k,l}\}) \triangleq \frac{\sum_{k=1}^{M}\sum_{l\neq k} r_{k,l}(N; \mathbf{P}_{k,l})}{M(M-1)} = \frac{\sum_{k=1}^{M}\sum_{l\neq k}\frac{1}{N}\,\text{tr}\left(\mathbf{P}_{k,l}^T\varphi_{L_k,L_l}\left[\mathbf{P}_{k,l}\right]\right)}{M(M-1)}, \tag{43}$$

where $\mathbf{P}_{k,l} = \mathbf{U}_k\mathbf{U}_l^T$ and the second equality follows by substituting (35) with $c_{k,l} = 1$.

A. 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 (43) is minimized. We pose this as the following

minimization problem,

$$\min_{\substack{\mathbf{U}_1,\ldots,\mathbf{U}_M \\ \mathbf{U}_k\mathbf{U}_k^T=\mathbf{I}\ \forall k}} \bar{r}(N,M;\{\mathbf{P}_{k,l}\}) = \min_{\substack{\mathbf{U}_1,\ldots,\mathbf{U}_M \\ \mathbf{U}_k\mathbf{U}_k^T=\mathbf{I}\ \forall k}} \frac{\sum_{k=1}^{M}\sum_{l\neq k}\frac{1}{N}\operatorname{tr}\left(\mathbf{P}_{k,l}^T\varphi_{L_k,L_l}\left[\mathbf{P}_{k,l}\right]\right)}{M(M-1)} \tag{44}$$

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 1, 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 $k=1,\ldots,M$. Thus, our objective is to find matrices $\{\mathbf{U}_1,\ldots,\mathbf{U}_M\}$ minimizing

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

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

$$h\left(\mathbf{U}_1,\ldots,\mathbf{U}_M\right)=\sum_{k=1}^{M}\sum_{l\neq k}\operatorname{tr}\left(\left(\mathbf{U}_k\mathbf{U}_l^T\right)^T\arcsin\left[\mathbf{U}_k\mathbf{U}_l^T\right]\right). \tag{45}$$

Finding orthogonal matrices (locally) minimizing (45) requires performing numerical search over the Lie group $\mathrm{SO}(N)$ defined in Section II-C. 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. For details on the method, see [18], [19] and [20].

1) Algorithm Description: The algorithm consists of the following steps.

1) Initialization: The matrices $\mathbf{U}_k^{(0)}$, $k=1,\ldots,M$, are Haar-distributed random orthogonal matrices generated independently. Set a constant step size μ.
2) Steps for a single matrix $\mathbf{U}_k^{(i)}$ in the ith iteration (in each iteration perform for $\mathbf{U}_k^{(i)}$, $k=1,\ldots,M$, in ascending order):
   a) Gradient computation:

$$\mathbf{\Gamma}_k^{(i)}=\frac{\partial h}{\partial\mathbf{U}_k}\left(\mathbf{U}_1^{(i+1)},\ldots,\mathbf{U}_{k-1}^{(i+1)},\mathbf{U}_k^{(i)},\ldots,\mathbf{U}_M^{(i)}\right)$$

$$\mathbf{G}_k^{(i)}=\mathbf{\Gamma}_k^{(i)}\mathbf{U}_k^{(i)^T}-\mathbf{U}_k^{(i)}\mathbf{\Gamma}_k^{(i)^T},$$

where for the objective function $h\left(\mathbf{U}_1,\ldots,\mathbf{U}_M\right)$ as given in (45),

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

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

   b) Update:

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

where $\exp(\mathbf{A})$ is the matrix exponential of $\mathbf{A}$.

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

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

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

TABLE V
THE ERROR CORRELATION COEFFICIENT FOR $M=2$ BRANCHES WITH $L_1=L_2=2$, WHEN $\mathbf{P}_{1,2}=\mathbf{H}_N$, FOR $N=2^n$, $n=1,\ldots,8$

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

B. Numerical Optimization Results

We were able to obtain very satisfying results for $M=2,3,4$ branches with dimension ranging up to $N=16$. Indeed, the set of transformations $\{\mathbf{U}_1,\ldots,\mathbf{U}_M\}$ that we obtained from the algorithm results in better performance than the expectation of both random ensembles, i.e., $\bar{r}_{\text{1-bit}}(N,M;\{\mathbf{P}_{k,l}\})\leq r_{\text{1-bit}}(N)$.

1) The Special Case of $M=2$ Branches and Appropriate Hadamard Dimensions: In Appendix F we show that 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. Thus, as we know the optimal solution, there is no need for numerical optimization. Nonetheless, precisely for this reason, this is an ideal case to test our algorithm. 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.[8] The corresponding expression for $\bar{r}(N,2;\mathbf{H}_N)$ (for quantizers of equal rate) is given by (74). Specifically, for $L_1=L_2=2$, substituting (28) in (74) we have

$$\bar{r}_{\text{1-bit}}(N,2;\mathbf{H}_N)=\frac{\sqrt{N}\arcsin(\frac{1}{\sqrt{N}})-1}{\pi/2-1},$$

as given in Table V. The algorithm indeed locates $\mathbf{U}_1,\mathbf{U}_2$ with $\mathbf{P}_{1,2}=\mathbf{H}_N$ for $N=2,4,8,12,16$.

2) General Number of Branches: 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

[8]We note that this structure is different from the randomized Hadamard ensemble studied in earlier sections.

small. This hints at the possibility of global optimality. For example: 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 $\bar{r}_{\text{1-bit}}(6, 2; \mathbf{P}_{1,2}) = 0.0644$. 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 $\bar{r}_{\text{1-bit}}(10, 2; \mathbf{P}_{1,2}) = 0.0342$, and $\bar{r}_{\text{1-bit}}(14, 2; \mathbf{P}_{1,2}) = 0.0233$ 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 $\bar{r}_{\text{1-bit}}(4, 3; \{\mathbf{P}_{k,l}\}) = \bar{r}_{\text{1-bit}}(4, 2; \mathbf{H}_4) = 0.0827$. Further examples of specific structured solutions may be found in [21]. 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 3: 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, yields an equivalent solution performance-wise.

Remark 4: 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 II-C3. 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 [16] 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 5: 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 [22].

VI. Conclusion

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 improved performance w.r.t. that attained with subtractive dithering. Further, the gain over the latter grows when considering multiple quantization branches. A further advantage is that it may equally be used with non-uniform quantizers, unlike subtractive dithered quantization.

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. It is interesting to prove the conjectured bound that motivates the objective function that we use for optimization in Section V.

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.

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 variance σ^2, i.e., $x \sim \mathcal{N}(0, \sigma^2)$. 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,

$$q_L(x) = \frac{\Delta}{2} \cdot \text{sign}(x)\left(1 + 2\sum_{i=1}^{\frac{L}{2}-1} 1_{\{i\cdot\Delta \le |x|\}}\right), \tag{46}$$

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.

1) Undithered Quantizer: Optimal Step Size: As shown by Max [11], the optimal step size satisfies the condition,

$$\mathbb{E}\{(x - q_L(x))q_L(x)\} = \mathbb{E}\{xq_L(x)\} - \mathbb{E}\{q_L^2(x)\} = 0. \tag{47}$$

The optimal step size $\Delta_{0,L} > 0$ may be found numerically by applying this condition.

Substituting the optimal step size, the achieved distortion is

$$\begin{aligned} D_{0,L} &= \mathbb{E}\{(x - q_L(x))^2\} \\ &= \sigma^2 - \mathbb{E}\{xq_L(x)\}, \end{aligned} \tag{48}$$

where the second equality follows from the optimality condition (47). Table VI gives the optimal step sizes and achieved distortions for $\sigma^2 = 1$ and integer bit rates up to $R = 8$.

TABLE VI
OPTIMAL STEP SIZES AND DISTORTIONS FOR A GAUSSIAN SOURCE WITH ZERO MEAN AND UNIT VARIANCE FOR $L = 2^R$, $R = 1, \ldots, 8$

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

2) Subtractive Dithered Quantizer: Optimal Step Size: For SDQ, the distortion (MSE) is given by

$$D_L^{\mathrm{SDQ}} = \mathbb{E}\{(x+u-q_L(x+u))^2\} \tag{49}$$
$$= \mathbb{E}\{(x+\Delta\cdot\tilde{u}-q_L(x+\Delta\cdot\tilde{u}))^2\}$$
$$= \int_{-\frac{1}{2}}^{\frac{1}{2}} \left[2\left(\sum_{i=1}^{\frac{L}{2}-1} \int_{(i-1-\tilde{u})\Delta}^{(i-\tilde{u})\Delta} \left(x+\Delta\cdot\tilde{u}-\frac{2i-1}{2}\Delta \right)^2 f_x(x)dx \right.\right.$$
$$\left.\left. +2\int_{(\frac{L}{2}-1-\tilde{u})\Delta}^{\infty} \left(x+\Delta\cdot\tilde{u}-\frac{L-1}{2}\Delta \right)^2 f_x(x)dx \right)\right] d\tilde{u},$$

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

The derivative of the MSE w.r.t. Δ can be written, using the Leibniz integral rule for the inner integrals w.r.t. x, as

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

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

$$\mathbb{E}\left\{(x+u-q_L(x+u))(u-q_L(x+u))\right\} = 0. \tag{50}$$

The optimal step size $\Delta_{0,L}^{\mathrm{SDQ}} > 0$ may be found using numerical search, such that condition (50) is satisfied. By imposing (50) in (49), the optimal distortion is given by

$$D_{0,L}^{\mathrm{SDQ}} = \mathbb{E}\left\{(x+u-q_L(x+u))\,x\right\}$$
$$= \sigma^2 - \mathbb{E}\left\{x q_L(x+u)\right\}, \tag{51}$$

where we used $\mathbb{E}\{x(x+u)\} = \mathbb{E}\{x^2\} = \sigma^2$ since x and u are independent and zero-mean. Table VII gives the optimal step sizes and the achieved distortions for $\sigma^2 = 1$ and integer bit rates up to $R = 8$.

TABLE VII
OPTIMAL STEP SIZES AND DISTORTIONS FOR SDQ OF A GAUSSIAN SOURCE WITH ZERO MEAN AND UNIT VARIANCE, FOR $L = 2^R$, $R = 1, \ldots, 8$

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

3) Subtractive Dithered Quantizer: Uncorrelated Distortion: To impose the condition that the input and error be uncorrelated, we scale the output by a constant κ_L to attain

$$\hat{x} = \kappa_L\left(q_L(x+u)-u\right) \tag{52}$$
$$= x - e,$$

where e is the uncorrelated error satisfying

$$\mathbb{E}\{xe\} = 0. \tag{53}$$

By substituting $e = x - \hat{x}$, condition (53) is equivalent to

$$\mathbb{E}\{x\hat{x}\} = \sigma^2, \tag{54}$$

since $\mathbb{E}\{x^2\} = \sigma^2$. The optimal step size that minimizes the distortion satisfies condition (50). Multiplying (50) by κ_L^2, and then substituting (52), we get the equivalent condition

$$\mathbb{E}\left\{(\hat{x}-\kappa_L x)\,\hat{x}\right\} = 0. \tag{55}$$

We therefore have that

$$\mathbb{E}\left\{\hat{x}^2\right\} = \kappa_L \mathbb{E}\left\{x\hat{x}\right\}$$
$$= \kappa_L\sigma^2, \tag{56}$$

where the last equality follows from (54).

Inserting (52) into (54), we obtain

$$\kappa_L = \frac{\sigma^2}{\mathbb{E}\{x q_L(x+u)\}}$$
$$= \frac{\sigma^2}{\sigma^2 - D_{0,L}^{\mathrm{SDQ}}}, \tag{57}$$

where the last equality follows from (51). The resulting uncorrelated distortion is given by

$$D_{0,L}^{\mathrm{SDQ,uc}} = \mathbb{E}\left\{e^2\right\}$$
$$= \mathbb{E}\left\{(x-\hat{x})^2\right\}$$
$$= \mathbb{E}\left\{\hat{x}^2\right\} - \sigma^2$$
$$= \frac{\sigma^2 D_{0,L}^{\mathrm{SDQ}}}{\sigma^2 - D_{0,L}^{\mathrm{SDQ}}}, \tag{58}$$

where the third equality follows from (54).

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 [17] for the bivariate Gaussian probability density function. Specifically,

for a Gaussian vector

$$(z_1, z_2) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \tag{59}$$

Mehler's formula states that

$$\begin{aligned} f_{z_1,z_2}(z_1,z_2) &= \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2}\frac{z_1^2-2\rho z_1 z_2+z_2^2}{1-\rho^2}} \\ &= \frac{1}{2\pi} e^{-\frac{1}{2}(z_1^2+z_2^2)} \sum_{n=0}^{\infty} \frac{1}{n!} He_n(z_1) He_n(z_2) \rho^n, \end{aligned} \tag{60}$$

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

$$He_n(z) = (-1)^n e^{\frac{z^2}{2}} \frac{d^n}{dz^n} e^{-\frac{z^2}{2}}, n \geq 0, \tag{61}$$

with the recursion relation,

$$\begin{aligned} He_{n+1}(z) &= z He_n(z) - n He_{n-1}(z), \\ He_0(z) &= 1, \\ He_1(z) &= z, \end{aligned} \tag{62}$$

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

Using (60), 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

$$a_{L_k,n} \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \kappa_{L_k} q_{L_k}(z) He_n(z) e^{-\frac{z^2}{2}} dz. \tag{63}$$

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 $n = 1$, substituting $He_1(z) = z$ into (63) we have from (21) that

$$\begin{aligned} a_{L_k,1} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \kappa_{L_k} q_{L_k}(z) z e^{-\frac{z^2}{2}} dz \\ &= \kappa_{L_k} \mathbb{E}\{z q_{L_k}(z)\} \\ &= 1 \end{aligned}$$

for all L_k. It follows that for $|\rho| \leq 1$,

$$\gamma_{L_k,L_l}(\rho) = \rho + \sum_{n=1}^{\infty} \frac{a_{L_k,2n+1} a_{L_l,2n+1}}{(2n+1)!} \rho^{2n+1}.$$

C. Properties of the Output Correlation Function

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

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 (18).

2) For the case of $L_k = L_l = 2$, Price's theorem [23] yields

$$\dot{\gamma}_{2,2}(\rho) = \kappa_2^2 \mathbb{E}\{\dot{q}_2(x_k') \dot{q}_2(x_l') | \rho\},$$

where $\dot{f}(x)$ denotes the derivative of $f(x)$ w.r.t. x. The derivative of $q_2(z)$ as given in (46), is a scaled Dirac delta function

$$\dot{q}_2(z) = \Delta_{0,2} \delta(z),$$

Performing the expectation for (z_1, z_2) as specified in (59) and substituting $\kappa_2 \cdot \Delta_{0,2} = \sqrt{2\pi}$ results in

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

so that

$$\gamma_{2,2}(\rho) = \arcsin(\rho) \quad \forall\, |\rho| \leq 1. \tag{64}$$

D. Properties of the Error Correlation Coefficient Function

We prove the properties of $\varphi_{L_k,L_l}(\rho)$ as stated in Corollary 1:

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

$$\begin{aligned} \varphi_{L_k,L_l}(\rho) &= \frac{\mathbb{E}\left\{\left(x_k' - \kappa_{L_k} q_{L_k}(x_k')\right)\left(x_l' - \kappa_{L_l} q_{L_l}(x_l')\right) | \rho\right\}}{\sqrt{D_k^{\mathrm{uc}}}\sqrt{D_l^{\mathrm{uc}}}} \\ &= \frac{\kappa_{L_k}\kappa_{L_l}\mathbb{E}\{q_{L_k}(x_k') q_{L_l}(x_l') | \rho\} - \rho}{\sqrt{D_k^{\mathrm{uc}}}\sqrt{D_l^{\mathrm{uc}}}} \\ &= \frac{\gamma_{L_k,L_l}(\rho) - \rho}{\sqrt{D_k^{\mathrm{uc}}}\sqrt{D_l^{\mathrm{uc}}}}, \end{aligned}$$

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 (21).

2) Substituting (26) into (28) results in the expansion

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

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

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

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

4) $\varphi_{L_k,L_l}(\rho)$ is an odd function, since its expansion (29) contains only odd powers of ρ. 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 consists only of positive coefficients.

5) Using (29) for the symmetric case $L_k = L_l$ yields,

$$|\varphi_{L_k,L_k}(\rho)| = \frac{1}{D_k^{\rm uc}} \sum_{n=1}^{\infty} \frac{a_{L_k,2n+1}^2}{(2n+1)!} |\rho|^{2n+1} \tag{65}$$

$$\leq \left(\frac{1}{D_k^{\rm uc}} \sum_{n=1}^{\infty} \frac{a_{L_k,2n+1}^2}{(2n+1)!} \right) |\rho|^3 \tag{66}$$

$$= \varphi_{L_k,L_k}(1)|\rho|^3$$

$$= |\rho|^3, \tag{67}$$

where (66) 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 $\varphi_{L_k,L_k}(1) = 1$ for all L_k. For the general case, we have

$$\begin{aligned} &|\varphi_{L_k,L_l}(\rho)| \\ &= \left| \sum_{n=1}^{\infty} \frac{1}{(2n+1)!} \frac{a_{L_k,2n+1} a_{L_l,2n+1}}{\sqrt{D_k^{\rm uc}} \sqrt{D_l^{\rm uc}}} \rho^{2n+1} \right| \\ &\leq \sum_{n=1}^{\infty} \frac{|a_{L_k,2n+1}||\rho|^{n+\frac{1}{2}}}{\sqrt{D_k^{\rm uc}(2n+1)!}} \frac{|a_{L_l,2n+1}||\rho|^{n+\frac{1}{2}}}{\sqrt{D_l^{\rm uc}(2n+1)!}} \qquad (68) \\ &\leq \sqrt{\sum_{n=1}^{\infty} \frac{a_{L_k,2n+1}^2 |\rho|^{2n+1}}{D_k^{\rm uc}(2n+1)!}} \sqrt{\sum_{n=1}^{\infty} \frac{a_{L_l,2n+1}^2 |\rho|^{2n+1}}{D_l^{\rm uc}(2n+1)!}} \qquad (69) \\ &= \sqrt{|\varphi_{L_k,L_k}(\rho)||\varphi_{L_l,L_l}(\rho)|}, \end{aligned}$$

where (68) follows from the triangle inequality, and (69) follows from the Cauchy–Schwarz inequality for sequences. Therefore, using (67), we have

$$\begin{aligned} |\varphi_{L_k,L_l}(\rho)| &\leq \sqrt{|\varphi_{L_k,L_k}(\rho)||\varphi_{L_l,L_l}(\rho)|} \\ &\leq |\rho|^3. \end{aligned} \tag{70}$$

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 II-C2. We have

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

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

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

where b_i' are i.i.d. and equal ± 1 with equal probability, and $s \sim B\left(N, \frac{1}{2}\right)$ (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,

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

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

$$\begin{aligned} \mathbb{E}\{\rho^2\} &= \frac{1}{N}, \\ \mathbb{E}\{\rho^4\} &= \frac{3N-2}{N^3}. \end{aligned} \tag{71}$$

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 (43), which for the considered scenario can be written as

$$\begin{aligned} \bar{r}_{\rm symm}(N, 2; \mathbf{P}_{1,2}) &= \frac{1}{N} \operatorname{tr}\left(\mathbf{P}_{1,2}^T \varphi_{L_1,L_1}[\mathbf{P}_{1,2}]\right) \qquad (72) \\ &= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \rho_{i,j}^{1,2} \varphi_{L_1,L_1}\left(\rho_{i,j}^{1,2}\right), \end{aligned}$$

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 (29) as

$$\varphi_{L_1,L_1}[\mathbf{P}_{1,2}] = \sum_{n=1}^{\infty} \frac{a_{L_1,2n+1}^2}{D_1^{\rm uc}(2n+1)!} [\mathbf{P}_{1,2}]^{2n+1}, \tag{73}$$

where $[\mathbf{P}_{1,2}]^n$ is an element-wise nth power of $\mathbf{P}_{1,2}$. Substituting (73) into (72) we obtain

$$\begin{aligned} &\bar{r}_{\rm symm}(N, 2; \mathbf{P}_{1,2}) \\ &\quad = \frac{1}{N} \sum_{n=1}^{\infty} \frac{a_{L_1,2n+1}^2}{D_1^{\rm uc}(2n+1)!} \operatorname{tr}\left(\mathbf{P}_{1,2}^T [\mathbf{P}_{1,2}]^{2n+1}\right). \end{aligned}$$

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

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

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

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

Setting the derivative to zero, we get

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

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 ith 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

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

and is minimized taking $N_i = N$ for all $i = 1, \ldots, N$, 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

$$\begin{aligned}\bar{r}_{\mathrm{symm}}(N, 2; \mathbf{H}_N) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{i,j}^{1,2}\varphi_{L_1,L_1}\left(\rho_{i,j}^{1,2}\right) \\ &= \sqrt{N}\varphi_{L_1,L_1}\left(\frac{1}{\sqrt{N}}\right), \qquad (74)\end{aligned}$$

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

REFERENCES

[1] L. Roberts, "Picture coding using pseudo-random noise," *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 145–154, 1962.

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

[3] E. Akyol and K. Rose, "On constrained randomized quantization," *IEEE Trans. Signal Process.*, vol. 61, no. 13, pp. 3291–3302, Jul. 2013.

[4] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74–93, Sep. 2001.

[5] S.-M. Yang and V. A. Vaishampayan, "Low-delay communication for rayleigh fading channels: An application of the multiple description quantizer," *IEEE Trans. Commun.*, vol. 43, no. 11, pp. 2771–2783, Nov. 1995.

[6] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. New York, NY, USA: Springer, 2012, vol. 159.

[7] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1250–1276, Jun. 2002.

[8] J. W. Lindeberg, "Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung," *Math. Z.*, vol. 15, no. 1, pp. 211–225, 1922.

[9] L. Schuchman, "Dither signals and their effect on quantization noise," *IEEE Trans. Commun. Technol.*, vol. 12, no. 4, pp. 162–165, Dec. 1964.

[10] J. M. Cioffi, G. P. Dudevoir, V. M. Eyuboglu, and D. G. Forney, Jr., "MMSE decision-feedback equalizers and coding. I. Equalization results," *IEEE Trans. Commun.*, vol. 43, no. 10, pp. 2582–2594, Oct. 1995.

[11] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inf. Theory*, vol. 6, no. 1, pp. 7–12, Mar. 1960.

[12] N. Jacobson, *Lie Algebras*. North Chelmsford, MA, USA: Courier Corporation, 2013.

[13] G. W. Stewart, "The efficient generation of random orthogonal matrices with an application to condition estimators," *SIAM J. Numer. Anal.*, vol. 17, no. 3, pp. 403–409, 1980.

[14] A. Stam, "Limit theorems for uniform distributions on spheres in high-dimensional Euclidean spaces," *J. Appl. Probab.*, vol. 19, pp. 221–228, 1982.

[15] E. Liberty, N. Ailon, and A. Singer, "Dense fast random projections and lean Walsh Transforms," *Discr. Comput. Geom.*, vol. 45, no. 1, pp. 34–44, 2011.

[16] A. C. Hung and T. H. Meng, "Multidimensional rotations for robust quantization of image data," *IEEE Trans. Image Process.*, vol. 7, no. 1, pp. 1–12, Jan. 1998.

[17] W. Kibble, "An extension of a theorem of Mehler's on Hermite polynomials," *Math. Proc. Camb. Phil. Soc.*, vol. 41, no. 01, pp. 12–15, 1945.

[18] Y. Nishimori, "Learning algorithm for independent component analysis by geodesic flows on orthogonal group," in *Proc. Int. Joint Conf. Neural Netw.*, 1999, vol. 2, pp. 933–938.

[19] M. Plumbley, "Lie group methods for optimization with orthogonality constraints," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Signal Separation*, 2004, pp. 1245–1252.

[20] D. A. Karpuk and C. Hollanti, "Rotating non-uniform and high-dimensional constellations using geodesic flow on lie groups," in *Proc. Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 5884-5889.

[21] R. Hadad, "Dithered quantization via orthogonal transformations and a Cauchy-Schwarz like inequality," Master's thesis, Dept. Elect. Eng., Tel Aviv Univ, Tel Aviv, Israel, 2016. [Online]. Available: http://www.eng.tau.ac.il/ uri/theses/hadad_msc.pdf

[22] T. E. Abrudan, J. Eriksson, and V. Koivunen, "Steepest Descent Algorithms for Optimization Under Unitary Matrix Constraint." *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1134–1147, Mar. 2008.

[23] R. Price, "A useful theorem for nonlinear devices having Gaussian inputs," *IRE Trans. Inf. Theory*, vol. 4, no. 2, pp. 69–72, Jun. 1958.

Ran Hadad was born in Rishon LeZion, Israel, on October 18, 1985. He received the B.Sc. (*cum laude*), and M.Sc. (*summa cum laude*) degrees in electrical engineering from Tel Aviv University, Tel Aviv, Israel, in 2013 and 2016, respectively. His research interests include signal processing and communications.

Uri Erez (M'09) was born in Tel Aviv, Israel, on October 27, 1971. He received the B.Sc. degree in mathematics and physics, and the M.Sc. and Ph.D. degrees in electrical engineering from Tel Aviv University, Tel Aviv, Israel, in 1996, 1999, and 2003, respectively. During 2003/2004, he was a Postdoctoral Associate at the Signals, Information, and Algorithms Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. Since 2005, he has been with the Department of Electrical Engineering-Systems,Tel-Aviv University. His research interests include the general areas of information theory and digital communication. From 2009 to 2011, he was an Associate Editor for Coding Techniques for the IEEE TRANSACTIONS ON INFORMATION THEORY.