# On Transform Coding with Dithered Quantizers [*]

Emrah Akyol and Kenneth Rose
Dept. of Electrical & Computer Engineering
University of California, Santa Barbara, CA 93106, USA
Email:{eakyol, rose}@ece.ucsb.edu

**Abstract**

This paper is concerned with optimal transform coding in conjunction with dithered quantization. While the optimal deterministic quantizer's error is uncorrelated with the reconstructed value, the dithered quantizer yields quantization errors that are correlated with the reconstruction but are white and independent of the source. These properties offer potential benefits, but also have implications on the optimization of the rest of the coder. We derive the optimal transform for consequent dithered quantization. For fixed rate coding, we show that the transform derived for dithered quantization is universally optimal (for all sources), unlike the conventional quantization case where optimality of the Karhunen-Loeve transform is guaranteed for Gaussian sources. Moreover, we establish variable rate coding optimality for Gaussian sources.

## 1 Introduction

Transform coding is a computationally attractive method of source coding and is widely used in audio, image and video compression. In the basic transform coding setting, an input vector is linearly transformed into a vector in the transform domain whose components (also called transform coefficients) are scalar-quantized. The decoder reconstructs the quantized coefficients and performs a linear (inverse) transformation to obtain an estimate of the source vector. The design goal is to find the optimal transform pairs and bit allocation to scalar quantizers, which minimize distortion. In general, transform coding underperforms optimal vector quantization due to space filling loss in scalar quantizers, even if the transform generates independent coefficients. However, due to its low complexity, transform coding is commonly employed in practical multimedia compression systems.

Transform coding of correlated Gaussian sources has been extensively studied. It is known that the Karhunen-Loeve transform (KLT) is the optimal transform for

[*] The work is supported in part by the NSF under grant CCF-0728986, the University of California MICRO program, Applied Signal Technology Inc., Cisco Systems Inc., Qualcomm Inc., and Sony Ericsson Inc.

1068-0314/09 $25.00 © 2009 IEEE
DOI 10.1109/DCC.2009.76

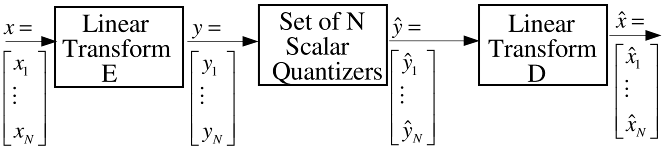

Figure 1: The general scheme of transform coding with scalar quantizers

Gaussian sources, in conjunction with optimal bit allocation and scalar quantization [1, 2]. The suboptimality of KLT has been demonstrated for several source distribution examples [3], and optimal transform coding for non-Gaussian sources remains a largely open problem. Derivations establishing the optimality of KLT for Gaussian sources typically involve the basic assumption on the quantization of transform coefficients, in particular, the quantization error is orthogonal to the coefficient reconstruction. While this assumption is valid for the optimal fixed rate quantization and also for high rate quantization theory approximations, it does not hold for dithered quantization.

Dithered (randomized) quantization[1] has useful properties such as producing white quantization noise independent of the source, and continuous reconstruction at the decoder [4]. A deterministic quantizer cannot produce source independent quantization noise in general or render the quantization error white if the source has memory. The properties of continuous output range and white quantization noise are known to be subjectively preferred by human audio-visual systems [4, 5]. Also, dithered quantization is used to find a lower bound in universal compression [6], since the quantization error is independent of the source distribution. In [7, 8, 9], many theoretical properties and extensions are studied for dithered lattice quantization based on a key result, namely, the optimal dithered lattice quantizer asymptotically yields Gaussian quantization noise thereby realizing the forward channel that achieves the rate distortion bound when applied to a Gaussian source.

We note again that in the case of deterministic optimal quantization, the optimal tranform is unknown for most distributions other than Gaussian [3]. A main premise of this work is that for fixed rate coding, dithered quantization enables universal transform coding, i.e., the optimality of the transform holds for all sources. (This is not the case for variable rate coding due to the dependence of the coding rate on the source distribution). Also, the quantization error is statistically orthogonal to the source and hence may be viewed as an additive independent noise term which in turn enables solving for the optimal transform by linear analysis (by solving a matrix equation) in terms of the second order statics of the source. This is not the case for optimal deterministic quantization, where difficulties include: i) the quantization error term can only be approximated as an additive uncorrelated noise at asymptotically high resolution; ii) the expected distortion depends on the type of distribution of each transform coefficient which, except in the simple Gaussian source case, depends non-trivially on the transform and makes it extremely challenging to minimize the distortion with respect to the transform. Because of these difficulties,

[1] Only subtractive dithering is considered in this paper

it is not straightforward to derive the optimal transform for non-Gaussian sources. The motivation for this work stems from the realization that dithered quantization holds considerable promise for circumventing the above difficulties and deriving the optimal transform for all sources.

# 2 Optimal Transform Derivation

## 2.1 Prior Work

The main scheme of transform coding is given in Figure-1. A jointly Gaussian vector $\mathbf{x}$ with covariance matrix $\mathbf{R_x}$ is first linearly transformed to obtain $\mathbf{y} = \mathbf{Ex}$, then quantized to get $\hat{\mathbf{y}} = \mathbf{Q}(\mathbf{y})$ which consists of scalar quantized samples. The quantization error vector is denoted as $\mathbf{n} = \mathbf{y} - \hat{\mathbf{y}}$. At the receiver side, a linear estimator is used to get an estimate of $\mathbf{x}$ as $\hat{\mathbf{x}} = \mathbf{D}\hat{\mathbf{y}}$ to minimize the mean square error,

$$J = E[(\mathbf{x} - \hat{\mathbf{x}})^{\mathbf{T}}(\mathbf{x} - \hat{\mathbf{x}})] = \mathbf{E}[\mathbf{Tr}((\mathbf{x} - \hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}})^{\mathbf{T}})] \tag{1}$$

Note that we include the “trace” formulation of the criterion as it will be convenient for matrix manipulations in the sequel. As is common, we assume for simplicity that the source is zero mean and that MSE is the distortion criterion. Under these assumptions, Huang and Schultheiss [10] first showed that the decoder should use the inverse of the transform used in the encoder (in Figure 1, $\mathbf{D} = \mathbf{E^{-1}}$ ). While this is a natural and intuitive choice, it is not as trivial as it seems due to the non-linear nature of quantization, and in fact depends on the optimality of the quantizers. The proof exploits the statistical properties of the quantization error of the optimal quantizer, specifically, the fact that the quantization error is uncorrelated with the output, i.e.,

$$E[\hat{\mathbf{y}}\mathbf{n}^{\mathbf{T}}] = \mathbf{0} \tag{2}$$

Next, they showed that the optimal choice for transform matrix $\mathbf{D}$ is the transpose of an orthogonal diagonalizing similarity transformation for $\mathbf{R_x}$, i.e., the KLT of the source, denoted as $\mathbf{S}$. In other words, $\mathbf{D} = \mathbf{S}$ satisfies

$$\mathbf{S}\mathbf{R_x}\mathbf{S}^{\mathbf{T}} = \mathbf{\Psi} \tag{3}$$

where $\mathbf{\Psi} = \mathbf{diag}(\lambda_{\mathbf{1}}, \lambda_{\mathbf{2}}, ..., \lambda_{\mathbf{N}})$ and $\lambda_i$'s are eigenvalues of $\mathbf{R_x}$. For convenience we will assume the ordering $\lambda_1 \geq \lambda_2 \geq, ..., \lambda_N$ with corresponding number of bits spent on coefficients $b_1 \geq b_2 \geq ... \geq b_N$. Note that, the optimality of KLT for the optimal fixed rate scalar quantizers does not require the high rate assumption. For a detailed tutorial discussion of KLT optimality see [1], which also covers optimality of KLT at high resolution in the case of variable rate (entropy) coding. More recent results [2] establish the optimality of KLT for Gaussian sources without high rate assumptions for both fixed and variable rate coding. However, that derivation minimizes the distortion in the transform domain, which is equivalent to minimizing it in the original domain if the transform is unitary. However, as will be shown shortly, in conjunction with dithered quantization the optimal transform is not unitary. It hence requires a method different from the one proposed in [2].

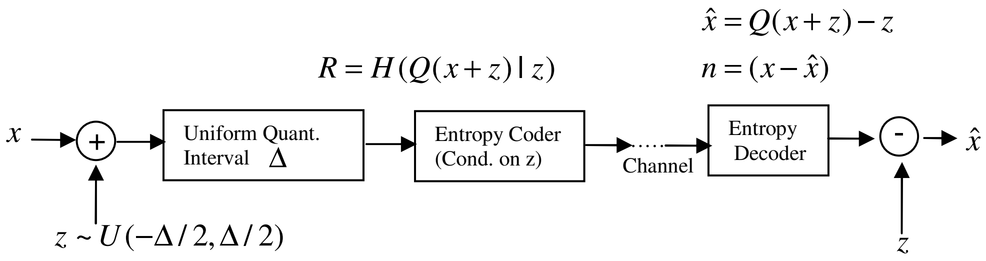

Figure 2: The general scheme of entropy coded dithered quantization

## 2.2 Dithered Quantization

The purpose of dithered quantization is to render the quantization error independent of the source, which can be achieved if certain conditions are met. Dithered quantization is performed in the framework where the quantizer is uniform $(-\Delta/2, \Delta/2)$ and the dither signal, $z$ is uniformly distributed on $(-\Delta/2, \Delta/2)$, matched to quantizer interval, as shown in Figure-2. A uniformly distributed dither signal, $z$, is added before quantization and the same dither signal is subtracted from the quantized value at the decoder (only "subtractive dithering" is considered in this paper). The quantized values are entropy coded, conditioned on the dither signal in the variable rate coding case. Note in this case, the rate can be approximated by the conditional entropy of the quantized values, $H(Q(x+z)|z)$, which is dependent on the source distribution. For fixed rate, both rate and distortion depend only on the second order statistics of the source. Randomized (dithered) quantizers have been studied in the past [4] due to their properties that differentiate them from deterministic quantizers, specifically: white quantization error that is independent of the source, $E[n^2] = \Delta^2/12$, and consequently have useful implications on universal compression bounds [6]. Note that source independent quantization noise cannot be achieved by a deterministic quantizer, but it is possible to enforce that the quantization error be uncorrelated with the source.

## 2.3 Simple Scalar Case

Some intuition is gained already from a simple scalar quantization setting. The dithered scalar quantizer is equivalent to the case where scalar source $x$ is corrupted by i.i.d (quantization) noise $n$, which is uncorrelated with $x$. At the receiver, $y = x+n$ is available and best linear estimate for $x$ is

$$\hat{x} = (\frac{\sigma_x^2}{\sigma_x^2 + \sigma_n^2})y \tag{4}$$

Note that an optimal deterministic quantizer would reconstruct $\hat{x} = y$ [1]. This simple observation of the scalar case already intuitively suggests that a unitary transform and specifically KLT will not be optimal for dithered quantization. Next, let us assume the source $x$ is Gaussian, and allow for scaling coefficients $\alpha$ before quantization and $\beta$ after quantization. Let also $f(b)$ be the distortion function of the dithered quantizer applied to unit variance, zero mean Gaussian at $b$ bits. Then, $\sigma_n^2 = \alpha^2 \sigma_x^2 f(b)$ and $\hat{x} = \beta(\alpha x + n)$. The optimal $\alpha$, $\beta$ will minimize $J$ where

$$\begin{array}{rcl} J & = & E[(x - \beta(\alpha x + n))^2] \\ & = & (1 - \beta\alpha)^2 \sigma_x^2 + \beta^2 \sigma_n^2 \\ & = & (1 - \beta\alpha)^2 \sigma_x^2 + \beta^2 \alpha^2 \sigma_x^2 f(b) \\ & = & (1 - 2\beta\alpha + (1 + f(b))(\beta\alpha)^2)\sigma_x^2 \end{array} \tag{5}$$

As expected, $J$ depends on the scaling coefficients only through the product $\beta\alpha$. By the optimality condition $\partial J / \partial(\beta\alpha) = 0$, we obtain the optimal scaling

$$\beta\alpha = 1/(1 + f(b)) \tag{6}$$

Generalizing to transform coding of signal blocks we intuitively expect that KLT followed by an appropriate diagonal scaling matrix would be optimal. The following section concretizes this intuition in a precise statement and formally proves it.

## 2.4 Optimal Transform for a Given Bit Allocation

Consider the problem: given bit allocation vector $\mathbf{b} = [b_1, b_2, ..., b_N]$ , find optimal $\mathbf{E}$ and $\mathbf{D}$ transform matrices to minimize the MSE. Without loss of generality we assume that the bit allocation vector is ordered, i.e., $b_1 \geq b_2 \geq ... \geq b_N$. From (1) the MSE cost can be written in trace form as

$$J = E[\mathbf{Tr}(\mathbf{x} - \mathbf{DEx} - \mathbf{Dn})(\mathbf{x} - \mathbf{DEx} - \mathbf{Dn})^{\mathbf{T}}] \tag{7}$$

Since we use a dithered quantizer, quantization error is uncorrelated with $\mathbf{y}$ and with $\mathbf{x}$, i.e., $E(\mathbf{x}\mathbf{n}^{\mathbf{T}}) = \mathbf{0}$. So we can write:

$$J = \mathbf{Tr}(\mathbf{DER_x E^T D^T} + \mathbf{R_x} + \mathbf{DR_n D^T} - \mathbf{2DER_x}) \tag{8}$$

As $\mathbf{R_x}$ does not depend on the transform, we may equivalently minimize

$$J_1 = \mathbf{Tr}(\mathbf{DER_x E^T D^T} + \mathbf{DR_n D^T} - \mathbf{2DER_x}) \tag{9}$$

Suppose there is a single function $f(.)$ to describe the rate distortion performance of the scalar dithered quantization of each transform coefficent through

$$E[(y_i - \hat{y}_i)^2] = \sigma_i^2 f(b_i) \tag{10}$$

where $b_i$ and $\sigma_i^2$ denote the number of bits allocated to coefficient $y_i$ and the variance of coefficient $y_i$ respectively, for $i = 1, 2, ..., N$. Note we can assume the form $f(b_i)$ for all transform coeficients irrespective of their distributions due to the basic properties of dithered quantization. Also, since the quantization error is independent of the source, we have

$$\mathbf{R_n} = \mathbf{diag}(\sigma_1^2 f(b_1), \sigma_2^2 f(b_2), ..., \sigma_N^2 f(b_N)) \tag{11}$$

Now, we define a convenient linear matrix operator, $\mathbf{d}(.)$ which sets to zero all off-diagonal entries of the argument matrix. Specifically, $\mathbf{d}(\mathbf{A}) = \mathbf{diag}(a_{11}, a_{22}, ...a_{NN})$ where $\mathbf{A}$ is some $N \times N$ matrix. Note that,

$$\mathbf{R_n} = \mathbf{d}(\mathbf{E}\mathbf{R_x}\mathbf{E^T}\boldsymbol{\Lambda}) \tag{12}$$

where $\boldsymbol{\Lambda} = \mathbf{diag}(f(b_1), f(b_2), ..., f(b_N))$. Also, it is straightforward to show (using matrix basic operations or the linear operator properties) that

$$\mathbf{d}(\mathbf{A}\boldsymbol{\Gamma}) = \boldsymbol{\Gamma}\mathbf{d}(\mathbf{A}) \tag{13}$$

for any diagonal $\boldsymbol{\Gamma}$ matrix. The following is a useful auxilary lemma.

**Lemma 1.** *For any arbitrary function matrix* $\mathbf{A}$*, variable matrix* $\mathbf{B}$ *and constant diagonal matrix* $\boldsymbol{\Gamma}$*,*

$$\partial(\mathbf{d}(\mathbf{A}\boldsymbol{\Gamma}))/\partial\mathbf{B} = \boldsymbol{\Gamma}\mathbf{d}(\partial\mathbf{A}/\partial\mathbf{B}) \tag{14}$$

*Proof.* Since both $\mathbf{d}(.)$ and differentiation are linear operators, they may be interchanged. Using (13) it is straightforward to obtain the lemma claim (14). □

Let $\mathbf{S}$ denote any unitary matrix that diagonalize $\mathbf{R_x}$ as defined in (3). We write $\mathbf{E} = \boldsymbol{\Phi}_\mathbf{1}\mathbf{S^T}$ and $\mathbf{D} = \mathbf{S}\boldsymbol{\Phi}_\mathbf{2}$ for any arbitrary $\boldsymbol{\Phi}_\mathbf{1}$, $\boldsymbol{\Phi}_\mathbf{2}$ matrices.

**Lemma 2.** *The optimal* $\boldsymbol{\Phi}_\mathbf{1}$ *and* $\boldsymbol{\Phi}_\mathbf{2}$ *matrices are diagonal.*

*Proof.* $\mathbf{DE} = \mathbf{S}\boldsymbol{\Phi}_\mathbf{2}\boldsymbol{\Phi}_\mathbf{1}\mathbf{S^T}$ and $\mathbf{E}\mathbf{R_x}\mathbf{E^T} = \boldsymbol{\Phi}_\mathbf{1}\boldsymbol{\Psi}\boldsymbol{\Phi}_\mathbf{1}^\mathbf{T}$. Substituting these expressions into (9) we obtain

$$\begin{aligned} J_1 &= \mathbf{Tr}(\boldsymbol{\Phi}_\mathbf{2}\boldsymbol{\Phi}_\mathbf{1}\boldsymbol{\Psi}\boldsymbol{\Phi}_\mathbf{1}^\mathbf{T}\boldsymbol{\Phi}_\mathbf{2}^\mathbf{T}) + \mathbf{Tr}(\boldsymbol{\Phi}_\mathbf{2}\mathbf{d}(\boldsymbol{\Lambda}\boldsymbol{\Phi}_\mathbf{1}\boldsymbol{\Psi}\boldsymbol{\Phi}_\mathbf{1}^\mathbf{T})\boldsymbol{\Phi}_\mathbf{2}^\mathbf{T}) \\ &- 2\mathbf{Tr}(\boldsymbol{\Phi}_\mathbf{2}\boldsymbol{\Phi}_\mathbf{1}\boldsymbol{\Psi}) \end{aligned} \tag{15}$$

Rearranging the terms using the trace equality $\mathbf{Tr}(\mathbf{AB}) = \mathbf{Tr}(\mathbf{BA})$,

$$\begin{aligned} J_1 &= \mathbf{Tr}(\boldsymbol{\Phi}_\mathbf{2}^\mathbf{T}\boldsymbol{\Phi}_\mathbf{2}(\boldsymbol{\Phi}_\mathbf{1}\boldsymbol{\Psi}\boldsymbol{\Phi}_\mathbf{1}^\mathbf{T} + \mathbf{d}(\boldsymbol{\Lambda}\boldsymbol{\Phi}_\mathbf{1}\boldsymbol{\Psi}\boldsymbol{\Phi}_\mathbf{1}^\mathbf{T}))) \\ &- 2\mathbf{Tr}(\boldsymbol{\Phi}_\mathbf{2}\boldsymbol{\Phi}_\mathbf{1}\boldsymbol{\Psi}) \end{aligned} \tag{16}$$

Setting $\partial J_1/\partial\boldsymbol{\Phi}_\mathbf{2} = \mathbf{0}$, yields

$$\mathbf{2}\boldsymbol{\Phi}_\mathbf{2}(\boldsymbol{\Phi}_\mathbf{1}\boldsymbol{\Psi}\boldsymbol{\Phi}_\mathbf{1}^\mathbf{T} + \mathbf{d}(\boldsymbol{\Lambda}\boldsymbol{\Phi}_\mathbf{1}\boldsymbol{\Psi}\boldsymbol{\Phi}_\mathbf{1}^\mathbf{T})) - \mathbf{2}\boldsymbol{\Psi}\boldsymbol{\Phi}_\mathbf{1}^\mathbf{T} = \mathbf{0} \tag{17}$$

Rearranging terms:

$$\boldsymbol{\Phi}_\mathbf{2} = \boldsymbol{\Psi}\boldsymbol{\Phi}_\mathbf{1}^\mathbf{T}(\boldsymbol{\Phi}_\mathbf{1}\boldsymbol{\Psi}\boldsymbol{\Phi}_\mathbf{1}^\mathbf{T} + \boldsymbol{\Lambda}\mathbf{d}(\boldsymbol{\Phi}_\mathbf{1}\boldsymbol{\Psi}\boldsymbol{\Phi}_\mathbf{1}^\mathbf{T}))^{\mathbf{-1}} \tag{18}$$

Setting $\partial J_1/\partial\boldsymbol{\Phi}_\mathbf{1} = \mathbf{0}$ and applying Lemma-1, we obtain

$$(\mathbf{2}\boldsymbol{\Psi}\boldsymbol{\Phi}_\mathbf{1}^\mathbf{T} + (\mathbf{2}\boldsymbol{\Lambda}\boldsymbol{\Psi}\mathbf{d}(\boldsymbol{\Phi}_\mathbf{1}))\boldsymbol{\Phi}_\mathbf{2}\boldsymbol{\Phi}_\mathbf{2}^\mathbf{T} - \mathbf{2}\boldsymbol{\Psi}\boldsymbol{\Phi}_\mathbf{2}^\mathbf{T} = \mathbf{0} \tag{19}$$

and hence

$$\boldsymbol{\Phi}_\mathbf{2} = (\boldsymbol{\Phi}_\mathbf{1} + \boldsymbol{\Lambda}\mathbf{d}(\boldsymbol{\Phi}_\mathbf{1}))^{\mathbf{-1}} \tag{20}$$

Note that we used the dithered quantization property that quantization noise is independent of the source (with Gaussian source assumption for the variable rate case) in this derivation, which implies $\partial\mathbf{\Lambda}/\partial\mathbf{\Phi_1} = \mathbf{0}$ and $\partial\mathbf{\Lambda}/\partial\mathbf{\Phi_2} = \mathbf{0}$. In conventional quantization, $\mathbf{\Lambda}$ depends on the distribution of the transform coefficient which is hard to track analytically. This point makes the solution difficult for non-Gaussian sources (note for a Gaussian source $y_i$'s are all Gaussian irrespective of the linear transform, so $\partial\mathbf{\Lambda}/\partial\mathbf{\Phi_1} = \mathbf{0}$ and $\partial\mathbf{\Lambda}/\partial\mathbf{\Phi_2} = \mathbf{0}$ hold.) Substituting (18) into (16) yields

$$J_1 = \mathbf{Tr}(\mathbf{\Phi_2^T}\mathbf{\Psi}\mathbf{\Phi_1^T}) - \mathbf{2Tr}(\mathbf{\Phi_2}\mathbf{\Phi_1}\mathbf{\Psi}) \tag{21}$$

Noting that $\mathbf{\Psi}$ is diagonal and using the trace equalities $\mathbf{Tr}(\mathbf{A}) = \mathbf{Tr}(\mathbf{A^T})$ and $\mathbf{Tr}(\mathbf{ABC}) = \mathbf{Tr}(\mathbf{CAB})$, we get $\mathbf{Tr}(\mathbf{\Phi_2^T}\mathbf{\Psi}\mathbf{\Phi_1^T}) = \mathbf{Tr}(\mathbf{\Phi_1}\mathbf{\Psi}\mathbf{\Phi_2}) = \mathbf{Tr}(\mathbf{\Phi_2}\mathbf{\Phi_1}\mathbf{\Psi})$ and hence

$$J_1 = -\mathbf{Tr}((\mathbf{\Phi_2}\mathbf{\Phi_1}\mathbf{\Psi}) \tag{22}$$

Plugging (20) into (22) we get

$$J_1 = -\mathbf{Tr}((\mathbf{\Phi_1} + \mathbf{\Lambda d}(\mathbf{\Phi_1}))^{\mathbf{-1}}\mathbf{\Phi_1}\mathbf{\Psi}) \tag{23}$$

Now, $J_1$ is a function of only $\mathbf{\Phi_1}$. Hence, setting the partial derivative with respect to $\mathbf{\Phi_1}$ to zero, $\partial J_1/\partial\mathbf{\Phi_1} = \mathbf{0}$ and using matrix inversion lemma [11]

$$\begin{array}{rcl}(\mathbf{I} + \mathbf{\Lambda})^{\mathbf{-1}} & = & \mathbf{\Phi_1}(\mathbf{\Phi_1} + \mathbf{\Lambda d}(\mathbf{\Phi_1}))^{\mathbf{-1}} \\ & = & \mathbf{I} - \mathbf{\Lambda d}(\mathbf{\Phi_1})(\mathbf{\Phi_1} + \mathbf{\Lambda d}(\mathbf{\Phi_1}))^{\mathbf{-1}}\end{array} \tag{24}$$

$(\mathbf{\Lambda} + \mathbf{I})^{\mathbf{-1}}$ is a diagonal matrix, since $\mathbf{\Lambda}$ and $\mathbf{I}$ are both diagonal. $\mathbf{\Lambda d}(\mathbf{\Phi_1})$ is also diagonal since it is the product of two diagonal matrices. The remaining factor $(\mathbf{\Phi_1} + \mathbf{\Lambda d}(\mathbf{\Phi_1}))^{\mathbf{-1}}$ must therefore be diagonal, which requires $\Phi_1$ to be diagonal, i.e., $\mathbf{\Phi_1} = \mathbf{d}(\mathbf{\Phi_1})$. Similar reasoning applied to (20) yields the conclusion that $\mathbf{\Phi_2}$ is also diagonal. □

Now, we can state the main theorem.

**Theorem.** *For given ordered bit allocations $b_1 \geq b_2 \geq, ..., b_N$, the* $\mathbf{E}$*,* $\mathbf{D}$ *transform matrices that minimize MSE for dithered scalar quantization are given by*

$$\mathbf{E} = \mathbf{\Phi_1}\mathbf{S^T}, \mathbf{D} = \mathbf{S}\mathbf{\Phi_2} \tag{25}$$

*for any* $\mathbf{\Phi_1}$*,* $\mathbf{\Phi_2}$ *diagonal matrices that satisfy*

$$\mathbf{\Phi_1}\mathbf{\Phi_2} = (\mathbf{I} + \mathbf{\Lambda})^{\mathbf{-1}} \tag{26}$$

*and* $\mathbf{S}$ *is the KLT matrix defined in Eq-3. Moreover, the distortion is*

$$J = \sum_{i=1}^{N} \lambda_i \frac{f(b_i)}{1 + f(b_i)} \tag{27}$$

*Proof.* Using Lemma-2 and $\mathbf{Tr}(\mathbf{\Gamma\Theta\Omega}) = \mathbf{Tr}(\mathbf{\Theta\Gamma\Omega})$ for any diagonal matrices $\mathbf{\Theta}$, $\mathbf{\Gamma}$, $\mathbf{\Omega}$, Eq-16 can be written as:

$$\mathbf{J} = \mathbf{Tr}((\mathbf{I} - \mathbf{\Phi_2\Phi_1})\mathbf{\Psi}) \tag{28}$$

MSE is only a function of $\mathbf{\Phi_1\Phi_2}$, not depending on individual values of $\mathbf{\Phi_1}$ or $\mathbf{\Phi_2}$. By (20) and Lemma-2 the optimality condition on $\mathbf{\Phi_1\Phi_2}$ follows directly: $\mathbf{\Phi_1\Phi_2} = (\mathbf{I} + \mathbf{\Lambda})^{-\mathbf{1}}$

There are possibly $N!$ distinct $\mathbf{S}$ matrices that satisfy (3) corresponding to $N!$ permutations of distinct eigenvalues of $\mathbf{R_x}$. To select the optimal $\mathbf{S}$ matrix, we need the optimal ordering of eigenvalues with respect to the ordering of bit allocations, as is standard practice with KLT and as is described, e.g., in [2]: higher rate should be allocated to the component that corresponds to larger eigenvalue. We are trying to minimize $J$, i.e., maximize $J_2$ where,

$$\begin{aligned} J_2 &= \mathbf{Tr}(\mathbf{\Phi_1\Phi_2\Psi}) \\ &= \mathbf{Tr}(\mathbf{\Psi}(\mathbf{I} + \mathbf{\Lambda})^{-\mathbf{1}}) \\ &= \sum_{i=1}^{N} \frac{\lambda_i}{1+f(b_i)} \end{aligned} \tag{29}$$

Both $\lambda_i$ and $1+f(b_i)$ terms are positive, the maximum is achived when $\lambda_i$ are in reverse order relative to $1 + f(b_i)$ [12]. Since $f(b_i)$ is a decreasing function of $b_i$, $\lambda_i$ should be ordered as is $b_i$, namely in decreasing order. Hence, the optimal permutation of the rows of $\mathbf{S}$ is the one that provides $(\lambda_1 \geq \lambda_2 \geq, ..., \lambda_N)$ when the bit allocation vector is ordered such that $(b_1 \geq b_2 \geq, ..., b_N)$ □

For the given $\mathbf{E}$, $\mathbf{D}$ matrices, the bit allocation should minimize MSE. If we write MSE in terms of quantization function and ordered eigenvalues using the main theorem,

$$J = \sum_{i=1}^{N} \lambda_i \frac{f(b_i)}{1 + f(b_i)} \tag{30}$$

Note that if standard KLT is used the distortion is

$$J_{KLT} = \sum_{i=1}^{N} \lambda_i f(b_i) \tag{31}$$

which is strictly larger than that of the proposed transform.

# 3 Discussion

We derived the optimal transform for subsequent dithered quantization. The optimal transform consists of KLT followed by a diagonal scaling matrix. For fixed rate coding, this transform is universally optimal (for all sources). In the case of variable rate coding, it is shown to be optimal for Gaussian sources. As future work, we will investigate the possible use of this transform in practical compression systems where the source distribution is not necessarily Gaussian.

# References

[1] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Springer, 1992.

[2] V. Goyal, J. Zhuang, and M. Vetterli, "Transform coding with backward adaptive updates," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1623–1633, 2000.

[3] M. Effros, H. Feng, and K. Zeger, "Suboptimality of the Karhunen-Loeve transform for transform coding," *IEEE Transactions on Information Theory*, vol. 50, no. 8, pp. 1605–1619, 2004.

[4] RM Gray and TG Stockham Jr, "Dithered quantizers," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 805–812, 1993.

[5] S.P. Lipshitz, R.A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *J. Audio Eng. Soc*, vol. 40, no. 5, pp. 355–375, 1992.

[6] J. Ziv, "On universal quantization," *IEEE Transactions on Information Theory*, vol. 31, no. 3, pp. 344–347, 1985.

[7] R. Zamir and M. Feder, "On universal quantization by randomized uniform/lattice quantizers," *IEEE Transactions on Information Theory*, vol. 38, no. 2 Part 2, pp. 428–436, 1992.

[8] R. Zamir and M. Feder, "Information rates of pre/post-filtered dithered quantizers," *IEEE Transactions on Information Theory*, vol. 42, no. 5, pp. 1340–1353, 1996.

[9] Y. Frank-Dayan and R. Zamir, "Dithered lattice-based quantizers for multiple descriptions," *IEEE Transactions on Information Theory*, vol. 48, no. 1, pp. 192–204, 2002.

[10] J. Huang and P. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Transactions on Communications*, vol. 11, no. 3, pp. 289–296, 1963.

[11] T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*, Prentice Hall Upper Saddle River, NJ, 2000.

[12] AW Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications* , Academic Press, New York, 1979.