Roundoff Noise Generated by Orthogonal Building Blocks in Signal Processing Structures

Vincent C. Liu and P. P. Vaidyanathan

Dept. of Electrical Engineering,

Caltech, Pasadena, CA 91125

Abstract: The statistics of roundoff errors produced by the multiplication of a random input vector with an orthogonal matrix can be expressed in terms of the characteristic function of the input vector. The case of a 2×2 orthogonal matrix is analyzed here by assuming that the input probability mass function follows a jointly Gaussian envelope. It is found that for small dynamic ranges the two errors are not necessarily uncorrelated. However, they are uncorrelated for large dynamic range.

I. Introduction

The roundoff error produced by the multiplication of a vector with orthogonal matrices is of interest due to presence of these orthogonal blocks in several digital filter structures. Fig. 1 shows the case of a 2×2 orthogonal matrix, where $(x_1, x_2)^T$ is the input vector quantized to B bits. The result of the matrix-vector multiplication, $(y_1, y_2)^T$, is then quantized to B bits also (which means the quantization level $\Delta = 2^{-B}$). The errors generated are defined to be $\epsilon_1 = \hat{y}_1 - y_1$ and $\epsilon_2 = \hat{y}_2 - y_2$. The type of quantization being considered in this paper is fixed point rounding. In the analysis of the roundoff errors, we shall make use of the results in [1].

We are interested in orthogonal matrices here, because they appear in several low-sensitivity implementations of digital filters, such as the Gray-Markel lattice [2], orthogonal filters [6], and the lossless FIR lattice structure [3]. Applications of these orthogonal matrices in multi-rate filter banks have also been reported recently [7], [8]. For example, consider the lossless FIR lattice as shown in Fig. 3. The location of the quantizers are exactly at the output of each 2×2 orthogonal matrix (denoted in the figure by R_i). According to common assumptions [4] [5], the noise generated at each quantizer is assumed to be uncorrelated to all the other noise sources. However since the entries of the 2×2 orthogonal matrix are related to each other as sines and cosines of the same angle θ, one might think that the roundoff errors would be correlated in some ways. In this paper, this issue is addressed in a quantitative way, and the answer is provided by simulation results.

II. The 2×2 Orthogonal Block

[1]Work supproted in part by the National Science Foundation grants DCI 8552579 and MIP 8604456

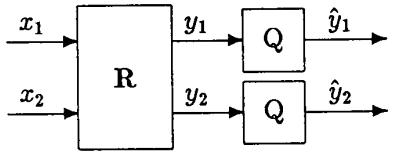

Figure 1: A 2×2 orthogonal building block with quantized outputs.

Consider the case of a signal passing through a multiplier followed by a quantizer with quantization level Δ, where $\Delta = 2^{-B}$, as shown in Fig 2. The error produced by the quantization process had been analyzed in [1]. The input u has a quantization level of Δ also. Let the multiplier value be N/L where N is an odd integer and L is a positive power of two. Thus the product y has a quantization level of $N\Delta/L$, which is then rounded off to B bits, producing $\hat{y}$. The quantization error ϵ (defined to be $\hat{y} - y$) can be expressed in terms of the input u as [1]

$$\epsilon = \sum_{k=0}^{L-1} \mathcal{E}_{L,N}(k) e^{j\frac{2\pi k}{L\Delta}u} \qquad (1)$$

The sequence $\mathcal{E}_{L,N}(k)$ is the set of Fourier coefficients of the periodic sequence $\{l(n)\}$ where $l(n)$ is the modulo L solution to the equation $l(n) + nN = 0$ Mod L.

In this paper, we shall apply (1) to the analysis of roundoff errors generated by orthogonal matrices. The case of the 2×2 orthogonal matrix is shown in Fig. 1. The vector $\mathbf{y} = (y_1, y_2)^T$ is related to the input by

$$\mathbf{y} = \mathbf{R}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} m_1 & m_2 \\ -m_2 & m_1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad (2)$$

The entries m_1 and m_2 are the quantized values of $\cos(\theta)$ and $\sin(\theta)$ repectively, where θ represents the angle of ro-

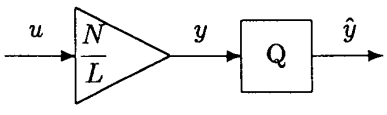

Figure 2: A single multiplier followed by a quantizer

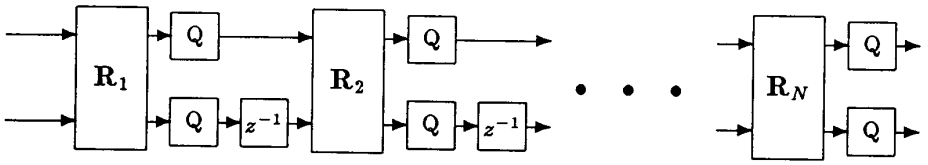

Figure 3: The lossless FIR lattice with quantizers

tation for the orthogonal matrix $\mathbf{R}$. Assume the level of quantization for these multipliers is $\Delta_1 = 2^{-B_1}$, then m_1 can be written as $M_1 M \Delta_1$ and m_2 as $M_2 M \Delta_1$ with M_1 and M_2 being relatively prime integers. Since x_1 and x_2 are quantized to B bits, let $x_1 = n_1\Delta$ and $x_2 = n_2\Delta$, then we get

$$\begin{aligned} y_1 &= (n_1 M_1 + n_2 M_2) M \Delta_1 \Delta \\ y_2 &= (-n_1 M_2 + n_2 M_1) M \Delta_1 \Delta \end{aligned} \tag{3}$$

Due to the fact that M_1 and M_2 are relatively prime, there exist integers n_1 and n_2 to make $n_1 M_1 + n_2 M_2 = 1$. This means y_1 has a quantization level of $M\Delta_1\Delta$. The same holds for y_2 . Both needs to be quantized back to a level of Δ, producing $\hat{y}_1$ and $\hat{y}_2$. Defining two numbers $u_1 = y_1/(M\Delta_1)$ and $u_2 = y_2/(M\Delta_1)$, it is clear that both u_1 and u_2 have B bits to the right of the binary point. The case in Fig. 2 becomes applicable here by making the identification of u_1 with u, y_1 with y and $M\Delta_1$ with the multiplier value $\frac{N}{L}$. As a result, (1) becomes

$$\epsilon_1 = \sum_{k=0}^{L-1} \mathcal{E}_{L,N}(k) e^{j\frac{2\pi k}{L\Delta}u_1} \tag{4}$$

Similarly, the expression for ϵ_2 is

$$\epsilon_2 = \sum_{k=0}^{L-1} \mathcal{E}_{L,N}(k) e^{j\frac{2\pi k}{L\Delta}u_2} \tag{5}$$

Assume that ϵ_1 and ϵ_2 have zero mean. From (4) and (5), expressions for the variances and cross-correlation between ϵ_1 and ϵ_2 can be found.

$$\begin{aligned} & E[\epsilon_1^2] \qquad (6) \\ &= E\left[\sum_{k_1,k_2=0}^{L-1} \mathcal{E}_{N,L}(k_1)\mathcal{E}_{N,L}(k_2) e^{j\frac{2\pi}{L\Delta}(k_1+k_2)u_1}\right] \\ &= \sum_{k_1,k_2=0}^{L-1} \mathcal{E}_{N,L}(k_1)\mathcal{E}_{N,L}(k_2) E\left[e^{j\frac{2\pi}{M\Delta_1 L\Delta}(k_1+k_2)y_1}\right] \\ &= \sum_{k_1,k_2=0}^{L-1} \mathcal{E}_{N,L}(k_1)\mathcal{E}_{N,L}(k_2) E\left[e^{j\frac{2\pi(k_1+k_2)}{L\Delta}(M_1x_1+M_2x_2)}\right] \end{aligned}$$

Defining the joint characteristic function of x_1 and x_2 as

$$\Phi(\omega_1,\omega_2) = E\left[e^{j(\omega_1 x_1 + \omega_2 x_2)}\right] \tag{7}$$

then (6) becomes

$$E[\epsilon_1^2] = \sum_{k_1=0}^{L-1}\sum_{k_2=0}^{L-1} \mathcal{E}_{N,L}(k_1)\mathcal{E}_{N,L}(k_2) \times \Phi\left(\frac{2\pi M_1(k_1+k_2)}{L\Delta}, \frac{2\pi M_2(k_1+k_2)}{L\Delta}\right). \tag{8}$$

Using (5), an expression for $E[\epsilon_2^2]$ can also be found

$$E[\epsilon_2^2] = \sum_{k_1=0}^{L-1}\sum_{k_2=0}^{L-1} \mathcal{E}_{N,L}(k_1)\mathcal{E}_{N,L}(k_2) \times \Phi\left(-\frac{2\pi M_2(k_1+k_2)}{L\Delta}, \frac{2\pi M_1(k_1+k_2)}{L\Delta}\right). \tag{9}$$

Similarly, the cross-correlation is

$$E[\epsilon_1\epsilon_2] = \sum_{k_1=0}^{L-1}\sum_{k_2=0}^{L-1} \mathcal{E}_{N,L}(k_1)\mathcal{E}_{N,L}(k_2) \times \Phi\left(\frac{2\pi(M_1k_1 - M_2k_2)}{L\Delta}, \frac{2\pi(M_2k_1 + M_1k_2)}{L\Delta}\right). \tag{10}$$

The common assumption being made about roundoff errors in digital filters is that they are uncorrelated with each other and the error is uncorrelated with the signal. For the case of the 2×2 orthogonal matrix, we will see how the cross-correlation of its errors behave in terms of the cross-correlation between the two inputs. Let the cross-correlation coefficient between ϵ_1 and ϵ_2 be defined as

$$\rho_{\epsilon_1,\epsilon_2} = \frac{E[\epsilon_1\epsilon_2]}{E[\epsilon_1^2]^{\frac{1}{2}} E[\epsilon_2^2]^{\frac{1}{2}}} \tag{11}$$

Obviously, $\rho_{\epsilon_1,\epsilon_2}$ will depend on the probability distribution of x_1 and x_2. However, it is not clear how $\rho_{\epsilon_1,\epsilon_2}$ is related to the correlation of the two inputs. If (11) is small enough for most inputs commonly encountered, then one may safely assume that ϵ_1 and ϵ_2 are uncorrelated. In order to calculate (11), assumptions need to be made about the joint probability distribution of x_1 and x_2. Let us assume that the joint probability density of x_1 and x_2 follows a Gaussian envelope,

$$f_{X_1,X_2}(x_1,x_2) = B \sum_{m,n=-\infty}^{\infty} \delta(x_1 - m\Delta, x_2 - n\Delta) \, e^{-(x_1^2 - 2\rho_0 x_1 x_2 + x_2^2)/2\sigma_0^2(1-\rho_0^2)} \tag{12}$$

The constant B is chosen such that the total probability is normalized to one. σ_0 is the variance of the Gaussian envelope. It is different from the variance of x_1 and of x_2. Similarly, ρ_0 is not the same as ρ_{x_1,x_2}.

For the input distribution given in (12), one can verify that the errors ϵ_1 and ϵ_2 have zero means. Furthermore, if the inputs are uncorrelated, i.e. $\rho_0 = 0$ in (12), then the errors are uncorrelated as well.

From (12), the characteristic function $\Phi(\omega_1, \omega_2)$ of x_1, x_2 can be found, and substituting it into (8) through (10) the variances and the cross-correlation of the two errors can be calculated. Note that for a given distribution function, such as the one in (11), the quantities $E[\epsilon_1^2]$, $E[\epsilon_2^2]$ and $E[\epsilon_1\epsilon_2]$ depends on the ratio σ_0/Δ. They do not depend on σ_0 and Δ individually. Therefore, we will use σ_0/Δ as a parameter for the input distribution.

With a quantization level of $\Delta = \Delta_1 = 2^{-8}$, we shall examine the behavior of $\rho_{\epsilon_1,\epsilon_2}$ as the rotational angle and the dynamic range vary. With a dynamic range of $(\sigma_0/\Delta) = 5$, the correlation coefficient $\rho_{\epsilon_1,\epsilon_2}$ is plotted in Fig. 4 as a funtion of θ for several values of ρ_0. As Fig. 4 shows, the cross-correlation between the two errors can be quite high for certain angles (such as $\theta = 36°$ and $38°$). Also, the correlation between the errors is large when the orthogonal matrix has small angles of rotation. Fig. 5 plots $\rho_{\epsilon_1,\epsilon_2}$ as a function of ρ_0 for small angles of θ. Here, the correlation of the errors becomes comparable to the cross-correlation of the inputs for $\theta \leq 3°$.

As we go to inputs with larger dynamic range the magnitude of $\rho_{\epsilon_1,\epsilon_2}$ tends to decrease, with the exception of a few particular values of θ. Fig. 6 represents the case of $\frac{\sigma_0}{\Delta} = 10$ and Fig. 7 shows the result for $\frac{\sigma_0}{\Delta} = 25$. Further numerical computation shows that $\rho_{\epsilon_1,\epsilon_2}$ goes to zero, as the dynamic range increases.

III. General Orthogonal Blocks

The expression in (4) and (5) can be generalized to any $K \times K$ orthogonal matrix. With $\mathbf{x} = (x_1, \ldots, x_K)^T$ as the input, the unquantized output is

$$\mathbf{y} = \mathbf{R}\mathbf{x} \tag{13}$$

Assuming each entry of $\mathbf{R}$ is quantized to B_1 bits, we can write $[\mathbf{R}]_{i_1,i_2} = M_{i_1,i_2}M\Delta_1$, where the $\gcd(M_{i_1,i_2}) = 1$. The quantization level of $\mathbf{y}$ is at least $M\Delta_1\Delta$. Let $M\Delta_1 = \frac{N}{L}$ as in (1), then following argument similar to the 2×2 case, the error ϵ_i can be written as

$$\epsilon_i = \sum_{k=0}^{L-1} \mathcal{E}_{L,N}(k) e^{j\frac{2\pi k}{L\Delta}\mathbf{x}^T\mathbf{m}_i}, \tag{14}$$

where $\mathbf{m}_i \triangleq (M_{i,1} \quad \ldots \quad M_{i,K})^T$. Defining the joint characteristic funtion of $\mathbf{x}$ to be $\Phi_K(\mathbf{w}) = E[e^{j\mathbf{w}^T\mathbf{x}}]$ with $\mathbf{w} = (\omega_1, \ldots, \omega_K)^T$, the error variance is given by

$$E[\epsilon_i^2] = \sum_{k_1=0}^{L-1}\sum_{k_2=0}^{L-1} \mathcal{E}_{L,N}(k_1)\mathcal{E}_{L,N}(k_2) \times \Phi_K\left(\frac{2\pi(k_1+k_2)}{L\Delta}\mathbf{m}_i\right) \tag{15}$$

Similar to (10), the cross-correlation between any two errors is

$$E[\epsilon_{i_1}\epsilon_{i_2}] = \sum_{k_1=0}^{L-1}\sum_{k_2=0}^{L-1} \mathcal{E}_{L,N}(k_1)\mathcal{E}_{L,N}(k_2) \times \Phi_K\left(\frac{2\pi}{L\Delta}(k_1\mathbf{m}_{i_1} + k_2\mathbf{m}_{i_2})\right) \tag{16}$$

IV. Summary

The roundoff errors generated by orthogonal matrix multiplications were analyzed. The variance for each error can be expressed in terms of the characteristic function of the input, and so can the cross-correlation between any pair of errors. For the 2×2 case, by assuming the input distribution to have a jointly Gaussian envelope, the cross-correlation between ϵ_1 and ϵ_2 is computed for various different input dynamic ranges. It is found that for most angles of rotation the cross-correlation between the two errors decreases rapidly as the dynamic range of the input goes up. However, for a few particular values of θ the cross-correlation between the errors remains high even for large dynamic range.

References

[1] C. W. Barnes, B. N. Tran and S. H. Leung, "On the Statistics of Fixed-Point Roundoff Error", IEEE Trans. ASSP, pp 595-606, June 1985.

[2] A. H. Gray, Jr. and J. D. Markel, "Digital lattice and ladder filter synthesis", IEEE Trans. AU, pp. 491-500, Dec. 1973.

[3] P. P. Vaidyanathan, "Passive Cascaded-Lattice Structures for Low-Sensitivity FIR Filter Design, with Applications to Filter Banks", IEEE Trans. CAS, pp. 1045-1064, Nov. 1986.

[4] L. B. Jackson, "On the interaction of roundoff noise and dynamic range in digital filters, Bell Systems Technical Journal, Vol. 49, pp. 159–184, 1970.

[5] C. T. Mullis and R. A. Roberts, "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters, IEEE Trans. CAS, pp. 551–562, Sept. 1976.

[6] R. A. Roberts and C. T. Mullis, *Digital Signal Processing*, Addison-Wesley Publishing Co., Inc., 1987.

[7] P. P. Vaidyanathan, "Theory and Design of M-Channel Maximally Decimated Quadrature Mirror Filters with Arbitrary M, Having the Perfect-Reconstruction Property", IEEE Trans. ASSP, pp. 476-492, April, 1987.

[8] P. P. Vaidyanathan and P.-Q. Hoang, "Lattice Structures for Optimal Design and Robust Implementation of Two-Channel Perfect-Reconstruction QMF Banks" IEEE Trans. ASSP, pp. 81-94, Jan. 1988.

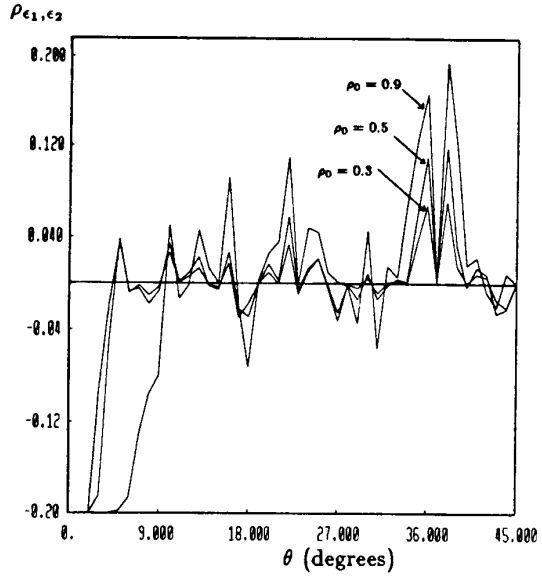

Figure 4: $\rho_{\epsilon_1,\epsilon_2}$ as a function of the rotational angle with dynamic range ($\sigma_0/\Delta = 5$)

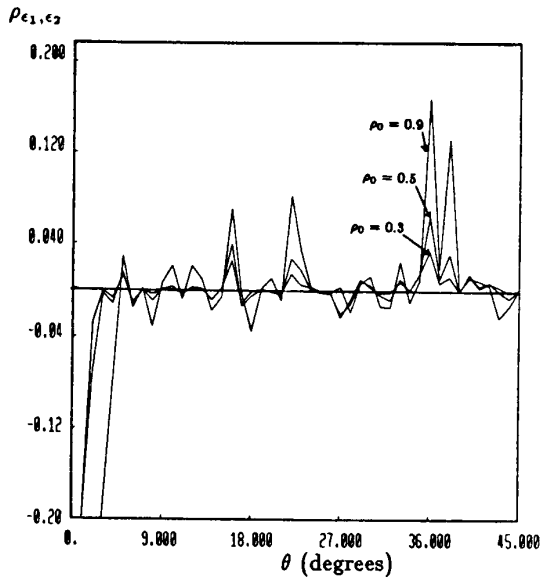

Figure 6: $\rho_{\epsilon_1,\epsilon_2}$ as a function of the rotational angle with dynamic range ($\sigma_0/\Delta = 10$)

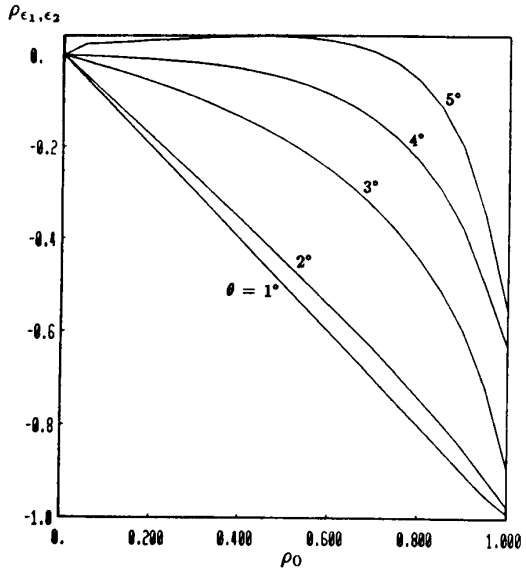

Figure 5: $\rho_{\epsilon_1,\epsilon_2}$ as a function of ρ_0 for small angles of rotation ($\sigma_0/\Delta = 5$)

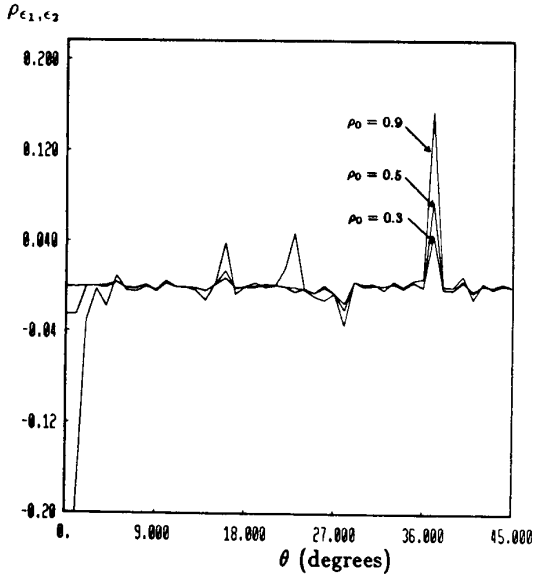

Figure 7: $\rho_{\epsilon_1,\epsilon_2}$ as a function of the rotational angle with dynamic range ($\sigma_0/\Delta = 25$)