

# On Constrained Randomized Quantization \*

Emrah Akyol and Kenneth Rose

*University of California, Santa Barbara, CA 93106, USA*

*Email:{eakyol, rose}@ece.ucsb.edu*

## Abstract

Randomized (dithered) quantization is a method capable of achieving white reconstruction error independent of the source. Dithered quantizers have traditionally been considered within their natural setting of uniform quantization. We extend conventional dithered quantization to nonuniform quantization, via a subterfuge: dithering is performed in the companded domain. Closed form necessary conditions for optimality of the compressor and expander mappings are derived for both fixed and variable rate randomized quantization. Numerically, mappings are optimized by iteratively imposing these necessary conditions. The framework is extended to include an explicit constraint that deterministic or randomized quantizers yield reconstruction error that is uncorrelated with the source. Surprising theoretical results show direct and simple connection between the optimal constrained quantizers and their unconstrained counterparts. Numerical results for the Gaussian source provide strong evidence that the proposed constrained randomized quantizer outperforms the conventional dithered quantizer, as well as the constrained deterministic quantizer.

## 1 Introduction

Dithered quantization is a randomized quantization method introduced in [1]. A central motivation for dithered quantization is its ability to yield quantization error that is white and independent of the source, which can be achieved if certain conditions, determined by Schuchman, are met [2]. Traditionally, dithered quantization has been studied in the framework where the quantizer is uniform (with step size  $\Delta$ ) and the dither signal is uniformly distributed over  $(-\frac{\Delta}{2}, \frac{\Delta}{2})$ , matched to the quantizer interval as shown in Figure 1. A uniformly distributed dither signal is added before quantization and the same dither signal is subtracted from the quantized value at the decoder side. Note that only subtractive dithering is considered in this paper. In the variable rate case, the quantized values are entropy coded, conditioned on the dither signal. Randomized (dithered) quantizers have been studied in the past due to important properties that differentiate them from deterministic quantizers, and were employed to characterize rate-distortion bounds for universal compression [3, 4]. Zamir and Feder provide extensive studies of the properties of dithered quantizers [5, 6].

---

\*The work is supported in part by the NSF under grants CCF-0728986, CCF-1016861 and CCF-1118075.

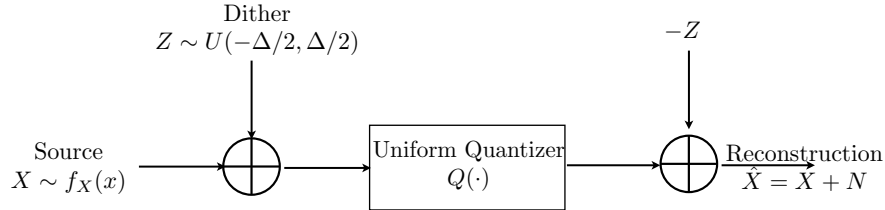


Figure 1: The basic structure of dithered quantization

Beyond its theoretical significance, randomized quantization is of practical interest. Many filter/system optimization problems in practical compression settings, such as the rate-distortion optimal filterbank design problem [7], or low rate filter optimization for DPCM compression of Gaussian auto-regressive processes [8], assume quantization noise that is independent of (or uncorrelated with) the source. Although this assumption is satisfied at asymptotically high rates [9], such systems are mostly useful for very low rate applications. For example, in [8], it is stated that the assumptions made in the paper are not satisfied by deterministic quantizers, and that dithered quantizers satisfy the assumptions exactly. However, conventional (uniform) dithered quantization suffers from suboptimal compression performance. Hence, a quantizer that mostly satisfies the assumptions, but at minimal cost in performance degradation, would have considerable impact on many such applications.

In this paper, we consider a generalization to enable effective dithering of nonuniform quantizers. To the best of our knowledge, this paper is the first attempt (other than our preliminary work in [10]) to consider dithered quantization in a nonuniform quantization framework. One immediate problem with nonuniform dithered quantization is how to apply dithering to unequal quantization intervals. In traditional dithered quantization, the dither signal is matched to the uniform quantization interval while maintaining independence of the source, but it is not clear how to match the generic dither to varying quantization intervals. As a remedy to this problem, we propose dithering in the companded domain. We derive the closed form necessary conditions for optimality of the compressor and expander mappings for both fixed and variable rate randomized quantization. We numerically optimize the mappings by iteratively imposing these necessary conditions

However, the resulting (unconstrained randomized) quantizer does not render reconstruction error orthogonal to the source. Therefore, we extend the framework to include an explicit such constraint. Surprising theoretical results show direct and simple connection between the optimally constrained random quantizers and their unconstrained counterparts. We note in passing that the nonuniform dithered quantizer subsumes the conventional uniform dithered quantizer as an extreme special case.

For the variable rate case, the proposed nonuniform dithered quantizer is expected to outperform the conventional dithered quantizer, most significantly at low rates where the optimal variable rate (entropy coded) quantizer is often far from uniform. We observe that a deterministic quantizer cannot render the quantization noise independent of the source but can make it uncorrelated with the source. We hence also present an alternative deterministic quantizer that provides quantization noise uncorrelated with the source. We derive the optimality conditions of such constrained quantizers, for both fixed and variable rate quantization, and compare their rate-distortion performance to that of randomized quantizers.

The paper is organized as follows: In Section II, we present a review of dithered

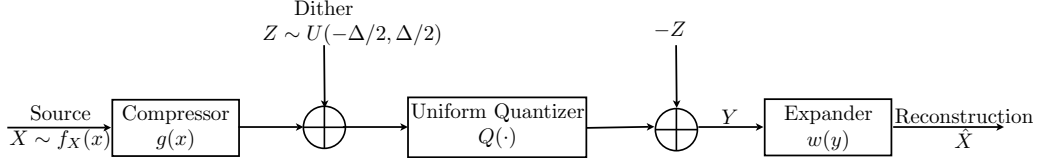


Figure 2: The proposed nonuniform dithered quantizer

quantization. In Section III, we present the proposed nonuniform randomized quantizer. In Section IV, the constrained quantizers are analyzed. Experimental results that compare the proposed quantizers to the conventional dithered quantizer are presented in Section V. We discuss the obtained results and summarize the contributions in Section VI.

## 2 Review of Dithered Quantization

We follow standard notation for information-theoretic quantities (see e.g., [11]). We assume zero mean sources throughout the paper. Zero-mean vectors  $\mathbf{X} \in \mathbb{R}^K$  and  $\mathbf{Y} \in \mathbb{R}^M$  are said to be uncorrelated if they are orthogonal:  $\mathbb{E}[\mathbf{Y}\mathbf{X}^T] = \mathbf{0}$  where  $\mathbf{0}$  is  $M \times K$  matrix of zeros. A quantizer is defined by a set of reconstruction points and a partition. The partition  $\mathcal{P} = \{\mathcal{P}_i\}$  associated with a quantizer is a collection of disjoint regions whose union covers  $\mathbb{R}^K$ . The reconstruction points  $\mathcal{R} = \{\mathbf{r}_i\}$  are typically chosen to minimize a distortion measure. The vector quantizer is a mapping  $Q : \mathbb{R}^K \rightarrow \mathcal{C} \subset \mathbb{R}^K$ , where  $\mathcal{C}$  is a countable set called the codebook, that maps every vector  $\mathbf{X} \in \mathbb{R}^K$  into the reconstruction point that is associated with the cell containing  $\mathbf{X}$ ,

$$Q(\mathbf{X}) = \mathbf{r}_i \text{ if } \mathbf{X} \in \mathcal{P}_i \quad (1)$$

While our theoretical results are general, for a vector quantizer of arbitrary dimensions, for presentation simplicity, we will primarily focus on scalar quantization in the treatment of numerical optimization of nonuniform dithered quantizer and for experimental results. The nonuniform dithered quantization approach is directly extendable to vector quantization by replacing the uniform quantizer with a lattice quantizer, although at the cost of significantly more challenging numerical optimization.

The scalar uniform quantizer, with reconstructions  $\{0, \pm\Delta, \pm2\Delta, \dots, \pm T\Delta\}$ , is a mapping  $Q : \mathbb{R} \rightarrow \mathbb{R}$  such that

$$Q(x) = i\Delta \text{ for } i\Delta - \Delta/2 < x \leq i\Delta + \Delta/2 \quad (2)$$

In fixed rate quantization, the range parameter  $T$  is determined by the rate  $R_f = \log(2T+1)$  while in variable rate quantization  $T$  need not, in principle, be finite and we will assume  $T \rightarrow \infty$ . In this case, uniform quantization is followed by lossless source encoding (entropy coder). Let dither  $Z$  be a random variable, distributed uniformly on the interval  $(-\Delta/2, \Delta/2)$ . Then, conventional dithered quantizer approximates the source  $X$  by

$$\hat{X} = Q(X + Z) - Z \quad (3)$$

It can be shown that the reconstruction error of this quantizer (denoted  $N$ ) is independent of the source value  $X = x$ , i.e.,  $N = \hat{X} - X = Q(X + Z) - Z - X$  is independent of  $X$  and uniformly distributed over  $(-\Delta/2, \Delta/2)$  for all  $X$ . Contrast

that with a deterministic quantizer, whose error is completely determined by the source value [9]. We note that for this property to hold, the quantizer should span the support of the source density i.e., there should be no overload distortion. While this is often the case for variable rate quantization, for fixed rate overload distortion is inevitable if the source has unbounded support such as a Gaussian source. For practical purposes though, it is common to assume that the source has finite support and we also follow this assumption in our analysis of fixed rate randomized quantization: the quantization error of conventional (uniform) dithered fixed rate quantization is assumed to be independent of the source.

The realization of the dither random variable  $Z$  is available to both the encoder and the decoder. Thus, assuming an optimal entropy coder, the rate of the variable rate quantizer tend to the conditional entropy of the reconstruction given the dither,

$$R_v = H(\hat{X}|Z) = H(Q(X + Z)|Z) \quad (4)$$

In [5], it was shown that the following holds:

$$H(Q(X + Z)|Z) = h(X + N) - \log \Delta \quad (5)$$

### 3 Nonuniform Dithered Quantizer

The main idea is to circumvent the main difficulty due to unequal quantization intervals by performing uniform dithered quantization in the companded domain (see Figure 2). The source  $X$  is transformed through compressor  $g(\cdot)$  before dithered uniform quantization. At the decoder side, the dither is subtracted to obtain  $Y$ . Since we perform uniform dithered quantization in the companded domain, it is easy to show that  $Y = g(X) + N$ , where  $N$  is uniformly distributed over  $(-\Delta/2, \Delta/2)$  and independent of the source. The reconstruction is obtained by applying the expander  $\hat{X} = w(Y)$ . The objective is to find the optimal compressor and expander mappings  $g(x), w(y)$  that minimize the expected distortion under the rate constraint. The MSE distortion can be written as:

$$D = \int \int [x - w(g(x) + n)]^2 f_X(x) f_N(n) dx dn \quad (6)$$

Note that the conventional dithered quantizer is a special case employing the trivial identity mappings, i.e.,  $g(x) = w(x) = x, \forall x$ .

#### 3.1 Optimal Expander

The conditional expectation  $w(y) = \mathbb{E}\{X|y\}$  minimizes MSE between the source and the estimate. Then, the optimal expander  $w$  is

$$w(y) = \frac{\int_{\gamma_-}^{\gamma_+} x f_X(x) dx}{\int_{\gamma_-}^{\gamma_+} f_X(x) dx} \quad (7)$$

where for fixed rate  $\gamma_+ = \min\{g^{-1}(\Delta K), g^{-1}(y + \Delta/2)\}$  and  $\gamma_- = \max\{g^{-1}(-\Delta K), g^{-1}(y - \Delta/2)\}$  while for variable rate  $\gamma_+ = g^{-1}(y + \Delta/2)$  and  $\gamma_- = g^{-1}(y - \Delta/2)$ .

## 3.2 Optimal Compressor

Unlike the expander, the optimal compressor cannot be written in closed form. However, a necessary optimality condition can be obtained by setting the functional derivative of the cost to zero. Thus, for a given expander  $w(y)$ , the functional derivative of the total cost,  $J$ , along the direction of any variation function  $\eta(x)$  vanishes [12], i.e.,

$$\nabla J = \left. \frac{\partial}{\partial \epsilon} \right|_{\epsilon=0} J[g(x) + \epsilon \eta(x)] = 0, \forall x \in \mathbb{R} \quad (8)$$

for the locally optimal compressor  $g(x)$ .

### 3.2.1 Fixed rate

For fixed rate, we have granular distortion, denoted  $D_g$ , and overload distortion, denoted  $D_{ol}$ . Note that we need the overload distortion terms here, because otherwise  $g(x)$  will grow unboundedly in the iterations of the proposed algorithm that enforces necessary conditions. Since the rate is fixed, the total cost is identical to the distortion in the fixed rate case, i.e.,  $J_f = D_g + D_{ol}$  and  $D_g$  and  $D_{ol}$  are:

$$D_g = \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} \int_{\gamma_-}^{\gamma_+} [x - w(g(x) + n)]^2 f_X(x) dx dn \quad (9)$$

$$D_{ol} = \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} \int_{-\infty}^{\gamma_-} [x - w(-K\Delta + n)]^2 f_X(x) dx + \int_{\gamma_+}^{\infty} [x - w(K\Delta + n)]^2 f_X(x) dx dn \quad (10)$$

where  $\gamma_+ = g^{-1}(\Delta K)$  and  $\gamma_- = g^{-1}(-\Delta K)$ .

### 3.2.2 Variable rate

To find the rate by (5), we need the distribution of  $Y = g(X) + N$ , which can be written as <sup>1</sup>

$$f_Y(y) = \frac{1}{\Delta} [F_X(g^{-1}(y + \Delta/2)) - F_X(g^{-1}(y - \Delta/2))] \quad (11)$$

where  $F_X(x)$  is the cumulative distribution function of  $X$ . The rate is then evaluated as

$$R_v = h(Y) - \log \Delta \quad (12)$$

The total cost for variable rate quantization is  $J_v = D + \lambda R$  where  $\lambda$  is the Lagrangian parameter that is adjusted to obtain the desired rate.

---

<sup>1</sup>We assume that  $g(x)$  is monotonically increasing, as usually done in conventional quantization.

### 3.3 Design Algorithm

The basic idea is to iteratively alternate between enforcing each necessary condition for optimality, thereby successively decreasing the total cost. Iterations are performed until the algorithm reaches a stationary point. Solving for the optimal expander is straightforward since the expander is expressed in closed form as a functional of the known quantities,  $g(x)$ ,  $f_X(x)$ . Since the compressor condition is not in closed form, we perform steepest descent, i.e., move in the direction of the functional derivative of the total cost with respect to the compressor mapping  $g$ .

$$g_{i+1}(x) = g_i(x) - \mu \nabla J[g] \quad (13)$$

By design, the total cost decreases monotonically as the algorithm proceeds iteratively. The compressor mapping is updated according to (13), where  $i$  is the iteration index,  $\nabla J[g]$  is the directional derivative and  $\mu$  is the step size. There is no guarantee that an iterative descent algorithm of this type will converge to the globally optimal solution. As a low complexity approach to mitigate the poor local minima problem, we used the “noisy channel relaxation” method of [13, 14].

## 4 Reconstruction Error Uncorrelated with the Source

In this section, we propose two quantization schemes (one deterministic, one randomized) that satisfy the assumption reconstruction error uncorrelated with the source.

### 4.1 Constrained Deterministic Quantizer

A deterministic quantizer cannot yield quantization noise independent of the source [9]. However, it is possible to render the quantization noise uncorrelated with the source. An early prior work along this line appeared in [15], where a uniform quantizer is converted to a quantizer whose quantization noise is uncorrelated with the source, by adjusting the reconstruction points. In this section, we derive the optimal (nonuniform in general) deterministic quantizer which is constrained to give quantization error uncorrelated with the source. Let  $\mathbf{r}_i$  and  $\hat{\mathbf{r}}_i$  be the reconstruction points and  $\mathcal{P}_i$  and  $\hat{\mathcal{P}}_i$  represent the  $i^{\text{th}}$  quantization region, for the constrained (i.e., whose quantization error is uncorrelated with the source) and unconstrained MSE optimal quantizer, respectively. Also, let  $p_i$  and  $\hat{p}_i$  denote the probability of  $X$  falling into the  $i^{\text{th}}$  cell of the constrained and unconstrained quantizers, respectively.

**Theorem 1.**  $\mathcal{P}_i = \hat{\mathcal{P}}_i$  and  $\mathbf{r}_i = \mathbf{C} \hat{\mathbf{r}}_i, \forall i$  where  $\mathbf{C} = \mathbf{R}_X \left( \sum_{i=1}^M p_i \hat{\mathbf{r}}_i \hat{\mathbf{r}}_i^T \right)^{-1}$

*Proof.* We start with the fixed rate analysis. Let  $M$  denote the number of quantization cells. The distortion can be expressed as

$$D = \sum_{i=1}^M \int_{\mathbf{x} \in \mathcal{P}_i} (\mathbf{x} - \mathbf{r}_i)^T (\mathbf{x} - \mathbf{r}_i) f_X(\mathbf{x}) d\mathbf{x} \quad (14)$$

and the ‘‘uncorrelatedness’’ constraint may be stated via the orthogonality principle

$$\sum_{i=1}^M \int_{\mathbf{x} \in \mathcal{P}_i} \mathbf{x}(\mathbf{x} - \mathbf{r}_i)^T f_X(\mathbf{x}) d\mathbf{x} = \mathbf{0} \quad (15)$$

Note further that (15) can be written as:

$$\sum_{i=1}^M \mathbf{r}_i \mathbf{l}_i^T = \mathbf{R}_X \text{ where } \mathbf{l}_i = \int_{\mathbf{x} \in \mathcal{P}_i} \mathbf{x} f_X(\mathbf{x}) d\mathbf{x} \quad (16)$$

The constrained problem of minimizing  $D$  subject to  $\sum_{i=1}^M \mathbf{r}_i \mathbf{l}_i^T = \mathbf{R}_X$  is equivalent to the unconstrained minimization of Lagrangian  $J$ , where

$$J = D + \sum_{k=1}^M \gamma(k)^T \left[ \mathbf{R}_X(k) - \sum_{i=1}^M \mathbf{r}_i \mathbf{l}_i(k) \right] \quad (17)$$

where  $\boldsymbol{\gamma} = [\gamma(1) \ \gamma(2) \dots \ \gamma(M)]$  denotes the  $M \times M$  Lagrangian matrix,  $\mathbf{R}_X(k)$  denotes the  $k^{\text{th}}$  column of  $\mathbf{R}_X$  and  $\mathbf{l}_i(k)$  denotes the  $k^{\text{th}}$  element of  $\mathbf{l}_i$ . By setting  $\nabla_{\mathbf{r}_i} J = \mathbf{0}$ , we obtain the condition:

$$\nabla_{\mathbf{r}_i} J = -2\mathbf{l}_i^T + 2p_i \mathbf{r}_i^T - \sum_{k=1}^M \gamma(k)^T \mathbf{l}_i(k) = \mathbf{0} \quad (18)$$

Noting that  $\sum_{k=1}^M \gamma(k)^T \mathbf{l}_i(k) = \mathbf{l}_i^T \boldsymbol{\gamma}$ , we obtain  $\mathbf{r}_i = \frac{1}{p_i} \mathbf{C} \mathbf{l}_i$  where  $\mathbf{C}$  is a constant  $M \times M$  matrix.  $\mathbf{C}$  is found by plugging this into (16):

$$\mathbf{C} = \mathbf{R}_X \left( \sum_{i=1}^M \frac{1}{p_i} \mathbf{l}_i \mathbf{l}_i^T \right)^{-1} \quad (19)$$

Note that  $\mathbf{l}_i/p_i$  is the MSE optimal reconstruction of an unconstrained quantizer that shares the same decision boundary with the constrained one,  $\mathcal{P}_i$ . Plugging (19) into (14) and after some algebraic manipulations, we obtain:

$$D = \frac{\sigma_X^2}{\sigma_X^2 - D^*} D^* \quad (20)$$

where  $D^*$  is the distortion associated with the quantizer given by  $\mathcal{P}_i$  and with corresponding optimal reconstruction points  $\mathbf{l}_i/p_i$ . (20) implies that  $D$  achieves its minimum whenever  $D^*$  is minimized. Hence,

$$\mathcal{P}_i = \widehat{\mathcal{P}}_i \Rightarrow \mathbf{l}_i = p_i \widehat{\mathbf{r}}_i \quad (21)$$

Plugging (21) into (4.1) and (19), we obtain the result. The proof for variable rate goes along similar lines, and we skip that part for brevity.  $\square$

## 4.2 Constrained Randomized Quantizer

Due to the effect of companding, the nonuniform randomized quantizer described above does not guarantee reconstruction error uncorrelated with the source even though it builds on the (conventional) dithered quantizer whose quantization error is independent of the source. We therefore explicitly constrain the randomized quantizer to generate uncorrelated reconstruction error, by adding a penalty term to the total cost function. The Lagrangian parameter  $\lambda_c \geq 0$  is set to ensure  $\mathbb{E}\{xw(g(x) + n)\} = \mathbb{E}\{x^2\}$ .

$$J_c = J + \lambda_c \mathbb{E}[x^2 - xw(g(x) + n)] \quad (22)$$

where  $J = J_v$  in the case of variable rate and  $J = J_f$  for fixed rate. We find the necessary conditions of optimality of constrained compressor and expander mappings at fixed and variable rate, by setting the functional derivative of the total cost ( $J_c$ ) to zero. Surprisingly, the optimally constrained compressor mapping remains unchanged (compared to the unconstrained optimal compressor) and the only modification of the optimally constrained expander mapping is simple scaling. We state this result in the following theorem.

**Theorem 2.** *Let  $g$  and  $w$  be the compressor and expander mappings of the unconstrained optimal randomized quantizer. Let  $g_c$  and  $w_c$  denote the optimal mappings subject to the constraint that the reconstruction error be uncorrelated with the source. Then,*

$$g_c(x) = g(x), w_c(y) = (1 - \lambda_c)w(y) \quad (23)$$

where  $\lambda_c$  is the Lagrangian multiplier of (22).

Note that this result applies to both fixed and variable rate.

*Proof.* The optimal expander is no longer the standard conditional expectation, since it is impacted by the constraint. By setting  $\left. \frac{\partial}{\partial \epsilon} J_c[w(y) + \epsilon \eta(y)] \right|_{\epsilon=0} = 0$ , we obtain the optimal expander in closed form as  $w_c(y) = (1 - \lambda_c)w(y)$ . The update rule for  $g_c(x)$  can be derived similarly. Setting  $\left. \frac{\partial}{\partial \epsilon} J_c[g(x) + \epsilon \eta(x)] \right|_{\epsilon=0} = 0$  and plugging  $w_c(y) = (1 - \lambda_c)w(y)$  yields, after straightforward algebra,  $g_c(x) = g(x)$ .  $\square$

## 5 Experimental Results

In this section, we numerically compare the proposed quantizers to the conventional (uniform) dithered quantizer and to the optimal quantizer, for a standard unit variance scalar Gaussian source. We implemented the proposed quantizers by numerically calculating the derived integrals. For that purpose, we sampled the distribution on a uniform grid. We also imposed bounded support ( $-3\sigma$  to  $3\sigma$ ) i.e., we numerically neglected the tails of the Gaussian. In this paper, we proposed three quantizers:

**Quantizer 1:** Unconstrained randomized quantizer.

**Quantizer 2:** Constrained randomized quantizer which renders the quantization error uncorrelated with the source.

**Quantizer 3:** Constrained deterministic quantizer which renders the quantization error uncorrelated with the source.

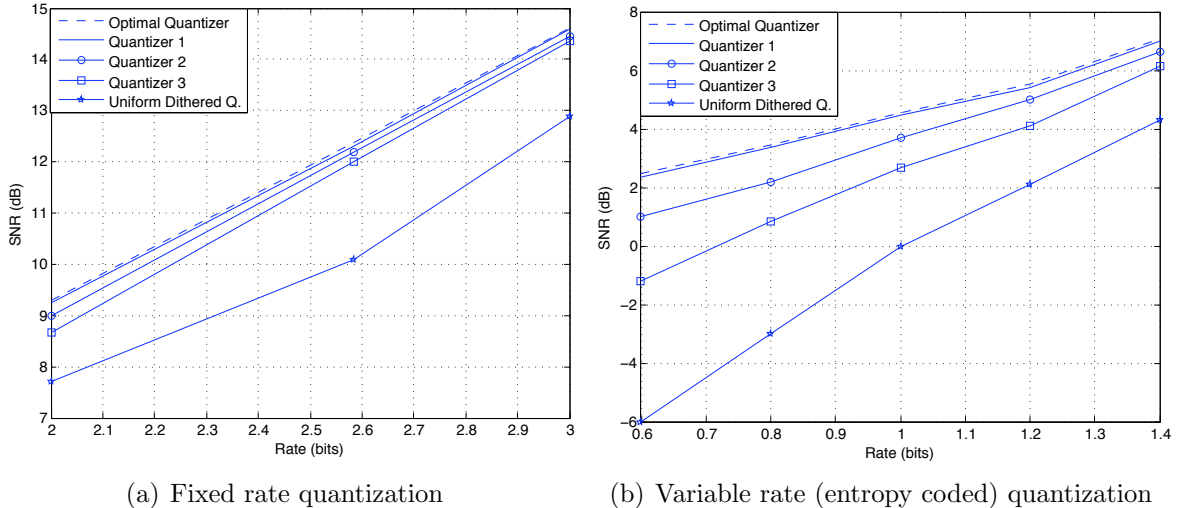


Figure 3: Performance comparison in terms of SNR versus rate.

Figure 3 demonstrates the performance comparisons among quantizers for fixed and variable rates. Note that for both fixed and variable rate, the optimal randomized quantizer performs very close to the optimal quantizer. However, it does not provide the statistical benefits of the other quantizers.

Note that for fixed rate, conventional (uniform) dithered quantization suffers significantly from the suboptimality of having equal quantization intervals irrespective of the rate region. However, at variable rate, the difference between the proposed and conventional dithered quantizer diminish at high rates, while at low rates the difference is quite significant. This is theoretically expected since at high rates, the optimal variable rate quantizer is very close to uniform, hence there is not much to gain from using a non-linear compressor-expander.

For both fixed and variable rate, the constrained randomized quantizer outperforms its deterministic counterpart, while both of them perform significantly better than the conventional dithered quantizer. Both of the proposed quantizers render quantization error *uncorrelated* with the source with low performance degradation while the dithered quantizer renders error *independent* of the source but (depending on the rate) at significant distortion penalty.

Numerical comparisons show that the proposed quantization schemes can significantly impact the design of compression systems such as [8, 7] where quantization error is assumed to be uncorrelated with the source. Note that the constrained randomized quantization satisfies this assumption exactly and significantly outperforms the conventional dithered quantization, which has been presented in such prior work as the viable option to satisfy these assumptions. In fact, besides the conventional dithered quantization, we derived additional quantization schemes that satisfy those assumptions: constrained deterministic quantization and constrained nonuniform random quantization. We also derived an unconstrained randomized quantizer, which performs almost as well as the optimal (deterministic) quantizer, yet offers perceptual benefits typical to dithered quantization.

## 6 Discussion

In this paper, we proposed a nonuniform randomized quantizer where dithering is performed in the companded domain to circumvent the problem of matching the dither range to varying quantization intervals. The optimal compressor and expander mappings that minimize the mean square error are found via a novel numerical method. Also, we discovered the connections between the optimal quantizer and the one whose reconstruction error is constrained to be orthogonal to the source, for both deterministic and randomized quantization. The proposed constrained randomized quantization outperforms conventional dithered quantization and also its deterministic counterpart, while still satisfying the requirement that the reconstruction error be uncorrelated with the source.

## References

- [1] L. Roberts, "Picture coding using pseudo-random noise," *IEEE Trans. on Inf. Theory*, vol. 8, no. 2, pp. 145–154, 1962.
- [2] L. Schuchman, "Dither signals and their effect on quantization noise," *IEEE Trans. on Communications*, vol. 12, no. 4, pp. 162–165, 1964.
- [3] J. Ziv, "On universal quantization," *IEEE Trans. on Inf. Theory*, vol. 31, no. 3, pp. 344–347, 1985.
- [4] R.M. Gray and T.G. Stockham Jr, "Dithered quantizers," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 805–812, 1993.
- [5] R. Zamir and M. Feder, "On universal quantization by randomized uniform/lattice quantizers," *IEEE Trans. on Inf. Theory*, vol. 38, no. 2 Part 2, pp. 428–436, 1992.
- [6] 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.
- [7] M.K. Mihcak et.al., "Rate-distortion-optimal subband coding without perfect-reconstruction constraints," *IEEE Trans. on Signal Pr.*, vol. 49, pp. 542–557, 2001.
- [8] O.G. Guleryuz and M.T. Orchard, "On the DPCM compression of Gaussian autoregressive sequences," *IEEE Trans. on Inf. Theory*, vol. 47, no. 3, pp. 945–956, 2001.
- [9] A. Gersho and R Gray, *Vector Quantization and Signal Compression*, Springer, 1992.
- [10] E. Akyol and K. Rose, "Nonuniform Dithered Quantization," in *Proceeding of the IEEE Data Compression Conference*, 2009, p. 435.
- [11] T. Cover and J. Thomas, *Elements of Information Theory*, J.Wiley NY, 1991.
- [12] D. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons Inc, 1969.
- [13] S. Gadkari and K. Rose, "Robust vector quantizer design by noisy channel relaxation," *IEEE Transactions on Communications*, vol. 47, no. 8, pp. 1113–1116, 1999.
- [14] P. Knagenhjelm, "A recursive design method for robust vector quantization," in *Proc. Int. Conf. Signal Processing Applications and Technology*, 1992, pp. 948–954.
- [15] A. Hjørungnes, *Optimal Bit and Power Constrained Filter Banks*, Ph.D. thesis, Norwegian University of Science and Technology, 2000.