On Constrained Randomized Quantization

Emrah Akyol, *Member, IEEE*, and Kenneth Rose, *Fellow, IEEE*

***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. In this paper we extend conventional dithered quantization to nonuniform quantization, via a subterfage: 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. Moreover, the proposed constrained quantizer renders the reconstruction error nearly white. In the second part of the paper, we investigate whether uncorrelated reconstruction error requires random coding to achieve asymptotic optimality. We show that for a Gaussian source, the optimal vector quantizer of asymptotically high dimension whose quantization error is uncorrelated with the source, is indeed random. Thus, random encoding in this setting of rate-distortion theory, is not merely a tool to characterize performance bounds, but a required property of quantizers that approach such bounds.**

***Index Terms*—Source coding, dithered quantization, subtractive dithering, compander, quantizer design, analog mappings.**

I. 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 independent of the source, which can be achieved if certain conditions, determined by Schuchman, are met [2].

The conventional dithered quantization framework involves a uniform quantizer, with step size Δ, and a dither signal uniformly distributed over $(-\frac{\Delta}{2}, \frac{\Delta}{2})$, matched to the quantizer interval as shown in Fig. 1. The dither signal is added before quantization. In subtractive dithering, the same dither signal is subtracted from the quantized value at the decoder side, while in non-subtractive dithering decoder does not have access to the dither signal. Subtractive dithering renders the quantization error independent of the source. We consider only subtractive dithering in this paper, while noting that the basic ideas are also applicable to non-substractive dithering.

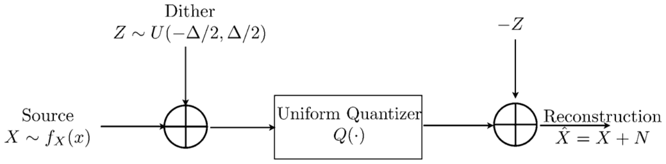

Fig. 1. The basic structure of dithered quantization.

Both subtractive and non-subtractive dithering play key roles in real world applications. Dithering is routinely used in audio and video applications where audio-visual properties are paramount, e.g., non-subtractive dithering is often one of the last stages of audio production for storage on compact disc [3]. Notably, Vanderkooy and Lipshitz, [4] studied various noise types which differ in their effect when used as dither signals and suggested optimal dither levels for audio. In [5] non-subtractive dither is studied in detail from a theoretical point of view. Li *et al.* [6] considered distribution preserving dithered quantization to improve the perceptual quality of mean square optimal quantizers in audio and video coding. The randomized quantizer in their model, outputs a signal with the same distribution as the source. Saldi et.al., very recently generalized this problem by requiring that the output of the quantizer have a prescribed distribution [7].

On the more theoretical side, 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 [8], [9]. The continued (theoretical) interest in dithered quantizers is due to their statistical properties. Zamir and Feder provide extensive studies of the properties of dithered quantizers [10], [11]. The results of these studies, specifically the fact that the dithered lattice quantizer at asymptotically high dimension realizes the Gaussian test channel, have led to the wide use of entropy coded dithered lattice quantization as a "structured method" to achieve fundamental bounds obtained via random (unstructured) coding arguments [12], see e.g., [13]–[15] for a sample of such applications and [16] for an overview.

Beyond its theoretical significance, randomized quantization is of significant practical interest. A main application area of dithered quantizers is the analysis of quantization in complex systems, due its simplicity in modeling quantization errors. For instance, Goyal recently investigated the performance of a collection of subtractively-dithered uniform scalar quantizers with the same step size, used in parallel as a model for the randomly varying uniform conventional quantizers [17]. Dithered

Manuscript received June 07, 2012; revised February 06, 2013; accepted April 18, 2013. Date of publication May 02, 2013; date of current version June 04, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tongtong Li. This work is supported by the NSF under the grants CCF-0728986, CCF-1016861 and CCF 1118075. The material in this paper was presented in part at the IEEE Data Compression Conferences, March 2009 and April 2012.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: eakyol@ece.ucsb.edu; rose@ece.ucsb.edu.

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.2013.2261296

quantizers have been found to be useful in analog-digital converters in general, particularly in delta-sigma modulators [18]–[20] where statistics of the quantization error is an important consideration. Another example pertains to distributed averaging algorithms (consensus) [21] which recently gained revived interest [22], [23]. In [24], it was shown that quantized consensus is significantly different than the original (unquantized) problem and the effect of quantization is arbitrary and hard to analyze. To alleviate the quantization problem, it is widely accepted practice to use dithered quantization due to its statistical advantages, see e.g., [25], [26]. Many filter/system optimization problems in practical compression settings, such as the "rate-distortion optimal filterbank design" problem [27], or low rate filter optimization for DPCM compression of Gaussian auto-regressive processes [28], assume quantization noise that is independent of (or uncorrrelated with) the source. Although this assumption is satisfied at asymptotically high rates [29], such systems are mostly useful for very low rate applications. For example, in [28], 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 [30], [31]) 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 connections 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.

Dithered quantization offers an interesting theoretical twist. Randomized quantization is an instance of the random encoding principle used to elegantly prove the achievability of coding bounds in rate distortion theory [32]. However, to actually achieve those bounds, a random encoding scheme is not necessary, as they can be approached by a sequence of deterministic quantizers of increasing block length. In the second part of the paper, we investigate the settings under which randomized quantization is asymptotically necessary. A trivial example involves requiring source-independent quantization error. It is obvious that the reconstruction (hence quantization error) is a deterministic function of the source when the quantizer is deterministic [29], while conventional dithered quantization produces quantization error that is independent of the source. Although a deterministic quantizer can never render the quantization error independent of the source, it can produce quantization error uncorrelated with the source. A natural question is whether the rate distortion bound, subject to the uncorrelated error constraint, can be achieved (asymptotically) with a deterministic quantizer.

The paper is organized as follows: In Section III, we present the proposed nonuniform randomized quantizers, along with its extension to constrained randomized quantizer that renders the quantization error orthogonal to the source. In Section IV, we derive the necessary conditions of optimality for the deterministic quantizer that generates reconstruction error uncorrelated with the source. In Section V, we study the asymptotic (in quantizer dimension) results, and show that for a Gaussian source, the optimal constrained quantizer must be randomized. Experimental results that compare the proposed quantizers to the conventional dithered quantizer are presented in Section VI. We discuss the results and summarize the contributions in Section VII.

II. Review of Dithered Quantization

A. Notation and Preliminaries

Let $\mathbb{R}$ and $\mathbb{R}^+$ denote the respective sets of real numbers, and positive real numbers. In general, lowercase letters (e.g., x) denote scalars, boldface lowercase (e.g., $\boldsymbol{x}$) vectors, uppercase (e.g., U, X) matrices and random variables, and boldface uppercase (e.g., $\boldsymbol{X}$) random vectors. R_X, and R_{XZ} denote covariance of $\boldsymbol{X}$ and cross covariance of $\boldsymbol{X}$ and $\boldsymbol{Z}$ respectively[1]. $\mathbb{E}(\,\cdot\,)$ and $\mathbb{P}(\,\cdot\,)$ denote the expectation and probability operators, respectively. ∇ denotes the gradient and ∇_x denotes the partial gradient with respect to $\boldsymbol{x}$. $f'(\,\cdot\,)$ denotes the first order derivative of the function $f(\,\cdot\,)$, i.e., $f'(x) = \frac{df(x)}{dx}$. All the logarithms in the paper are natural logarithms and may in general be complex. Integrals are, in general, Lebesgue integrals. $\mathcal{N}(\boldsymbol{\mu}, \mathcal{K})$ denotes the Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix $\mathcal{K}$, and $U[a, b]$ denotes the uniform density over [a,b]. Let $\mathcal{S}^+$ denote the set of monotonically increasing, Borel measurable functions $\{g(\,\cdot\,) : \mathbb{R} \to \mathbb{R}\}$.

[1] We assume zero mean random variables. This assumption is not necessary, but it considerably simplifies the notation. Therefore, it is kept throughout the paper.

The entropy of a discrete random vector $\boldsymbol{X} \in \mathbb{R}^K$ taking values in $\mathcal{X}$ is

$$H(\boldsymbol{X}) = -\sum_{\boldsymbol{x} \in \mathcal{X}} \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) \log \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) \tag{1}$$

where logarithm is base 2 to measure it in bits. The differential entropy of a continuous random vector $\boldsymbol{X}$ with probability density function $f_X(\boldsymbol{x})$ is

$$h(\boldsymbol{X}) = -\int f_X(\boldsymbol{x}) \log f_X(\boldsymbol{x}) d\boldsymbol{x}. \tag{2}$$

The divergence between two densities f_X and g_X, is given by

$$\mathcal{D}(f_X \,\|\, g_X) = \int f_X(\boldsymbol{x}) \log \frac{f_X(\boldsymbol{x})}{g_X(\boldsymbol{x})} d\boldsymbol{x}. \tag{3}$$

The divergence definition above can be extended to conditional densities. For joint densities, f_{XY} and g_{XY} the conditional divergence $\mathcal{D}(f_{X\,|\,Y} \,\|\, g_{X\,|\,Y})$ is defined as the divergence between the conditional distributions $f_{X\,|\,Y}$ and $g_{X\,|\,Y}$ averaged over the density $f_Y(\boldsymbol{y})$:

$$\mathcal{D}(f_{X\,|\,Y} \,\|\, g_{X\,|\,Y}) = \int f_Y(\boldsymbol{y}) \int f_{X\,|\,Y}(\boldsymbol{x}, \boldsymbol{y}) \times \log \frac{f_{X\,|\,Y}(\boldsymbol{x}, \boldsymbol{y})}{g_{X\,|\,Y}(\boldsymbol{x}, \boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{y}. \tag{4}$$

The mutual information between two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ with marginal densities $f_X(\boldsymbol{x})$ and $f_Y(\boldsymbol{y})$ and a joint density $f_{XY}(\boldsymbol{x}, \boldsymbol{y})$ is given by

$$I(\boldsymbol{X}, \boldsymbol{Y}) = \int \int f_{X,Y}(\boldsymbol{x}, \boldsymbol{y}) \log \frac{f_{X,Y}(\boldsymbol{x}, \boldsymbol{y})}{f_X(\boldsymbol{x}) f_Y(\boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{y}. \tag{5}$$

Zero-mean vectors $\boldsymbol{x} \in \mathbb{R}^K$ and $\boldsymbol{y} \in \mathbb{R}^M$ are said to be uncorrelated if they are orthogonal:

$$\mathbb{E}[\boldsymbol{y}\boldsymbol{x}^T] = 0 \tag{6}$$

where the right hand size is $M \times K$ matrix of zeros.

B. Dithered Quantization

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 $\mathcal{R} = \{\boldsymbol{r}_i\}$ are typically chosen to minimize a distortion measure. The vector quantizer is a mapping $Q_K : \mathbb{R}^K \to \mathbb{R}^K$ that maps every vector $\boldsymbol{X} \in \mathbb{R}^K$ into the reconstruction point that is associated with the cell containing $\boldsymbol{X}$, i.e.,

$$Q_K(\boldsymbol{X}) = \boldsymbol{r}_i \;\text{ if } \boldsymbol{X} \in \mathcal{P}_i. \tag{7}$$

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 companded domain uniform quantizer with a lattice quantizer, although at the cost of significantly more challenging numerical optimization.

The scalar uniform quantizer, with reconstructions $\{0, \pm\Delta, \pm 2\Delta, \ldots, \pm T\Delta\}$, is a mapping $Q : \mathbb{R} \to \mathbb{R}$ such that

$$Q(x) = i\Delta \text{ for } i\Delta - \Delta/2 < x \leq i\Delta + \Delta/2. \tag{8}$$

In fixed rate quantization, the range parameter T is determined by the rate R_f

$$R_f = \log(2T + 1) \tag{9}$$

while in variable rate quantization T need not, in principle, be finite and we will assume $T \to \infty$. 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

$$\hat{X} = Q(X + Z) - Z. \tag{10}$$

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 [29].

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 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 tends to the conditional entropy of the reconstruction given the dither, i.e.,

$$R_v = H(\hat{X} \,|\, Z) = H(Q(X + Z) \,|\, Z). \tag{11}$$

In [10], it was shown that the following holds:

$$H(Q(X + Z) \,|\, Z) = h(X + N) - \log\Delta. \tag{12}$$

III. 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 Fig. 2). The source X is transformed through compressor $g(\cdot)$ before undergoing 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

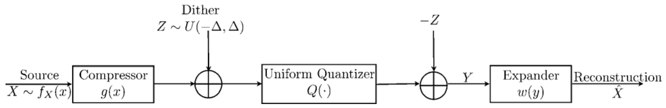

Fig. 2. The proposed nonuniform dithered quantizer.

objective is to find the optimal compressor and expander mappings $g(\cdot), w(\cdot)$ that minimize the expected distortion under the rate constraint. The MSE distortion can be written as:

$$D = \int\int [x - w(g(x) + n)]^2 f_X(x) f_N(n) dx dn \tag{13}$$

where $f_N(n)$ is uniform over $(-\Delta/2, \Delta/2)$. Interestingly, this problem bears some similarity to the joint source channel mapping problem where the optimal analog encoding and decoding mappings are studied [33]. In our setting, the quantization error is analogous to the channel noise and the rate constraint in variable rate quantization plays a role similar to that of the power constraint. Similar to [33], we develop an iterative procedure that enforces the necessary conditions for optimality of the mappings. Note that the conventional (uniform) dithered quantizer is a special case employing the trivial identity mappings, i.e., $g(x) = w(x) = x$ almost everywhere (*a.e.*) in x.

A. Optimal Expander

The conditional expectation $\mathbb{E}\{X \mid Y = y\}$ minimizes MSE between the source and the estimate, hence

$$w(y) = \mathbb{E}\{X \mid y\}. \tag{14}$$

Plugging the expressions for expectation, we obtain

$$w(y) = \int x \, f_{X \mid Y}(x, y) \, \mathrm{d}x. \tag{15}$$

Applying Bayes' rule

$$f_{X \mid Y}(x, y) = \frac{f_X(x) f_{Y \mid X}(x, y)}{\int f_X(x) \, f_{Y \mid X}(x, y) \, \mathrm{d}x} \tag{16}$$

and noting that $f_{Y \mid X}(x, y) = f_N(y - g(x))$, the optimal expander can be written, in terms of known quantities, as

$$w(y) = \frac{\int\limits_{-\infty}^{\infty} x f_X(x) f_N(y - g(x)) dx}{\int\limits_{-\infty}^{\infty} f_X(x) f_N(y - g(x)) dx} \tag{17}$$

where $f_N(\cdot)$ is uniform density over $[-\Delta/2, \Delta/2]$, hence

$$w(y) = \frac{\int\limits_{\gamma_-}^{\gamma_+} x f_X(x) dx}{\int\limits_{\gamma_-}^{\gamma_+} f_X(x) dx} \tag{18}$$

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

Note: We restrict the discussion to regular quantizers throughout this paper, hence $g(\cdot) \in \mathcal{S}^+$, i.e., $g(\cdot)$ is monotonically increasing.

B. 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, a locally optimal compressor $g(\cdot)$, for a given expander $w(\cdot)$, requires that the functional derivative of the total cost, J, along the direction of any admissible[2] variation function $\eta(\cdot)$ vanishes [34], i.e.,

$$\left.\frac{\partial}{\partial \epsilon}\right|_{\epsilon=0} J[g(x) + \epsilon \eta(x)] = 0, \tag{19}$$

a.e. in x, for all admissible perturbation functions $\eta(\cdot)$.

1) Fixed Rate: For fixed rate, we have granular distortion, denoted D_g, and overload distortion, denoted D_{ol}. Note that we must account for the overload distortion here, as this constrains $g(x)$ from growing unboundedly in the iterations of the proposed algorithm. Since the rate is fixed, the total cost only measures the distortion, i.e., $J_f = D_g + D_{ol}$ where D_g and D_{ol} are:

$$D_g = \frac{1}{\Delta} \int\limits_{-\Delta/2}^{\Delta/2} \int\limits_{g^{-1}(-\Delta T)}^{g^{-1}(\Delta T)} [x - w(g(x)+n)]^2 f_X(x) dx dn. \tag{20}$$

$$D_{ol} = \frac{1}{\Delta} \int\limits_{-\Delta/2}^{\Delta/2} \left\{ \int\limits_{-\infty}^{g^{-1}(-\Delta T)} [x - w(-T\Delta+n)]^2 f_X(x) dx + \int\limits_{g^{-1}(\Delta T)}^{\infty} [x - w(T\Delta + n)]^2 f_X(x) dx \right\} dn. \tag{21}$$

2) Variable Rate: The rate is obtained via (11) and (12), which require the distribution of $Y = g(X) + N$:

$$f_Y(y) = \frac{1}{\Delta} \left[F_X(g^{-1}(y + \Delta/2)) - F_X(g^{-1}(y - \Delta/2)) \right] \tag{22}$$

where $F_X(x)$ is the cumulative distribution function of X. The rate is then evaluated using (12) as

$$R_v = h(Y) - \log \Delta. \tag{23}$$

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

C. Design Algorithm

The basic idea is to iteratively alternate between the imposition of individual necessary conditions for optimality, and thereby successively decrease the total Lagrangian cost. Iterations are performed until the algorithm reaches a stationary point. Imposing optimality condition for the expander (18) is straightforward, since the expander can be expressed as closed form functional of known quantities, $g(\cdot)$, $f_X(\cdot)$. The compressor optimality condition (19) is not in closed form

[2] Admissibility here need not be overly restrictive since it is used to derive a necessary condition. Hence, we only require that admissible functions be (Borel) measurable, that integrals exist, and that we can change the order of integration and differentiation.

and we perform steepest descent search in the direction of the functional derivative of the Lagrangian with respect to the compressor mapping $g(\cdot)$. By design, the Lagrangian cost decreases monotonically as the algorithm proceeds iteratively. The update for the compressor is stated generically as

$$g_{i+1}(x) = g_i(x) - \mu \nabla J[g] \tag{24}$$

where i is the iteration index, $\nabla J[g(\cdot)]$ is the directional derivative and μ is the step size. The precise expressions for $\nabla J[g]$ for fixed an variable rate are given in the Appendix along with their derivations.

Note that there is no guarantee that an iterative descent algorithm of this type will converge to the globally optimal solution. The algorithm will converge to a local minimum and hence, initial conditions are important in such greedy optimizations. A low complexity approach to mitigate the poor local minima problem, is to embed within the solution the "noisy channel relaxation" method of [35], [36]. We initialize the compressor mapping with random initial conditions and run the algorithm for a very low rate (large value for the Lagrangian parameter λ). Then, we gradually increase the rate (decrease λ) while tracking the minimum. Note that local minima problem is more pronounced at multi-dimensional optimizations, which hence requires more powerful non-convex optimization tools such as deterministic annealing [37]. In our design and experiments, we focus on scalar compressor and expander and we did not observe significant local minima problems.

IV. Reconstruction Error Uncorrelated With the Source

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

A. Constrained Deterministic Quantizer

A deterministic quantizer cannot yield quantization noise independent of the source [29]. However, it is possible to render the quantization noise uncorrelated with the source. An early prior work along this line appeared in [38], 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 vector quantizer which is constrained to give quantization error uncorrelated with the source.

Let M denote the number of quantization cells. Let $\boldsymbol{r}_i$ and $\widehat{\boldsymbol{r}_i}$ be the reconstruction points and $\mathcal{P}_i$ and $\widehat{\mathcal{P}_i}$ represent the ith quantization region, R_Q and $\hat{R}_Q$ the covariance of the reconstructions, R_E and $\hat{R}_E$ the covariance of the reconstruction error 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 $\boldsymbol{X}$ falling into the ith cell of these respective quantizers.

Theorem 1: $\mathcal{P}_i = \widehat{\mathcal{P}_i}$ and $\boldsymbol{r}_i = R_X \hat{R}_Q^{-1} \hat{\boldsymbol{r}}_i \quad \forall i$.

Proof: We start with the fixed rate analysis. The distortion can be expressed as

$$D = \sum_{i=1}^{M} \int_{\boldsymbol{x} \in \mathcal{P}_i} (\boldsymbol{x} - \boldsymbol{r}_i)^T (\boldsymbol{x} - \boldsymbol{r}_i) f_X(\boldsymbol{x})\, d\boldsymbol{x} \tag{25}$$

and the "uncorrelatedness" constraint may be stated via the orthogonality principle

$$\sum_{i=1}^{M} \int_{\boldsymbol{x} \in \mathcal{P}_i} \boldsymbol{x} (\boldsymbol{x} - \boldsymbol{r}_i)^T f_X(\boldsymbol{x})\, d\boldsymbol{x} = 0 \tag{26}$$

where the right hand side is $M \times M$ matrix of zeros. Note further that (26) can be written as:

$$\sum_{i=1}^{M} \boldsymbol{r}_i \boldsymbol{l}_i^T = R_X \tag{27}$$

where $\boldsymbol{l}_i = \int\limits_{\boldsymbol{x} \in \mathcal{P}_i} \boldsymbol{x} f_X(\boldsymbol{x})\, d\boldsymbol{x}$. The constrained problem of minimizing D subject to (27) is equivalent to the unconstrained minimization of Lagrangian J:

$$J = D + \sum_{k=1}^{K} \boldsymbol{\gamma}(k)^T \left[R_X(k) - \sum_{i=1}^{M} \boldsymbol{r}_i \boldsymbol{l}_i(k) \right], \tag{28}$$

where $\boldsymbol{\gamma} = [\boldsymbol{\gamma}(1)\ \boldsymbol{\gamma}(2) \ldots, \boldsymbol{\gamma}(K)]$ denotes the $K \times K$ multiplier Lagrangian matrix, $R_X(k)$ and $l_i(k)$ respectively denote the kth column of R_X and the kth element of $\boldsymbol{l}_i$. By setting $\nabla_{r_i} J = \boldsymbol{0}$, we obtain the condition:

$$\nabla_{r_i} J = -2\boldsymbol{l}_i^T + 2p_i \boldsymbol{r}_i^T - \sum_{k=1}^{K} \gamma(k)^T l_i(k) = \boldsymbol{0}. \tag{29}$$

Noting that

$$\sum_{k=1}^{K} \gamma(k)^T l_i(k) = \boldsymbol{l}_i^T \boldsymbol{\gamma}, \tag{30}$$

we obtain

$$\boldsymbol{l}_i^T \gamma = 2p_i \boldsymbol{r}_i^T - 2\boldsymbol{l}_i^T, \tag{31}$$

or

$$\boldsymbol{l}_i^T = 2p_i \boldsymbol{r}_i^T (\gamma + 2I)^{-1}. \tag{32}$$

Plugging (32) in (27), we obtain:

$$\gamma = 2R_X^{-1} R_Q - 2I. \tag{33}$$

Plugging (33) in (32), we have

$$\boldsymbol{r}_i = R_Q {R_X}^{-1} \boldsymbol{l}_i / p_i. \tag{34}$$

Note that $\boldsymbol{l}_i/p_i$ is the MSE optimal reconstruction of an unconstrained quantizer that shares the same decision boundary with the constrained one, $\mathcal{P}_i$. Next, we show that $\mathcal{P}_i = \widehat{\mathcal{P}_i}$.

We note that the optimal boundaries of the optimal constrained quantizer must satisfy the following nearest neighbor rule for $\mathcal{P}_i$:

$$\begin{aligned} \boldsymbol{x} \in \mathcal{P}_i \Rightarrow & (\boldsymbol{x} - \boldsymbol{r}_i)^T (\boldsymbol{x} - \boldsymbol{r}_i) + \gamma (\boldsymbol{x} - \boldsymbol{r}_i)^T \boldsymbol{x} \\ & \leq (\boldsymbol{x} - \boldsymbol{r}_j)^T (\boldsymbol{x} - \boldsymbol{r}_j) + \gamma (\boldsymbol{x} - \boldsymbol{r}_j)^T \boldsymbol{x} \ \ \forall j \neq i. \end{aligned} \tag{35}$$

Adding $\frac{1}{4} \boldsymbol{x}^T \gamma^T \gamma \boldsymbol{x}$ to both sides, we have

$$\begin{aligned} \boldsymbol{x} \in \mathcal{P}_i \Rightarrow & \left(\left(I + \frac{1}{2}\gamma \right) \boldsymbol{x} - \boldsymbol{r_i} \right)^T \left(\left(I + \frac{1}{2}\gamma \right) \boldsymbol{x} - \boldsymbol{r_i} \right) \\ & \leq \left(\left(I + \frac{1}{2}\gamma \right) \boldsymbol{x} - \boldsymbol{r_j} \right)^T \left(\left(I + \frac{1}{2}\gamma \right) \boldsymbol{x} - \boldsymbol{r_j} \right) \ \ \forall j \neq i. \end{aligned} \tag{36}$$

Noting that $(I + \frac{1}{2}\gamma) = R_X{}^{-1} R_Q$ and using (34), we rewrite (36) as

$$\boldsymbol{x} \in \mathcal{P}_i \Rightarrow (\boldsymbol{x} - \boldsymbol{r_i})^T (\boldsymbol{x} - \boldsymbol{r_i}) \leq (\boldsymbol{x} - \boldsymbol{r_j})^T (\boldsymbol{x} - \boldsymbol{r_j}) \ \forall j \neq i \tag{37}$$

which is the unconstrained nearest neighbor rule and hence

$$\mathcal{P}_i = \widehat{\mathcal{P}_i} \ \forall i \tag{38}$$

which implies that

$$\boldsymbol{l}_i = p_i \hat{\boldsymbol{r}}_i. \tag{39}$$

Plugging (39) in (27), we obtain

$$\hat{R}_Q = R_X R_Q^{-1} R_X \tag{40}$$

and finally using (40), (34) and (39), we obtain

$$\boldsymbol{r}_i = R_X \hat{R}_Q^{-1} \hat{\boldsymbol{r}}_i. \tag{41}$$

The proof for variable rate follows similar lines, with the only modification that we now have to account for the rate term $R = -\sum_{i=1}^{M} p_i \log p_i$. The uncorrelatedness constraint is identical to the one in fixed rate, hence the overall Lagrangian cost can be expressed as:

$$J = D + \sum_{k=1}^{M} \boldsymbol{\gamma}(k)^T \left[R_X(k) - \sum_{i=1}^{M} \boldsymbol{r}_i \boldsymbol{l}_i(k) \right] - \lambda \sum_{i=1}^{M} p_i \log p_i \tag{42}$$

for some $\lambda \in \mathbb{R}^+$. By setting $\nabla_{r_i} J = \mathbf{0}$, noting that the term $\sum_{i=1}^{M} p_i \log p_i$ does not depend on r_i and hence, following the same steps, we obtain (34). Towards showing (38), we again consider the nearest neighbor rule, for variable rate case (see e.g., [39]): if $\boldsymbol{x} \in \mathcal{P}_i$, then the following must hold for all $j \neq i$:

$$\begin{aligned}(\boldsymbol{x} - \boldsymbol{r}_i)^T (\boldsymbol{x} - \boldsymbol{r}_i) - \lambda p_i \log p_i + \gamma (\boldsymbol{x} - \boldsymbol{r}_i)^T \boldsymbol{x} \\ \leq (\boldsymbol{x} - \boldsymbol{r}_j)^T (\boldsymbol{x} - \boldsymbol{r}_j) - \lambda p_j \log p_j + \gamma (\boldsymbol{x} - \boldsymbol{r}_j)^T \boldsymbol{x}.\end{aligned} \tag{43}$$

Again, following the same steps that led to (37), we show that (43) is equivalent to the nearest neighbor rule of the optimal unconstrained quantizer and hence (38) holds. ■

Remark 1: The second part of Theorem 1 holds more generally, not necessarily for the optimal quantizers. Given a quantizer $\hat{Q}_K$ with reconstruction points chosen to minimize distortion for some (not necessarily optimal) partition $\mathcal{P}_i$, the minimum distortion quantizer that renders reconstruction error orthogonal to the source and shares the same partition $\mathcal{P}_i$ is obtained by scaling the reconstruction points as $\boldsymbol{r}_i = R_X \hat{R}_Q^{-1} \hat{\boldsymbol{r}}_i$. This is due to the fact that the assumption that $\mathcal{P}_i$ is the optimal partition was not used in any of the steps in deriving (34), hence (34) holds for any $\mathcal{P}_i$. Given that the partition is not changed, $\mathcal{P}_i = \widehat{\mathcal{P}}_i$, we obtain $\boldsymbol{r}_i = R_X \hat{R}_Q^{-1} \hat{\boldsymbol{r}}_i$.

Corollary 1: The reconstruction error covariances of the optimal constrained and unconstrained quantizers satisfy the following:

$$\hat{R}_E^{-1} = R_X^{-1} + R_E^{-1}. \tag{44}$$

Proof: Note that error is orthogonal to reconstructions

$$\hat{R}_E = R_X - \hat{R}_Q, \tag{45}$$

for the unconstrained optimal quantizer. Plugging (40) in (45), we have

$$\hat{R}_E = R_X - R_X R_Q{}^{-1} R_X. \tag{46}$$

Reconstruction error of the optimal constrained quantizer is orthogonal to the source:

$$R_Q = R_X + R_E. \tag{47}$$

Invoking matrix inversion lemma, see e.g., [40], we obtain

$$(R_X + R_E)^{-1} = R_X^{-1} - R_X^{-1} (R_X^{-1} + R_E{}^{-1})^{-1} R_X^{-1}. \tag{48}$$

Plugging (48) in (46), we obtain (44). ■

B. Constrained Randomized Quantizer

Due to the effect of companding, the nonuniform randomized quantizer we derived in Section III 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 $\mathbb{E}\{xw(g(x) + n)\} = \mathbb{E}\{x^2\}$.

$$J_c = J + \lambda_c \mathbb{E}[x^2 - xw(g(x) + n)] \tag{49}$$

where $J = J_v$ in the case of variable rate and $J = J_f$ for fixed rate. Let $g(\,\cdot\,)$ and $w(\,\cdot\,)$ be the compressor and expander mappings of the unconstrained optimal randomized quantizer. Let $g_c(\,\cdot\,)$ and $w_c(\,\cdot\,)$ denote the optimal mappings subject to the constraint that the reconstruction error be uncorrelated with the source. In the following, we present the relationship between the optimal constrained and unconstrained randomized quantizers.

Theorem 2: For both fixed and variable rate,

$$g_c(x) = g(x), w_c(y) = \frac{\sigma_X^2}{\mathbb{E}\{w^2(y)\}} w(y) \tag{50}$$

a.e. in x and y.

Proof: We focus on fixed rate as the variable rate case follows directly from the fixed rate proof. The optimal expander $w(\,\cdot\,)$ is no longer the standard conditional expectation, since it is impacted by the constraint. Towards finding the optimal $w_c(\,\cdot\,)$, we apply the standard method in variational calculus [34]: The following must hold

$$\left.\frac{\partial}{\partial \epsilon}\right|_{\epsilon=0} J_c[w_c(y) + \epsilon \eta(y)] = 0 \tag{51}$$

for any perturbation function $\eta(\,\cdot\,)$. Plugging (49) in (51), and evaluating the derivative, we have

$$\begin{aligned}&\left.\frac{\partial}{\partial \epsilon}\right|_{\epsilon=0} J_c[w_c(y) + \epsilon \eta(y)] \\ &= \int \left\{ \int (-2(x - w_c(y)) - \lambda_c x) f_X(x) f_N(y - g_c(x)) \right\} \eta(y) dy \\ &= 0\end{aligned} \tag{52}$$

for all $\eta(y)$. The quantization error is denoted as $N \sim U[-\Delta/2, \Delta/2]$. Equality for all admissible variation functions, $\eta(\cdot)$, requires the expression in braces to vanish *a.e.*. This gives the necessary condition for optimality as:

$$\begin{aligned} w_c(y) &= \left(1+\frac{\lambda_c}{2}\right) \frac{\int x f_X(x) f_N(y-g_c(x))dx}{\int f_X f_N(y-g_c(x))(x)dx} \\ &= \left(1+\frac{\lambda_c}{2}\right) w_s(y) \end{aligned} \tag{53}$$

where $w_s(\cdot)$ is the optimal unconstrained expander when $g_c(\cdot)$ is used as the compressor. Next, we find λ_c. We first note that

$$\mathbb{E}\{(X - w_c(g(X)+N))X\} = 0 \tag{54}$$

or

$$\mathbb{E}\{(w_c(g(X)+N))X\} = \sigma_X^2. \tag{55}$$

From orthogonality principle, we also have

$$\mathbb{E}\{(X - w_s(g(X)+N))w_s(g(X)+N)\} = 0. \tag{56}$$

Plugging (53) in (55) and (56), we obtain

$$\lambda_c = 2\frac{\sigma_X^2}{\mathbb{E}\left\{w_s^2\right\}} - 2. \tag{57}$$

Towards deriving the update rule for $g_c(\cdot)$, we perturb the overall cost in $g_c(\cdot)$,

$$\left.\frac{\partial}{\partial \epsilon}\right|_{\epsilon=0} J_c[g_c(x)+\epsilon\eta(x)] = 0 \tag{58}$$

which yields,

$$\begin{aligned} &\left.\frac{\partial}{\partial \epsilon}\right|_{\epsilon=0} J_c[g_c(x)+\epsilon\eta(x)] \\ &= \int \left\{ \int (-2(x-w_c(g_c(x)+n))-\lambda_c x)))w_c'(g_c(x)+n) \right. \\ &\qquad \left. f_N(n)dn \right\} \eta(x) f_X(x)dx \\ &= 0 \end{aligned}$$

for all $\eta(x)$. Equality for all admissible variation functions, $\eta(\cdot)$, requires the expression in braces to vanish *a.e.*, i.e.,

$$\begin{aligned} \int (-2(x-w_c(g_c(x)+n))-\lambda_c x) \\ \times w_c'(g_c(x)+n) f_N(n)dn = 0 \end{aligned} \tag{59}$$

almost everywhere in x. Plugging (53) in (59), we have

$$\int (x - w_s(g_c(x)+n))w_s'(g_c(x)+n) f_N(n)dn = 0 \tag{60}$$

which is precisely the necessary condition to be satisfied by the optimal unconstrained compressor, $g(\cdot)$. Hence,

$$g_c(x) = g(x) \tag{61}$$

a.e in x which yields $w_s(y) = w(y)$ and hence

$$w_c(y) = \frac{\sigma_X^2}{\mathbb{E}\{w^2(y)\}} w(y) \tag{62}$$

a.e in y. ■

Remark 2: Surprisingly, the optimally constrained compressor mapping remains unchanged (from the unconstrained optimal compressor) and the only modification to the expander mapping is simple scaling. This result parallels Theorem 1 which shows the decision boundaries, determined by the compressor $g(\cdot)$ do not change and all reconstructions are scaled by the same adjustment factor.

V. Asymptotic Analysis

A. Rate-Distortion Functions

To quantify theoretically how much a source[3] can be compressed under the independent/uncorrelated reconstruction error constraint, we define two rate-distortion functions in which we respectively constrain the reconstructions error to be i) uncorrelated with the source: $R_U(D)$, and ii) independent of the source: $R_I(D)$.

Assume that we have source X with density $f_X(\cdot)$ that produces the independent identically distributed (i.i.d.) sequence $X_1, X_2, \ldots, X_n$ denoted as X^n. Similarly, let $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$ be the reconstruction sequence, denoted as $\hat{X}^n$. Let $S^n = X^n - \hat{X}^n$ be the i.i.d. sequence of reconstruction errors with marginal density $f_S(\cdot)$. Let $f_{XS}(x,s)$ denote joint distribution of X and S and $d(X^n, \hat{X}^n)$ denote the distortion measure between sequences X^n and $\hat{X}^n$ defined as

$$d(X^n, \hat{X}^n) = \frac{1}{n}\sum_{i=1}^{n} d(X_i, \hat{X}_i). \tag{63}$$

Let us recall the classical rate-distortion result in information theory:

Rate-Distortion Theorem: (eg. [32]): Let $R(D)$ be the infimum of all achievable rates R with average distortion $\mathbb{E}[d(X^n, X^n + S^n)] \leq D$ as $n \to \infty$. Then,

$$R(D) = \inf_{S:\mathbb{E}[d(X,X+S)]\leq D} I(X; X+S). \tag{64}$$

We next focus on our problem: let $R_U(D)$ be the infimum of all achievable rates R with average distortion

$$\lim_{n\to\infty} \mathbb{E}[d(X^n, X^n + S^n)] \leq D \tag{65}$$

subject to the constraint

$$\mathbb{E}[X_i S_i] = 0, \forall i. \tag{66}$$

Similarly, let $R_I(D)$ be the infimum of all achievable rates R with average distortion $\mathbb{E}[d(X^n, X^n + S^n)] \leq D$ subject to the constraint S_i is independent of X_i for all i, as $n \to \infty$. Then, we have the following result characterizing the fundamental limits of source compression under the constraints that reconstruction error is uncorrelated with or independent of the source.

Theorem 3:

$$R_U(D) = \inf_{\substack{S:\mathbb{E}[d(X,X+S)]\leq D \\ \mathbb{E}[XS]=0}} I(X; X+S) \tag{67}$$

$$R_I(D) = \inf_{\substack{S:\mathbb{E}[d(X,X+S)]\leq D \\ \mathcal{D}(f_{XS}(X,S)\,\|\,f_X(X)f_S(S))=0}} I(X; X+S) \tag{68}$$

[3] The notation in this section is limited to scalar sources for simplicity, it is trivial to extend the results to vector sources albeit with more complicated notation.

Proof: Consider the distortion measures

$$d_U(x, x+s) = d(x, x+s) + \beta \|xs\| \tag{69}$$

$$d_I(x, x+s) = d(x, x+s) + \beta \log \frac{f_{XS}(x,s)}{f_X(x) f_S(s)} \tag{70}$$

for some $\beta > 0$.

We next consider the rate-distortion functions (denoted $R_U^*(D)$ and $R_I^*(D)$) associated with these distortion measures. By replacing d with d_U and d_I in (64), we obtain the following expressions:

$$R_U^*(D) = \inf_{S:\mathbb{E}[d_U(X,X+S)]\le D} I(X; X+S). \tag{71}$$

$$R_I^*(D) = \inf_{S:\mathbb{E}[d_I(X,X+S)]\le D} I(X; X+S). \tag{72}$$

We note that the achievability and the converse proofs are straightforward extensions of the standard achievability and the converse proofs for regular rate distortion function.

We next consider the distortion measures d_U and d_I and associated rate-distortion functions $R_U^*(D)$ and $R_I^*(D)$ when $\beta \to \infty$. As $\beta \to \infty$, $\mathbb{E}[d_U(X^n, X^n + S^n)] \le D$ implies $\mathbb{E}[d(X^n, X^n + S^n)] \le D$ while $\mathbb{E}[X_i S_i] \to 0$ for all i. Similarly, as $\beta \to \infty$, $\mathbb{E}[d_I(X^n, X^n + S^n)] \le D$ implies $\mathbb{E}[d(X^n, X^n + S^n)] \le D$ under the constraint that X_i and S_i are asymptotically independent for all i. Hence, as $\beta \to \infty$, the distortion measures under consideration satisfy the respective requirements of uncorrelatedness or independence, i.e., $R_U(D) = R_U^*(D)$ and $R_I(D) = R_I^*(D)$.

Hence, (67) and (68) are indeed the information theoretic characterization of the limits of encoding a source with uncorrelated and independent reconstruction error respectively. ■

B. Gaussian Vector Source With MSE Distortion

In this section, we examine a special case where the source is vector Gaussian and the distortion measure is squared error. We start with an auxiliary lemma without proof (see eg. [32] for the proof).

Lemma 1 ([32]): Let $\boldsymbol{S} \sim f_S$ and $\boldsymbol{S}_G \sim f_{S_G}$ be random vectors in $\mathbb{R}^K$ with the same covariance matrix R_S. If $\boldsymbol{S}_G \sim \mathcal{N}(\boldsymbol{0}, R_S)$ then

$$\mathbb{E}_{S_G}[\log(f_{S_G}(\boldsymbol{S}))] = \mathbb{E}_S[\log(f_{S_G}(\boldsymbol{S}))] \tag{73}$$

where $\mathbb{E}_{S_G}$ and $\mathbb{E}_S$ denote the expectations with respect to f_{S_G} and f_S respectively.

Let us present a key lemma regarding the mutual information of two correlated random vectors constrained to have a fixed cross covariance matrix.

Lemma 2: Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, R_X)$ and $\boldsymbol{S}_G \sim \mathcal{N}(\boldsymbol{0}, R_S)$ be jointly Gaussian random vectors in $\mathbb{R}^K$. Let $\boldsymbol{S} \in \mathbb{R}^K$ and $\boldsymbol{S}_G$ have the same covariance matrix, R_S and the same cross covariance matrix with $\boldsymbol{X}$, R_{SX}. Then,

$$I(\boldsymbol{X}, \boldsymbol{X}+\boldsymbol{S}) \ge I(\boldsymbol{X}, \boldsymbol{X}+\boldsymbol{S}_G) \tag{74}$$

with equality if and only if $\boldsymbol{S} \sim \mathcal{N}(\boldsymbol{0}, R_S)$.

Proof: Consider $\gamma = I(\boldsymbol{X}, \boldsymbol{X}+\boldsymbol{S}) - I(\boldsymbol{X}, \boldsymbol{X}+\boldsymbol{S}_G)$. Plugging the expressions, we obtain:

$$\gamma = h(\boldsymbol{X}|\boldsymbol{S_G}+\boldsymbol{X}) - h(\boldsymbol{X} \,|\, \boldsymbol{S}+\boldsymbol{X}) \tag{75}$$

Noting that $h(\boldsymbol{X} \,|\, \boldsymbol{S_G}+\boldsymbol{X}) = h(\boldsymbol{S_G} \,|\, \boldsymbol{S_G}+\boldsymbol{X})$ and $h(\boldsymbol{X} \,|\, \boldsymbol{S}+\boldsymbol{X}) = h(\boldsymbol{S} \,|\, \boldsymbol{S}+\boldsymbol{X})$ and plugging $\boldsymbol{Y} = \boldsymbol{X}+\boldsymbol{S}$ and $\boldsymbol{Y}_G = \boldsymbol{X}+\boldsymbol{S}_G$, we obtain:

$$\gamma = h(\boldsymbol{S}_G|\boldsymbol{Y}_G) - h(\boldsymbol{S}|\boldsymbol{Y}) = \tag{76}$$

$$\iint \Big\{ f_{S,Y}(\boldsymbol{s},\boldsymbol{y}) \times \log f_{S\,|\,Y}(\boldsymbol{s},\boldsymbol{y}) - f_{S_G,Y_G}(\boldsymbol{s},\boldsymbol{y}) \log f_{S_G\,|\,Y_G}(\boldsymbol{s},\boldsymbol{y}) \Big\} d\boldsymbol{s}d\boldsymbol{y} \tag{77}$$

Using Lemma 1 and the fact that the joint distribution f_{S_G,Y_G} is Gaussian:

$$= \iint f_{S,Y}(\boldsymbol{s},\boldsymbol{y}) \left[\log f_{S\,|\,Y}(\boldsymbol{s},\boldsymbol{y}) - \log f_{S_G\,|\,Y_G}(\boldsymbol{s},\boldsymbol{y})\right] d\boldsymbol{s}d\boldsymbol{y} \tag{78}$$

$$= \int f_Y(\boldsymbol{y}) \int f_{S\,|\,Y}(\boldsymbol{s},\boldsymbol{y}) \log \frac{f_{S\,|\,Y}(\boldsymbol{s},\boldsymbol{y})}{f_{S_G\,|\,Y_G}(\boldsymbol{s},\boldsymbol{y})} d\boldsymbol{s}d\boldsymbol{y} \tag{79}$$

$$= \mathcal{D}(f_{S\,|\,Y}, f_{S_G\,|\,Y_G}) \tag{80}$$

$\mathcal{D} \ge 0$ with equality if and only if $\boldsymbol{S} \sim \mathcal{N}(\boldsymbol{0}, R_S)$. ■

Next, we present our main result on this topic:

Theorem 4: For a Gaussian vector source $\boldsymbol{X} \in \mathbb{R}^K$ and MSE distortion $d(\boldsymbol{x}, \hat{\boldsymbol{x}}) = (\boldsymbol{x} - \hat{\boldsymbol{x}})^T (\boldsymbol{x} - \hat{\boldsymbol{x}})$, the following holds:

$$R_I(D) = R_U(D). \tag{81}$$

Proof: Generally, $R_I(D) \ge R_U(D)$, since independent reconstruction error is also uncorrelated. Note that the uncorrelated error constraint dictates $R_{SX} = 0$ and the distortion constraint is $Tr(R_S) = D$. Lemma 2 states that under these constraints, for a Gaussian source, Gaussian reconstruction error minimizes the mutual information between the source and the reconstruction, i.e., $I(\boldsymbol{X}_G, \boldsymbol{X}_G + \boldsymbol{S})$ achieves its minimum when $\boldsymbol{S} \sim \mathcal{N}(\boldsymbol{0}, R_S)$. Then, $\boldsymbol{X}_G$ and $\boldsymbol{S}$ are uncorrelated and jointly Gaussian and are, thereby, also independent. ■

We next pose the question: Are there cases where the best possible vector quantizer at asymptotically high dimension that renders the reconstruction error uncorrelated with the source, is necessarily a randomized one? The next corollary answers in the affirmative, as is proved by the Gaussian case.

Corollary 2: For a Gaussian source, at asymptotically high quantizer dimension, the quantizer that achieves minimum distortion subject to the uncorrelated error constraint is necessarily a randomized one.

Proof: From Theorem 4, the reconstruction error for the Gaussian source subject to the uncorrelated error constraint is independent of the source. No deterministic quantizer can render the quantization noise independent from the source by definition; hence, the optimal quantizer is a randomized one. ■

Note that our results hold only asymptotically, and it is still an open question whether or not they hold at finite dimensions. The numerical results in the next section support the thesis that randomized quantizers are better at finite dimensions.

VI. 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 to 3) i.e., we numerically neglected the tails of the Gaussian. In this paper, we proposed three new quantizers:

Quantizer 1: Unconstrained randomized quantizer. This quantizer does not render the reconstruction error uncorrelated with the source.

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.

Figs. 3 and 4 demonstrate the performance comparisons among quantizers for fixed and variable rates respectively. 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 diminishes at high rates, while at low rates the difference is quite significant. This is theoretically anticipated 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 uniform quantizer renders error *independent* of the source but at significant distortion penalty.

An additional benefit of the proposed random quantizers pertains to the correlation of the reconstruction errors when correlated sources are quantized. The conventional dithered quantizer renders quantization error independent of the source hence, when two correlated sources are quantized with a dithered quantizer, the reconstruction errors are uncorrelated. For deterministic quantizers (Quantizer 3 and the optimal quantizer), the reconstruction error is a deterministic function of the source hence, intuitively randomized quantizers are expected to have lower reconstruction error correlation. Figs. 5 and 6 demonstrate the correlation of the reconstruction error for different values of source correlation for a bivariate Gaussian source for fixed and variable rate quantization respectively. These numerical results illustrate this intuitive conclusion: randomization is significantly useful in decreasing the correlations of reconstruction errors. Specifically, the constrained randomized quantizer (Quantizer 2) yields extremely low error correlation, very close to that of the conventional dithered quantization. This property is particularly useful in practical applications such as image-video compression where white reconstruction error is preferred due to audio-visual considerations, see eg. deblocking filters commonly used in video coding [41]. Also note that, the unconstrained randomized quantizer (Quantizer 1) significantly decreases the error correlation compared to the optimal quantizer, with negligible sacrifice in rate distortion performance. Hence, this statistical benefit of randomization comes with no significant penalty.

Numerical comparisons show that the proposed quantization schemes can significantly impact the design of compression systems such as [27], [28] where quantization error is assumed to be uncorrelated with the source. Note that the constrained randomized quantization satisfies this assumption precisely 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, as an alternative to the conventional dithered quantization that satisfies these assumptions at considerable performance cost, we derived additional quantization schemes that satisfy those assumptions: constrained deterministic quantization and constrained nonuni-

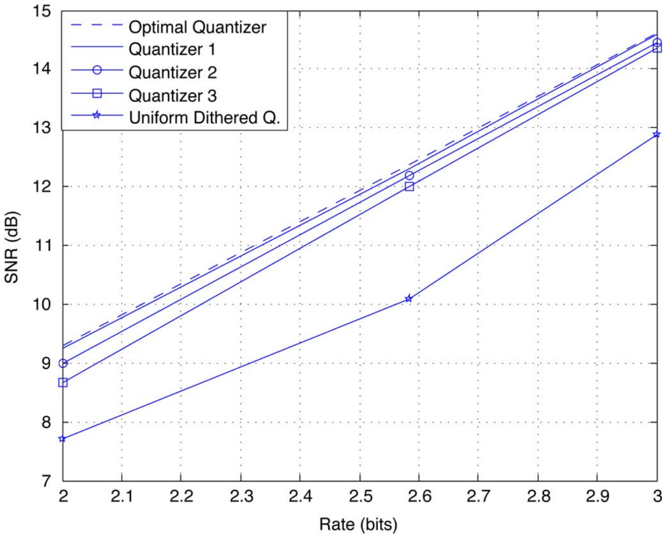

Fig. 3. Performance comparison in terms of SNR versus rate for fixed rate quantization.

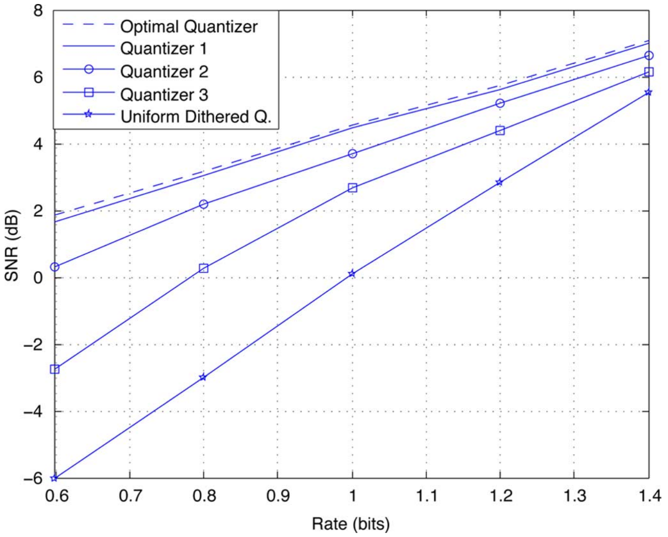

Fig. 4. Performance comparison in terms of SNR versus rate for variable rate quantization.

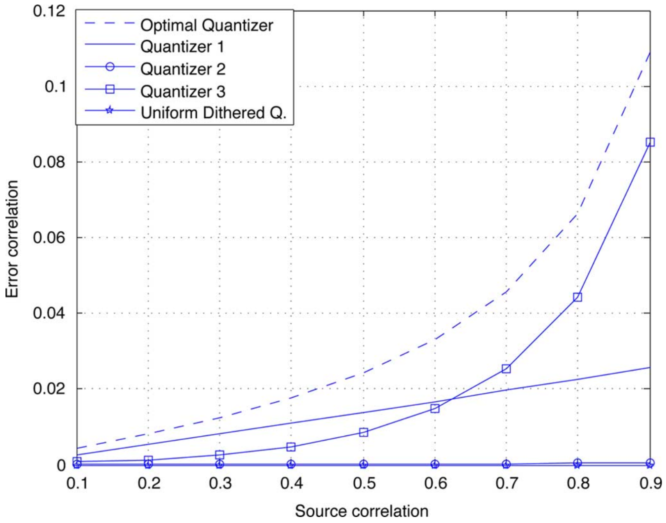

Fig. 5. Correlation of the reconstruction error versus source correlation for fixed rate quantization at rate $R = 2$ bits/sample.

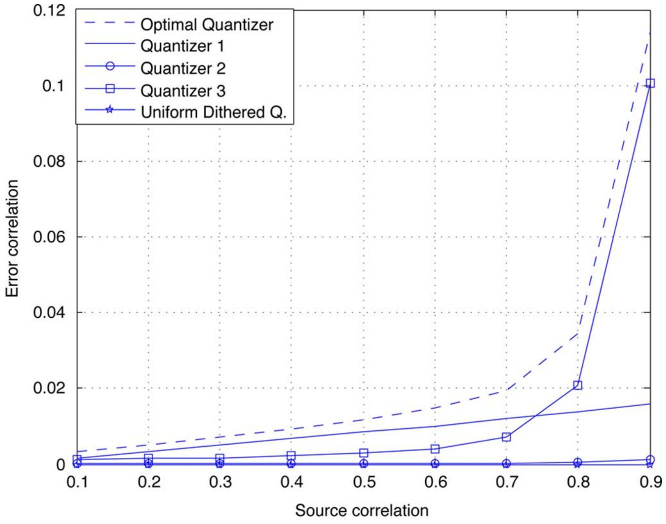

Fig. 6. Correlation of the reconstruction error versus source correlation for variable rate quantization at rate $R = 1.4$ bits/sample.

form 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.

While it is difficult to prove the strict superiority of these new quantizers over the conventional dithered quantizer, we numerically show it, for both fixed and variable rate quantizers, in Figs. 3 and 4. The numerical results motivate a theoretical proposition: the optimal vector quantizer that renders the reconstruction error orthogonal to the source is necessarily randomized. While we proved this result at asymptotically high dimensions, it remains a conjecture at finite dimensions, based on the numerical results in this section.

VII. 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 the constrained deterministic quantizer proposed in this paper, while still satisfying the requirement that the reconstruction error be uncorrelated with the source. Moreover, the proposed randomized quantizers significantly reduce the correlations across reconstruction errors when correlated samples, i.e., sources with memory, are quantized. The design complexity of proposed nonuniform dithered scalar quantizers is not significantly different from that of the optimal conventional quantizers. We also showed that at asymptotically high dimensions, the MSE optimal vector quantizer designed for a vector Gaussian source, which renders the reconstruction error uncorrelated with the source, must be a randomized quantizer. As future work, we will investigate the applicability of this result to a broader class of problems (e.g., non-Gaussian sources, finite dimensional vector quantizers, different statistical constraints) where random encoding is not merely a tool to derive rate-distortion bounds, but a necessary element in practical systems approaching such bounds.

Appendix I
Functional Derivatives

In this section, we derive ∇J for fixed and variable rate cases. We repeatedly use the Leibniz rule, which is stated here for completeness.

Leibniz Rule: Let $f(u,y) : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a function whose partial derivative with respect to u, exists and is continuous everywhere in u. Also assume that $a(\cdot), b(\cdot)$ are functions $\mathbb{R} \rightarrow \mathbb{R}$ which are differentiable everywhere. Then,

$$\frac{d}{du}\int_{a(u)}^{b(u)} f(u,y)dy = \int_{a(u)}^{b(u)} \frac{df(u,y)}{du}dy + f(u,b(u))b'(u) - f(u,a(u))a'(u) \quad (82)$$

Fixed Rate: Recall that $J_f = D_g + D_{ol}$. Applying Leibniz rule, we get

$$\begin{aligned}
&\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} D_g(g(x)+\epsilon\eta(x)) \\
&= \frac{-2}{\Delta}\int_{-\Delta/2}^{\Delta/2}\int_{g^{-1}(-\Delta T)}^{g^{-1}(\Delta T)} [x - w(g(x)+n)]w'(g(x)+n)f_X(x)\eta(x)dxdn \\
&+ \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} [g^{-1}(\Delta T) - w(g(\Delta T)+n)]^2 (g^{-1})'(\Delta T)\eta(\Delta T)dn \\
&- \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} [g^{-1}(-\Delta T) - w(g(-\Delta T)+n)]^2 (g^{-1})'(-\Delta T)\eta(-\Delta T)dn.
\end{aligned} \quad (83)$$

$$\left.\frac{\partial J_v(g(x)+\epsilon\eta(x))}{\partial\epsilon}\right|_{\epsilon=0}$$
$$=\frac{-2}{\Delta}\int_{-\Delta/2}^{\Delta/2}\int[x-w(g(x)+n)]w'(g(x)+n)f_X(x)\eta(x)dndx-\frac{\lambda}{\Delta}\int\left.\frac{\partial\left[F_X\left((g+\epsilon\eta)^{-1}\left(x+\frac{\Delta}{2}\right)\right)-F_X\left((g+\epsilon\eta)^{-1}\left(x-\frac{\Delta}{2}\right)\right)\right]}{\partial\epsilon}\right|_{\epsilon=0}$$
$$\times\log\left(\frac{1}{\Delta}\left[F_X\left(g^{-1}\left(x+\frac{\Delta}{2}\right)\right)-F_X\left(g^{-1}\left(x-\frac{\Delta}{2}\right)\right)\right]\right)dx. \tag{92}$$
$$=\frac{-2}{\Delta}\int\int[x-h(g(x)+n)]w'(g(x)+n)f_X(x)\eta(x)dxdn-\frac{\lambda}{\Delta}\int\left(\log\frac{1}{\Delta}\left[F_X\left(g^{-1}\left(x+\frac{\Delta}{2}\right)\right)-F_X\left(g^{-1}\left(x-\frac{\Delta}{2}\right)\right)\right]\right)$$
$$\times\left(\frac{f_X\left(g^{-1}\left(x+\frac{\Delta}{2}\right)\right)}{g'\left(x+\frac{\Delta}{2}\right)}\eta\left(x+\frac{\Delta}{2}\right)-\frac{f_X\left(g^{-1}\left(x-\frac{\Delta}{2}\right)\right)}{g'\left(x-\frac{\Delta}{2}\right)}\eta\left(x-\frac{\Delta}{2}\right)\right)dx. \tag{93}$$

Let us now consider D_{ol} given in (21). Applying Leibniz rule:

$$\left.\frac{d}{d\epsilon}\right|_{\epsilon=0}D_{ol}(g(x)+\epsilon\eta(x))$$
$$=\frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2}[g^{-1}(-\Delta T)-w(g(-\Delta T)+n)]^2(g^{-1})'(-\Delta T)\eta(-\Delta T)dn$$
$$-\frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2}[g^{-1}(\Delta T)-w(g(\Delta T)+n)]^2(g^{-1})'(\Delta T)\eta(\Delta T)dn. \tag{84}$$

Clearly, the last two terms in (83) are cancelled by (84). Hence,

$$\left.\frac{\partial}{\partial\epsilon}\right|_{\epsilon=0}J_f[g(x)+\epsilon\eta(x)]$$
$$=\frac{-2}{\Delta}\int_{-\Delta/2}^{\Delta/2}\int_{g^{-1}(-\Delta T)}^{g^{-1}(\Delta T)}[x-w(g(x)+n)]w'(g(x)+n)f_X(x)\eta(x)dxdn$$

which implies

$$\nabla J_f=\frac{-2}{\Delta}\int_{-\Delta/2}^{\Delta/2}[x-w(g(x)+n)]w'(g(x)+n)f_X(x)dn.$$

Variable rate: First, we obtain the density of $Y=g(X)+N$ where $N\sim U[-\Delta/2,\Delta/2]$ in $f_X(\cdot)$ and $g(\cdot)$. In general,

$$F_Y(y)=\int\int_{y\geq g(x)+n}f_{X,N}(x,n)dxdn \tag{85}$$

where F_Y is the cumulative distribution function of Y. Since $g(\cdot)$ is monotonically increasing and N is independent of X, we have

$$F_Y(y)=\frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2}\int_{g^{-1}(y-n)}^{\infty}f_X(x)dxdn \tag{86}$$
$$=\frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2}[F_X(\infty)-F_X(g^{-1}(y-n))]dn. \tag{87}$$
$$f_Y(y)=\frac{dF_Y(y)}{dy} \tag{88}$$
$$=\frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2}-(g^{-1})'(y-n)f_x(g^{-1}(y-n))dn \tag{89}$$
$$=\frac{1}{\Delta}\left[F_X\left(g^{-1}\left(y+\frac{\Delta}{2}\right)\right)\right.$$
$$\left.-F_X\left(g^{-1}\left(y-\frac{\Delta}{2}\right)\right)\right]. \tag{90}$$

Next, we derive the update rule for the variable rate:

$$J_v=\int_{-\Delta/2}^{\Delta/2}\int[x-w(g(x)+n)]^2f_X(x)dxdn+\lambda R_v. \tag{91}$$

Plugging (90) and (23) in (91), we have [see (92)-(93), shown at the top of the page]. By changing variables and manipulating (93):

$$\nabla J_v=\frac{-2}{\Delta}\int_{-\Delta/2}^{\Delta/2}[x-w(g(x)+n)]w'(g(x)+n)f_X(x)dn$$
$$-\frac{\lambda}{\Delta}\left(\frac{f_X(g^{-1}(x))}{g'(x)}\right)\log\left(\frac{F_X(g^{-1}(x))-F_X(g^{-1}(x-\Delta))}{F_X(g^{-1}(x+\Delta))-F_X(g^{-1}(x))}\right).$$

References

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

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

[3] K. C. Pohlmann and K. C. Pohlman, *Principles of Digital Audio*. New York, NY, USA: McGraw-Hill, 1995, vol. 4.

[4] J. Vanderkooy and S. P. Lipshitz, "Dither in digital audio," *J. Audio Eng. Soc.*, vol. 35, no. 12, pp. 966–975, 1987.

[5] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, and J. N. Wright, "A theory of nonsubtractive dither," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 499–516, 2000.

[6] M. Li, J. Klejsa, and W. B. Kleijn, “Distribution preserving quantization with dithering and transformation,” *IEEE Signal Process. Lett.*, vol. 17, no. 12, pp. 1014–1017, 2010.

[7] N. Saldi, T. Linder, and S. Yuksel, “Randomized quantization and optimal design with a marginal constraint,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2013.

[8] J. Ziv, “On universal quantization,” *IEEE Trans. Inf. Theory*, vol. 31, no. 3, pp. 344–347, 1985.

[9] R. M. Gray and T. G. Stockham, Jr., “Dithered quantizers,” *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 805–812, 1993.

[10] R. Zamir and M. Feder, “On universal quantization by randomized uniform/lattice quantizers,” *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pt. 2, pp. 428–436, 1992.

[11] R. Zamir and M. Feder, “Information rates of pre/post-filtered dithered quantizers,” *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1340–1353, 1996.

[12] C. E. Shannon, “The mathematical theory of information,” *Bell Syst. Tech. J.*, vol. 27, no. 6, pp. 379–423, 1949.

[13] J. Ostergaard and R. Zamir, “Multiple-description coding by dithered delta-sigma quantization,” *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4661–4675, 2009.

[14] R. Zamir, Y. Kochman, and U. Erez, “Achieving the Gaussian rate-distortion function by prediction,” *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3354–3364, 2008.

[15] U. Erez and R. Zamir, “Achieving $1/2\log(1+\mathrm{snr})$ on the AWGN channel with lattice encoding and decoding,” *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2293–2314, 2004.

[16] R. Zamir, “Lattices are everywhere,” in *Proc. IEEE Inf. Theory Appl. Workshop*, 2009, pp. 392–421.

[17] V. K. Goyal, “Scalar quantization with random thresholds,” *IEEE Signal Process. Lett.*, vol. 18, no. 9, pp. 525–528, 2011.

[18] W. Chou and R. M. Gray, “Dithering and its effects on sigma-delta and multistage sigma-delta modulation,” *IEEE Trans. Inf. Theory*, vol. 37, no. 3, pp. 500–513, 1991.

[19] I. Galton, “Granular quantization noise in a class of delta-sigma modulators,” *IEEE Trans. Inf. Theory*, vol. 40, no. 3, pp. 848–859, 1994.

[20] S. Pamarti, J. Welz, and I. Galton, “Statistics of the quantization noise in 1-bit dithered single-quantizer digital delta–sigma modulators,” *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 54, no. 3, pp. 492–503, 2007.

[21] J. N. Tsitsiklis, “Problems in decentralized decision making and computation,” LID, MIT, Cambridge, MA, USA, Tech. Rep., DTIC Document, 1984.

[22] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, “Randomized gossip algorithms,” *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[23] A. G. Dimakis *et al.*, “Gossip algorithms for distributed signal processing,” *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, 2010.

[24] A. Kashyap, T. Başar, and R. Srikant, “Quantized consensus,” *Automatica*, vol. 43, no. 7, pp. 1192–1203, 2007.

[25] T. C. Aysal, M. J. Coates, and M. G. Rabbat, “Distributed average consensus with dithered quantization,” *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4905–4918, 2008.

[26] S. Kar and J. M. F. Moura, “Distributed consensus algorithms in sensor networks: Quantized data and random link failures,” *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1383–1400, 2010.

[27] M. K. Mihcak, P. Moulin, M. Anitescu, and K. Ramchandran, “Rate-distortion-optimal subband coding without perfect-reconstruction constraints,” *IEEE Trans. Signal Process.*, vol. 49, no. 3, pp. 542–557, 2001.

[28] O. G. Guleryuz and M. T. Orchard, “On the DPCM compression of Gaussian autoregressive sequences,” *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 945–956, 2001.

[29] A. Gersho and R. M. Gray*, Vector Quantization and Signal Compression*. New York, NY, USA: Springer, 1992.

[30] E. Akyol and K. Rose, “Nonuniform dithered quantization,” in *Proc. IEEE Data Compression Conf.*, 2009, p. 435.

[31] E. Akyol and K. Rose, “On constrained randomized quantization,” in *Proc. IEEE Data Compress. Conf.*, 2012, pp. 212–222.

[32] T. M. Cover and J. A. Thomas*, Elements of Information Theory*. New York, NY, USA: Wiley, 1991.

[33] E. Akyol, K. Rose, and T. A. Ramstad, “Optimal mappings for joint source channel coding,” in *Proc. IEEE Inf. Theory Workshop*, 2010.

[34] D. G. Luenberger*, Optimization by Vector Space Methods*. New York, NY, USA: Wiley, 1969.

[35] S. Gadkari and K. Rose, “Robust vector quantizer design by noisy channel relaxation,” *IEEE Trans. Commun.*, vol. 47, no. 8, pp. 1113–1116, 1999.

[36] P. Knagenhjelm, “A recursive design method for robust vector quantization,” in *Proc. Int. Conf. Signal Process. Appl. Technol.*, 1992, pp. 948–954.

[37] K. Rose, “Deterministic annealing for clustering, compression, classification, regression, and related optimization problems,” *Proc. IEEE*, vol. 86, no. 11, pp. 2210–2239, 1998.

[38] A. Hjorungnes, “Optimal bit and power constrained filter banks,” Ph.D. dissertation, Norwegian Univ. of Sci. and Technol., Trondheim, Norway, 2000.

[39] P. A. Chou, T. Lookabaugh, and R. M. Gray, “Entropy-constrained vector quantization,” *IEEE Trans. Acoust, Speech, Signal Process.*, vol. 37, no. 1, pp. 31–42, 1989.

[40] R. A. Horn and C. R. Johnson*, Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[41] P. List *et al.*, “Adaptive deblocking filter,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 614–619, 2003.

Emrah Akyol (S’03–M’12) received the Ph.D. degree in 2011 from the University of California at Santa Barbara. From 2006 to 2007, he held positions at Hewlett-Packard Laboratories and NTT Docomo Laboratories, both in Palo Alto, CA, USA. Currently, Dr. Akyol is a postdoctoral researcher at UC Santa Barbara. His research focuses on networked source coding, joint source-channel coding, and low-delay communications with applications to optimization and control of smart grids.

Kenneth Rose (S’85–M’91–SM’01–F’03) received the Ph.D. degree in 1991 from the California Institute of Technology. He then joined the Department of Electrical and Computer Engineering, University of California at Santa Barbara, where he is currently a Professor. His main research activities are in the areas of information theory and signal processing, and include rate-distortion theory, source and source-channel coding, audio and video coding and networking, pattern recognition, and nonconvex optimization. He is interested in the relations between information theory, estimation theory, and statistical physics, and their potential impact on fundamental and practical problems in diverse disciplines.

Dr. Rose was co-recipient of the 1990 William R. Bennett Prize Paper Award of the IEEE Communications Society, as well as the 2004 and 2007 IEEE Signal Processing Society Best Paper Awards.