Staggered Quantizers for Perfect Perceptual Quality: A Connection between Quantizers with Common Randomness and Without

Ruida Zhou
Department of Electrical Engineering
University of California, Los Angeles
Los Angeles, CA
ruida@g.ucla.edu

Chao Tian
Department of Electrical and Computer Engineering
Texas A&M University
College Station, TX
chao.tian@tamu.edu

***Abstract*—The rate-distortion-perception (RDP) framework has attracted significant recent attention due to its application in neural compression. It is important to understand the underlying mechanism connecting procedures with common randomness and those without. Different from previous efforts, we study this problem from a quantizer design perspective. By analyzing an idealized setting, we provide an interpretation of the advantage of dithered quantization in the RDP setting, which further allows us to make a conceptual connection between randomized (dithered) quantizers and quantizers without common randomness. This new understanding leads to a new procedure for RDP coding based on staggered quantizers.**

I. Introduction

Compression plays an important role in the efficient representation of information content, particularly visual content. Traditionally, the tradeoff between the compression rate and the incurred distortion has been studied under two different but related frameworks: the quantization framework [1] and the rate-distortion theory [2] framework. In the former, the focus is on the design of quantizers that compress data samples one at a time (i.e., scalar quantization) or a few at a time (i.e., vector quantization), while the latter focuses on the fundamental limits of lossy compression by allowing an asymptotically large number of samples to be encoded together.

Largely driven by the recent emergence of the neural compression, the issue of perceptual quality has led to the study of the problem of rate-distortion-perception (RDP) tradeoff [3]–[10]. In this formulation, a new quality constraint, which was introduced to capture the perceptual quality loss due to compression, is further imposed in addition to the existing objective distortion constraint. Mathematically, this formulation [3] requires the probability distribution of the content after decompression to be close to that of the source content before compression; the case when the two distributions are exactly the same is often referred to as "perfect perceptual quality", which is our focus in this work.

The RDP problem has attracted significant recent research attention, and several studies in this area revealed that common randomness plays an important role in this setting [5], [6]. More precisely, the lack of common randomness can cause significant performance loss compared to methods that have such common randomness at their disposal, and this loss is particularly severe for scalar quantization. There are two known prevailing methods of introducing common randomness for RDP coding. The first is based on probabilistic sampling [11], and the second is through universal dithered quantization [12], [13]. The first approach requires the knowledge of a target joint distribution between the samples and the compressed version, and furthermore, involves a rather complex sampling procedure. The dither-based approach, on the other hand, is simpler to implement and thus more attractive, however, its architecture places an inherent constraint on the eventual probability distribution, and though widely used, it is not clear what actually makes it suitable for the RDP setting.

One piece of the puzzle has thus far been missing between the compression procedures without common randomness (e.g., scalar quantization with deterministic encoder) and those with a large amount of common randomness (dithered quantizers), particularly from a quantizer design perspective. That is, quantizers with deterministic encoders require no common randomness, and the dither-based approach will utilize common randomness on an uncountable set in a less transparent manner. What exactly is the underlying mechanism that lends the dither-based approach the advantage, and is there an effective procedure with an intermediate amount of common randomness? Although these questions have previously been studied under the rate-distortion framework with asymptotic large sample block size [14], the asymptotic nature of such analysis makes the mechanism rather opaque.

In this work, we develop a better understanding of these issues under the quantization framework. Using a decomposition perspective, we provide a new way to understand the mechanism from which procedures utilizing common randomness obtain the advantage. We first focus on an idealized setting on the unit circle, and provide a complete analysis of the performance. Based on these understandings, we provide a new approach to introduce common randomness using staggered quantizers. We further discuss the application of such an approach to other sources. It should be noted that staggered quantizers have been previously used for multiple description coding [15]–[17] which offered surprisingly competitive performance compared to more sophisticated approaches.

II. Backgrounds

A. Rate-distortion function and quantizers

Let the data source X be a real-valued random variable, with a distribution P_X on the alphabet $\mathcal{X}$. The reconstruction

alphabet is denoted as $\hat{\mathcal{X}}$. Given a distortion measure $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$, e.g., the squared error distortion $d(x, \hat{x}) = (x - \hat{x})^2$ when $\mathcal{X} = \hat{\mathcal{X}} = \mathbb{R}$, the (informational) rate-distortion function under a distortion constraint D is defined as

$$R(D) = \min_{P_{\hat{X}|X}:\mathbb{E}d(X,\hat{X})\leq D} I(X; \hat{X}),$$

where $I(\cdot;\cdot)$ is the mutual information function.

Rate-distortion theory deals with the setting when an infinite number of samples is allowed to be encoded together. In practice, samples are usually encoded one or few at a time, referred to as scalar quantization and vector quantization, respectively. In particular, a scalar quantizer consists of an encoding mapping $f : \mathcal{X} \to \mathbb{Z}$ which determines the representation index to assign to a sample, and a decoding function $g : \mathbb{Z} \to \hat{\mathcal{X}}$ which assigns a reconstruction point to each representation index. Therefore, $\hat{X} = g(f(X))$. Indices are allowed to be further entropy-coded, e.g., using Huffman code. When entropy coding is allowed, it is usually referred to as entropy-constrained scalar quantization (ECSQ), whereas when the number of quantization levels is fixed, it is usually referred to as fixed-rate quantization.

Universal dithered quantizer utilizes a uniform quantizer with stepsize Δ in the encoding and decoding process [18]. Different from classic deterministic quantizers, a random noise Z, independent of the data samples and uniformly distributed on the base interval $(-\Delta/2, \Delta/2]$, is available at both the encoder and the decoder. The noise Z is first added to the sample as $X + Z$, which is then quantized to its nearest neighbor using the deterministic uniform quantizer, and finally the same dither noise Z is subtracted at the decoder. It was shown [12], [13] that using this procedure $\hat{X} = X + \tilde{Z}$, where $\tilde{Z}$ has the same marginal probability distribution as Z and is also independent of X, and conditioned on the common randomness, the optimal entropy coding rate (of the lattice index) is exactly $H(f(X + Z)|Z) = I(X; X + Z)$. Note that such a rate is impossible to achieve in practice, since it requires one entropy code for a specific realization of the noise $Z = z$: Firstly, the usual technique of universal compression becomes unrealistic because it is unlikely (with zero probability) to have identical noise realizations and therefore very few samples to estimate the corresponding probability distribution; secondly, unless the distribution is analytically simple, storing the distribution or the entropy coding codewords for each noise realization is also unrealistic. Entropy coding of $f(X+Z)$ can be considered instead, resulting in a rate of $H(f(X + Z))$.

B. Rate-distortion-perception function and RDP coding

The (informational) rate-distortion-perception function can be viewed as a generalization of the rate-distortion function, which under a given distortion constraint D and a given perception constraint P, is defined as

$$R(D, P) = \min_{P_{\hat{X}|X}:\mathbb{E}d(X,\hat{X})\leq D, w(P_X,P_{\hat{X}})\leq P} I(X; \hat{X}), \quad (1)$$

where $w(\cdot, \cdot)$ is a measure quantifying the distance between two probability distributions, e.g., KL divergence, total variation, or Wasserstein metric. We are mainly interested in the case of perfect perception, i.e.,

$$R(D, 0) = \min_{P_{\hat{X}|X}:\mathbb{E}d(X,\hat{X})\leq D, P_{\hat{X}}=P_X} I(X; \hat{X}), \quad (2)$$

which is independent of the choice of $w(\cdot, \cdot)$ measure. Similar to the rate-distortion setting, it was shown [19] that the RDP function is also the fundamental limit of any encoding and decoding function pairs in the RDP setting. It was established in [20] that under the MSE distortion measure, $R(D, 0) = R(\frac{D}{2}, \infty)$. These results are again asymptotic in nature, meaning the corresponding codes are allowed to encode a large number of samples together.

For scalar quantization (also called one-shot coding), it is possible to achieve the following coding rate [19] $R(D, P) + \log(R(D, P) + 1) + 4$, using the sampling-based approach mentioned earlier, which is at a higher rate than the RDP function. The loss can be significant at the usual range of practical compression applications, e.g., at a target rate of 4bits with a potential loss of more than 4bits. It is not known whether this is the best rate possible for one-shot coding.

It has been shown that quantizers without common randomness can suffer significantly in RDP coding, and common randomness is important. Dithered quantizer appears to be a natural match and can be utilized. However, the output of the original dithered quantizer has a distribution the same as $X+\tilde{Z}$, and therefore, there is a mismatch with the target RDP-optimal distribution. Particularly, for the perfect perceptual quality setting, the distribution of $X + \tilde{Z}$ may be different from P_X, and a distribution shaping procedure is needed at the decoder, at the expense of increased distortion. This shaping can be accomplished using a nonlinear function $\phi(\cdot)$ operating on the output of the dithered quantizer $X + \tilde{Z}$, and neural networks can be used to fulfill this role.

C. Quantization on the unit circle

Consider the following idealized *unit-circle* setting: the data signal X to be compressed is uniformly distributed over the unit circle $\mathcal{X} = \{x \in \mathbb{R}^2 : \|x\|_2 = 1\}$. The distortion is measured using the square error function $d(x, \hat{x}) = \|x - \hat{x}\|_2^2$, the coding rate is set at 1 bit per sample, and the reconstruction $\hat{X}$ is required to be of perfect perception quality, i.e., $\hat{X} \overset{d}{=} X$. Since the signal has its domain is the unit circle, we can represent any $x \in \mathcal{X}$ by its angle $\theta(x) \in \Theta \triangleq (-\pi, \pi]$ such that $x = (\cos(\theta(x)), \sin(\theta(x)))$. Fixed-rate quantization at rate 1 on this data source was previously considered in [5] to illustrate the advantage of stochastic (dithered) encoders. Two types of quantizers were considered in [5]:

- Quantizer with a deterministic encoder: Since there is no common randomness, to obtain perfect perception quality, decoder side noise must be injected. It was shown that the optimal quantization procedure in this case is as follows:
$$f(\theta(x)) = \begin{cases} 1 & \theta(x) \in [0, \pi) \\ -1 & \text{otherwise} \end{cases}, \quad g(i) = \frac{i \times \pi}{2} - \bar{Z},$$
where $\bar{Z}$ is a private random variable at the decoder side, independent of X, distributed uniformly on $[-\pi/2, \pi/2)$.

We here view $g(i)$ as a random function, and therefore did not include $\bar{Z}$ as part of the function input. This procedure gives a distortion $2 - 8/\pi^2$.

- Dithered quantizer: Let Z be distributed uniformly over $[-\pi/2, \pi/2)$ independent of X, dithered quantization operates as follows:
$$f(Y) = \begin{cases} 1 & Y \in [0, \pi) \bmod 2\pi \\ -1 & \text{otherwise} \end{cases}, \; g(i) = \frac{i \times \pi}{2} - Z,$$
where $Y = \theta(x) + Z$ and $\theta(\hat{x}) = g(f(\theta(x) + Z))$. By the property of the dither quantizer, we have $\theta(\hat{X}) = \theta(X) + \tilde{Z} \mod 2\pi$, where $\tilde{Z} \overset{d}{=} Z$ and is independent of X. The distortion thus induced is $2 - 4/\pi$, which is about 38.9% lower than that using the deterministic encoder.

The dithered quantizer performs better here for two reasons: 1) The distribution of $\theta(X) + \tilde{Z} \mod 2\pi$ is uniform on the unit circle, and thus naturally matches the perceptual requirement; 2) If the perception consideration were not present, the first approach could choose a single reconstruction point to minimize the distortion, however now it is forced to utilize private randomness at the decoder, over $1/2$ of the unit circle, to produce the desired distribution; this private randomness thus induces additional distortion. Fig. 1 (a) and (b) illustrate this effect of the two procedures.

III. Quantization on the unit circle

A. Noise realization and staggered quantizers

Consider again the unit circle setting at rate 1. An alternative view of a quantizer with common randomness is to consider the quantizer induced by fixing a realization of the common randomness $Z = z$, which is illustrated in Fig. 1 (c). It is seen that the partitions of these quantizers are in fact congruent to that shown in Fig. 1 (a). Since Z is uniformly distributed on $[-\pi/2, \pi/2)$, the dithered quantization procedure is in fact mixing an uncountably many such quantizers, one for each $z \in [-\pi/2, \pi/2)$. Due to the common randomness Z, there is no need to inject decoder side randomness, which helps reduce the resultant distortion.

The two types of quantizers considered in [5] can then be viewed as two extremes of a class of quantizers: the former is a single quantizer with a deterministic encoder that relies solely on decoder side randomness for perception, while the latter is mixing (randomly selected using the common randomness) among uncountably many quantizers each with a deterministic encoder that requires no decoder side randomness. In between the two extremes, we can consider mixing staggered quantizers with deterministic encoders, which will need to rely on decoder side randomness to some extent. One such example with $N = 4$ quantizers is illustrated in Fig. 2. It can be seen that each individual quantizer only requires the decoder side randomness to be uniformly distributed on $1/8$ of the unit circle, instead of $1/2$ of the unit circle. As discussed earlier, decoder side randomness induces additional distortion, and this reduction in its range helps to reduce the distortion. As we increase the number of quantizers, the distortion is further reduced, eventually approaching that of the dithered quantizer.

B. Staggered quantizers on the unit circle

Generalizing the idea shown in Fig. 2, we can use N staggered L-level quantizers, each of which uniformly partitions the unit circle. The N quantizers are obtained by offsetting sequentially by an amount of $2\pi/(LN)$ in terms of the angle on the unit circle. The common randomness uniformly selects one of N quantizers, and the decoder adds private random noise uniformly distributed on $1/(2N)$ of the unit circle.

Theorem 1. *In the unit-circle setting, at perfect perceptual quality, N staggered quantizers each with L levels achieve the following rate-distortion pair.*
$$(R, D) = \left(\log L, 2 - 2\frac{\sin(\pi/(LN))}{\pi/(LN)}\frac{\sin(\pi/L)}{\pi/L}\right).$$

The result subsumes the special case $N = 1$ and $L = 2$ given in [5].

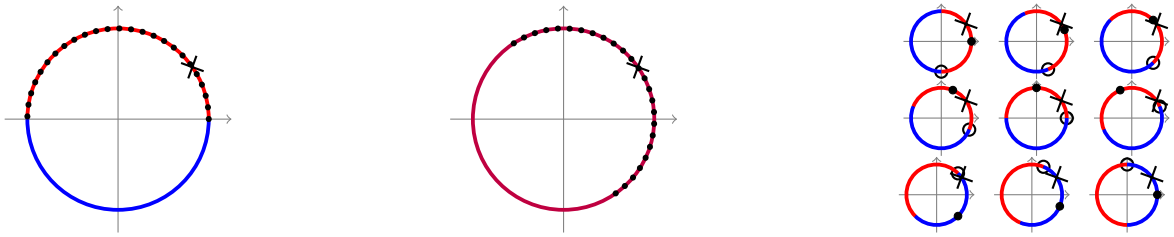
Proof of Theorem 1. Since each of N quantifiers is uniform with L levels, the rate for the corresponding quantization procedure is $\log L$. Due to symmetry, we analyze the distortion with a fixed quantizer. The arc (in angle) that the samples are quantized to the same index on has a length $(2\pi)/L$ since there are L levels, and the inserted decoder noise is placed at the center of the arc uniformly distributed with a length $(2\pi)/(NL)$ since there are also N quantizers. Since $\|(\cos(\theta), \sin(\theta)) - (\cos(\alpha), \sin(\alpha))\|^2 = 2(1 - \cos(\theta - \alpha))$, the distortion can then be calculated as
$$= \frac{L}{2\pi}\frac{LN}{2\pi}\int_{-\pi/L}^{\pi/L}\left(\int_{-\pi/(NL)}^{\pi/(NL)} 2(1 - \cos(\theta - \alpha))\mathrm{d}\alpha\right)\mathrm{d}\theta$$
$$= 2 + \frac{L^2 N}{2\pi^2}\int_{-\pi/L}^{\pi/L} \sin(\theta - \pi/(NL)) - \sin(\theta + \pi/(NL))\mathrm{d}\theta$$
$$= 2 + \frac{L^2 N}{\pi^2}\left(\cos\left(\frac{\pi}{L}\frac{N+1}{N}\right) - \cos\left(\frac{\pi}{L}\frac{N-1}{N}\right)\right)$$
$$= 2 - 2\frac{\sin(\pi/(NL))}{\pi/(NL)}\frac{\sin(\pi/L)}{\pi/L},$$
which is the desired result. □

The next two theorems provide the fundamental limits of RDP coding and single-shot coding in the unit-circle setting.

Theorem 2. *In the unit-circle setting, the information-theoretic rate-distortion trade-off with perfect perceptual quality $R(D, 0)$ is given by the pairs parametrized by $\lambda > 0$*
$$\Big\{(R, D) = \Big(\log(2\pi) - h(Z), \mathbb{E}[2 - 2\cos(Z)]\Big) : Z \sim p(z; \lambda) = \frac{e^{\lambda\cos(z)}}{\int_{-\pi}^{\pi} e^{\lambda\cos(z')}\mathrm{d}z'}, \; \lambda > 0\Big\}.$$

Note that this is the best that can be achieved using infinitely large coding blocks, and it is in general impossible to achieve using single-shot coding.

Proof of Theorem 2. We aim to minimize the rate-distortion Lagrangian with perfect perceptual quality for any Lagrange multiplier $\lambda > 0$, i.e,
$$\min_{p_{\hat{X}|X}: \hat{X} \overset{d}{=} X} I(X; \hat{X}) + \lambda\mathbb{E}[\|X - \hat{X}\|^2]. \tag{3}$$



(a) Quantizer - private randomness (b) Dithered quantization (c) Dissection of dithered quantization

Fig. 1: 1-bit quantizers on the unit-circle with perfect perceptual quality: "×" indicates a sample realization of X; "•" indicate the distribution of reconstruction $\hat{X}$; red and blue regions indicate the partition region associated with indices $+1$ and -1, respectively. In (a), the deterministic encoder is used. The sample is encoded as $+1$ and its reconstruction is distributed uniformly over the red region. In (b), the dithered approach is used, and the reconstruction would be distributed uniformly over the arc centered at the sample. There are no clear partitions in this case, and thus purple is used as a mixture of red and blue regions. In (c), "∘" indicates realizations of negative common randomness $-Z$, and the dithered quantization is viewed as a mixture of uncountably many deterministic quantizers, each associated with a realization of Z.

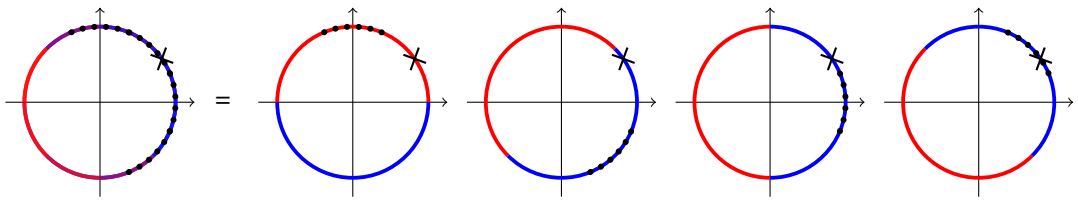


Fig. 2: Staggered quantizers with 1 bit coding rate and 2 bits common randomness.

Due to perfect perceptual quality, the reconstructed signal $\hat{X}$ must lie on the unit circle, and we can represent $\hat{X}$ by its angle $\theta(\hat{X})$. The MSE distortion term $\|X-\hat{X}\|_2^2$ can be written as $2(1-\cos\Big(\theta(X)-\theta(\hat{X})\Big))$. The mutual information can be lower bounded by

$$\begin{aligned} I(X;\hat{X}) &= h(X)-h(X|\hat{X}) \geq h(X)-h(X-\hat{X}) \\ &= h(\theta(X))-h(\theta(X)-\theta(\hat{X})). \end{aligned} \tag{4}$$

For simplicity, from here on we will write $\theta=\theta(X)$ and $\hat{\theta}=\theta(\hat{X})$, and denote $\beta:=\theta-\hat{\theta}$.

Since $h(\theta(X))$ is a constant, we can consider the optimization problem below, equivalent to lower-bounding (3)

$$\text{minimize}_{p(\beta)} -h(\beta)+2\lambda\mathbb{E}[(1-\cos(\beta))]. \tag{5}$$

Using simple calculus of variation, it can be verified that the optimal distribution of β for the optimization above is $p(\beta)=\frac{e^{2\lambda\cos(\beta)}}{\int_{-\pi}^{\pi}e^{2\lambda\cos(\beta')}\mathrm{d}\beta'}$. Since β is independent of θ, the sum $\hat{\theta}=\theta+\beta$ has a uniform distribution over $[-\pi,\pi]$. Thus this distribution indeed provides a lower bound to (3).

To show that they are in fact equal, we only need to observe that in (4), the only inequality can be written as

$$\begin{aligned} I(X;\hat{X}) &= h(X)-h(X|\hat{X}) = h(\theta)-h(\theta|\hat{\theta}) \\ &= h(\theta)-h(\beta|\hat{\theta}) \geq h(\theta)-h(\beta). \end{aligned} \tag{6}$$

However, observe that we have

$$\begin{aligned} p_{\beta|\hat{\theta}}(\beta|\hat{\theta}) &= \frac{p_{\beta,\hat{\theta}}(\beta,\hat{\theta})}{p_{\hat{\theta}}(\hat{\theta})} = \frac{p_{\beta,\theta}(\beta,\hat{\theta}-\beta)}{p_{\hat{\theta}}(\hat{\theta})} \\ &= \frac{p_{\beta}(\beta)p_{\theta}(\hat{\theta}-\beta)}{p_{\hat{\theta}}(\hat{\theta})} = p_{\beta}(\beta), \end{aligned}$$

where the last step is because both θ and $\hat{\theta}$ are uniformly distributed marginally. This implies β is in fact independent of $\hat{\theta}$, and $h(\beta|\hat{\theta})=h(\beta)$, and therefore (6) becomes equality, which establishes the overall equality. Thus the rate-distortion pairs are indeed characterized by that given in Theorem 2.

It is not difficult to verify that the curve (or function) above is continuous, and its epigraph is non-empty and closed lying in the upper right quadrant. Each point on the curve naturally has a supporting hyperplane, since it is a solution of optimizing the corresponding Lagrangian. Thus by the partial converse of supporting hyperplane theorem the curve is convex. □

Theorem 3. *In the unit-circle setting, the optimal scalar quantization (single shot coding) trade-off between the coding rate and the distortion with perfect perceptual quality is the piece-wise linear function with the following extreme points*

$$\left\{(R,D)=\left(\log L, 2-2\frac{\sin(\pi/L)}{\pi/L}\right): L=1,2,3,\dots\right\},$$

which can be achieved by dithered quantizations.

As $N\to\infty$, we see that $\frac{\sin(\pi/(LN))}{\pi/(LN)}\to 1$, therefore, the performance of the staggered quantizer approaches that of dithered quantization in this setting. Due to the uniform data source distribution, dithered quantizers are optimal, and N staggered quantizers each with L levels each does not offer any advantage over dithered quantizers. However, as we will discuss in the next section, this is not the case in general, since the flexibility in entropy coding can lead to an additional edge.

Proof of Theorem 3. Any codecs (f,g) can be represented by $f:\mathcal{X}\times\mathbb{R}\to\mathbb{Z}$ and $g:\mathbb{Z}\times\mathbb{R}\to\mathcal{X}$. The signal X is encoded by $f(X,V)$ to some integer and then reconstructed by $\hat{X}=g(f(X,V),V)$, where V is the common randomness.

Due to the perfect perceptual quality requirement, the reconstructed signal $\hat{X}$ must lie on the unit circle. Without considering perceptual quality, we first characterize the scalar optimal quantization under the condition that reconstruction $\hat{X}$ lies on the unit circle. Take any Lagrange multiplier $\lambda>0$,

consider minimizing the following rate distortion Lagrangian with decision variables (f, g, V)

$$H(f(X;V)|V) + \lambda \mathbb{E}_{X,V}[d(X, g(f(X;V);V))]$$
$$= \mathbb{E}_V[\mathbb{E}_X[-\log(\mathbb{P}(f(X;V)|V)) + \lambda d(X, g(f(X;V);V))|V]]$$

It suffices to study the deterministic quantizer, since for any stochastic quantizer (f, g, V), there exists a deterministic quantizer $(f(;v), g(;,v))$ with some realization of $V = v$ such that its Lagrangian is at most that of the stochastic quantizer.

It is straightforward to verify that the optimal deterministic quantizer in this setting must have contiguous regions (pathological cases may exist for complex distributions [21]), i.e., the region in $\mathcal{X}$ of the same index $f(\cdot, v)$ should be contiguous. For such a quantizer with L levels, i.e., $|f(\cdot, v)| = L$, it can then be shown using calculus that it must be a uniform quantizer. The optimal scalar quantization (single shot coding) trade-off between the coding rate and the distortion is the piece-wise linear function with the following extreme points

$$\left\{(R, D) = \left(\log L, 2 - 2\frac{\sin(\pi/L)}{\pi/L}\right) : L = 1, 2, 3, \ldots\right\}.$$

This piece-wise linear function is a lower bound, when considering perfect perceptual quality. However, it is straightforward to verify that dithered quantization has the perfect perceptual quality and can achieve the extreme points and thus match the lower bound. Thus the optimal scalar quantization trade-off between the coding rate and the distortion with perfect perceptual quality is also the piece-wise linear function above and can be achieved by time-sharing dithered quantizers. □

IV. Design of Staggered Quantizers for General Scalar Sources

Consider applying the staggered quantization approach to a general scalar source. Assuming there are N uniform quantizers to be staggered, the encoding function $f_n(x)$ for the n-th quantizer with stepsize Δ is

$$f_n(x) = \left\lfloor \frac{x}{\Delta} - \frac{n}{N} \right\rceil, \quad n = 0, 1, 2, \ldots, N-1, \tag{7}$$

where $\lfloor \cdot \rceil$ is the operation that rounds to the nearest integer.

To achieve perfect perceptual quality, decoder side randomness must be used, yet due to the potential non-uniformity of the distribution, it is more involved than simply subtracting certain random values. To present the procedure, first denote the density of the data source X as $p_X(x)$ and denote by $F_X(x) = \mathbb{P}(X \leq x)$ its cumulative distribution function. Denote its inverse as $F_X^{-1}(t) \triangleq \inf\{x : F_X(x) > t\}$ for any $t \in [0, 1)$. Let us introduce a density function on $[a, b]$ as $q_{a,b}(x) \triangleq \frac{p_X(x)}{\int_a^b p_X(t)dt}$. A random variable generated privately at the decoder side according to this distribution is denoted as $\tilde{Z}_{a,b}$, which is independent of all the other random variables.

Define an indexing function $m(x, n) = N \cdot f_n(x) + n$, which essentially specifies an order of all the quantization cells in all these N quantizers. Define its inverse at input x as $m_X^{-1}(j) \triangleq \inf\{x : \exists n \in [0 : N-1], m(x, n) = j\}$. Intuitively, for each quantizer and quantizer cell index pair $(n, f_n(x))$, the reconstruction at the decoder is a random variable that follows a distribution that matches the data sample distribution

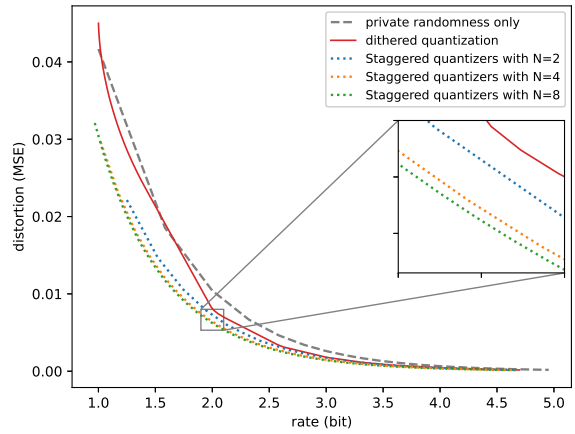


Fig. 3: Quantization of a uniformly distributed source on an interval

in an interval. Now to specify the specific interval, we define a sequence of boundaries $(a(j), b(j))_{j \in \mathbb{Z}}$ as

$$a(j) \triangleq F_X^{-1}\left(\sum_{k=1}^{N} \frac{F_X(m_x^{-1}(j-k))}{N}\right), \; b(j-1) \triangleq a(j).$$

The encoding and reconstruction process can now be described as follows. Given data source X at the encoder side, the encoding procedure uniformly at random selects one of the N encoders $\{f_0, f_1, \ldots, f_{N-1}\}$ with stepsize Δ. The index n of the selected encoder is a common randomness shared by the decoder, and the data sample is encoded as $f_n(X)$. At the decoder, we compute the index j using $f_n(x)$ and n by the indexing function $m(\cdot)$, and the reconstruction is a random sample $\hat{X} = \tilde{Z}_{a(j),b(j)}$. More formally, the decoding function upon receiving code $f_n(X) = i$ is

$$g(i) = \tilde{Z}_{a(j),b(j)}, \quad \text{with } j = Ni + n, \tag{8}$$

where n is the common randomness of the offset quantizer index. We remark here that the offsets can be viewed as a random dither which takes discrete values in $\{0, 1/N, 2/N, \ldots, (N-1)/N\}$. However, for each realization, the reconstruction is sampled in an interval, unlike in classic deterministic quantizers or dithered quantizers.

Since the number of staggered quantizers is small, it is possible to design tailored entropy code for each, whereas this is impossible for dithered quantizers, resulting in a rate close to $H(f(X + Z))$. Dithered quantization also suffers because $X + \tilde{Z}$ induces loss of perception, and an additional shaping step is required. As shown in Fig. 3, the proposed approach can sometimes outperform both dithered quantizers and deterministic encoders. Particularly, even mixing 2 quantizers appears to provide competitive performance.

V. Conclusion

We consider RDP coding from a quantizer design perspective. By decomposing dithered quantization, we obtain staggered quantizers as intermediates between the two extremes of dithered quantization and quantization without common randomness. This new perspective provides a new way to understand one-shot coding for RDP.

REFERENCES

[1] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Springer, 1992, vol. 159.

[2] T. Berger, *Rate distortion theory: A mathematical basis for data compression*. Prentice-Hall series in information and system sciences, 1971.

[3] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *International Conference on Machine Learning*. PMLR, 2019, pp. 675–685.

[4] A. B. Wagner, "The rate-distortion-perception tradeoff: The role of common randomness," *arXiv preprint arXiv:2202.04147*, 2022.

[5] L. Theis and E. Agustsson, "On the advantages of stochastic encoders," in *Neural Compression: From Information Theory to Applications–Workshop@ ICLR 2021*, 2021.

[6] J. Chen, L. Yu, J. Wang, W. Shi, Y. Ge, and W. Tong, "On the rate-distortion-perception function," *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 4, pp. 664–673, 2022.

[7] G. Serra, P. A. Stavrou, and M. Kountouris, "Computation of rate-distortion-perception function under f-divergence perception constraints," *arXiv preprint arXiv:2305.04604*, 2023.

[8] Y. Hamdi and D. Gündüz, "The rate-distortion-perception trade-off with side information," *arXiv preprint arXiv:2305.13116*, 2023.

[9] S. Salehkalaibar, J. Chen, A. Khisti, and W. Yu, "Rate-distortion-perception tradeoff based on the conditional-distribution perception measure," *arXiv preprint arXiv:2401.12207*, 2024.

[10] X. Niu, D. Gündüz, B. Bai, and W. Han, "Conditional rate-distortion-perception trade-off," *arXiv preprint arXiv:2305.09318*, 2023.

[11] C. T. Li and A. El Gamal, "Strong functional representation lemma and applications to coding theorems," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 6967–6978, 2018.

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

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

[14] N. Saldi, T. Linder, and S. Yüksel, "Randomized quantization and source coding with constrained output distribution," *IEEE Transactions on Information Theory*, vol. 61, no. 1, pp. 91–106, 2014.

[15] C. Tian, "A new class of multiple description scalar quantizer and its application to image coding," *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 329–332, 2005.

[16] ——, "Staggered lattices in multiple description quantization," in *Data Compression Conference*, 2005, pp. 398–407.

[17] U. Samarawickrama, J. Liang, and C. Tian, "m-channel multiple description coding with two-rate coding and staggered quantization," *IEEE transactions on circuits and systems for video technology*, vol. 20, no. 7, pp. 933–944, 2010.

[18] R. Zamir, *Lattice Coding for Signals and Networks: A Structured Coding Approach to Quantization, Modulation, and Multiuser Information Theory*. Cambridge University Press, 2014.

[19] L. Theis and A. B. Wagner, "A coding theorem for the rate-distortion-perception function," in *Neural Compression: From Information Theory to Applications–Workshop@ ICLR 2021*, 2021.

[20] Z. Yan, F. Wen, R. Ying, C. Ma, and P. Liu, "On perceptual lossy compression: The cost of perceptual reconstruction and an optimal training framework," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 682–11 692.

[21] A. Gyorgy and T. Linder, "On the structure of optimal entropy-constrained scalar quantizers," *IEEE transactions on information theory*, vol. 48, no. 2, pp. 416–427, 2002.

APPENDIX

Optimality of Uniform Quantizers in the Proof of Theorem 3. Consider two adjacent Voronoi cells. Suppose the two adjacent regions have a total size (in terms of the angle spanned) $2\pi r$ for some $r \in (0, 1]$, moreover, suppose the first Voronoi is of size $2\pi\alpha$ for some $\alpha \in (0, r)$. For optimal partitions, α must be a minimizer of the following function

$$\begin{aligned} l(\alpha; r) = &(r-\alpha)\ln(r-\alpha) + \frac{\lambda}{\pi}\sin(\pi(r-\alpha)) \\ &+ \alpha\ln(\alpha) + \frac{\lambda}{\pi}\sin(\pi\alpha). \end{aligned}$$

Its derivative is

$$\begin{aligned} l'(\alpha; r) = &-\ln(r-\alpha) - \lambda\cos(\pi(r-\alpha)) \\ &+ \ln(\alpha) + \lambda\cos(\pi\alpha) \end{aligned}$$

and its second derivative is

$$l''(\alpha; r) = \frac{1}{r-\alpha} + \frac{1}{\alpha} - \lambda\pi(\sin(\pi(r-\alpha)) + \sin(\pi\alpha)).$$

It is not hard to verify that l and l'' are even functions, and l' is an odd function. There are two circumstances

1) λ is small, and $l''(\alpha; r) \geq 0$. Then $l(\alpha; r)$ is a non-constant symmetric convex function whose optimal value is achieved by $\alpha \to 0$ or $\alpha \to r$, which conflicts with the fact that the optimal quantizer has non-empty Voronoi.
2) λ is large, and $l''(\alpha; r)$ will be positive on both ends and negative in the middle. $l'(\alpha; r)$ is increasing, decreasing and increasing. $l(\alpha; r)$ will either have a maximum with $\alpha = r/2$ or the maximum is approached by $\alpha \to 0$ or $\alpha \to r$.

Therefore any two adjacent non-empty Voronoi cells have the same size. The optimal quantizer thus must have equal-sized Voronoi cells, thus a uniform quantizer. □