MULTIQUAN RDP: RATE-DISTORTION-PERCEPTION CODING VIA OFFSET QUANTIZERS

Anonymous authors
Paper under double-blind review

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 on 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 multiple quantizers with offsets. Though the procedure can be viewed as intermediates between the two extremes, its explicit structure can be advantageous in some cases. Experimental results are given on both simple data sources and images to illustrate its behavior.

1 INTRODUCTION

Compression plays an important role in efficient representation of information content, particularly visual content. Traditionally, the tradeoff between the compression rate and the distortion caused by such compression has been studied under two different but related frameworks: the quantization framework (Gersho & Gray, 1992) and the rate-distortion theory (Berger, 1971) framework. In the former, the focus is on the design of quantizers that compress the data samples one a time (i.e., scalar quantization) or 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. While the latter approach is able to provide sharp theoretical guarantees in the information theoretic sense, the former has arguably more practical impact leading to near-optimal quantizers. In both approaches, the distortion is measured in an objective yet potentially artificial manner, e.g., in terms of the mean squared error (MSE).

Largely driven by the recent emergence of the neural compression, the issue of perceptual quality has led to the formulation and the study of the problem of rate-distortion-perception (RDP) tradeoff. In this formulation, a new quality constraint, which is introduced to capture the perceptual quality loss due to compression, is further imposed in addition to the existing objective distortion constraint. Mathematically, this formulation (Blau & Michaeli, 2019) 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".

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 (Theis & Agustsson, 2021; Chen et al., 2022). More precisely, the lack of the common randomness can cause significant performance loss comparing 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 (Li & El Gamal, 2018), and the second is through universal dithered quantization (Ziv, 1985; Zamir & Feder, 1992). The probabilistic sampling-based method requires the knowledge of a target joint distribution between the samples and the compressed version, and furthermore, involves a rather complex sampling procedure. This is hardly surprising, since the approach was originally proposed to provide strong information theoretic bounds instead of as an immediately practical coding 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 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 dither-based approach will utilizes a 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 (Saldi et al., 2014), the asymptotic nature of such analysis makes the mechanism rather opaque. In this work, we set out to develop understanding on these issues under the quantization framework, and the main contribution of this work is as follows:

- Using a decomposition perspective, we provide a new way to understand the mechanism from which procedures utilizing common randomness obtain the advantage. A simple setting is analyzed in detail to illustrate the benefit in a quantitative manner.
- Based on these understandings, we provide a new approach to introduce common randomness using multiple quantizers with offsets, which can be viewed as intermediates between the two extreme cases of with full common randomness and without common randomness.
- The new approach is applied on simple data sources, and on neural network based image compression to illustrate its behavior. The explicit coding structure can be advantageous in some settings, e.g., for nonuniform distributions and/or distributions with bounded support.

2 BACKGROUNDS

2.1 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

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

where $I(\cdot; \cdot)$ is the mutual information function. Explicit solutions for this optimization problem are only known for some limited probability distributions and distortion measures, however for discrete random variables on finite alphabets, the Blahut-Arimoto algorithm can be used to compute the solution efficiently (Cover & Thomas, 2006). The significance of this function is that it characterizes the best possible rate that any encoding function and decoding function pair can accomplish.

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 quantization (ECQ), whereas when the number of quantization level is fixed, it is usually referred to fixed-rate quantization. The encoding and decoding functions can be optimized using iterative algorithms, such as Lloyd algorithm (Lloyd, 1982) or generalized Lloyd algorithms (Chou et al., 1989).

Universal dithered quantizer utilizes a uniform quantizer with stepsize Δ in the encoding and decoding process (Zamir, 2014). 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 on top of the sample resulting $X + Z$, which is then quantized to its nearest neighbour using the deterministic uniform

quantizer, and the finally the same dither noise Z is subtracted from this point at the decoder. It was shown (Ziv, 1985; Zamir & Feder, 1992) 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 in practice, such rate is not possible to achieve, since it requires one entropy code for a specific realization of the noise $Z = z$. Entropy coding with respect to the (marginal) distribution of $f(X+Z)$ is usually applied for a rate $H(f(X+Z))$, which is higher than $H(f(X+Z)|Z)$.

2.2 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

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

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.,

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

which is independent of the choice of $w(\cdot,\cdot)$ measure. Similar to the rate-distortion setting, it was shown (Theis & Wagner, 2021) that the RDP function is also the fundamental limits of any encoding and decoding function pairs in the RDP setting. It was established in Yan et al. (2021) 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.

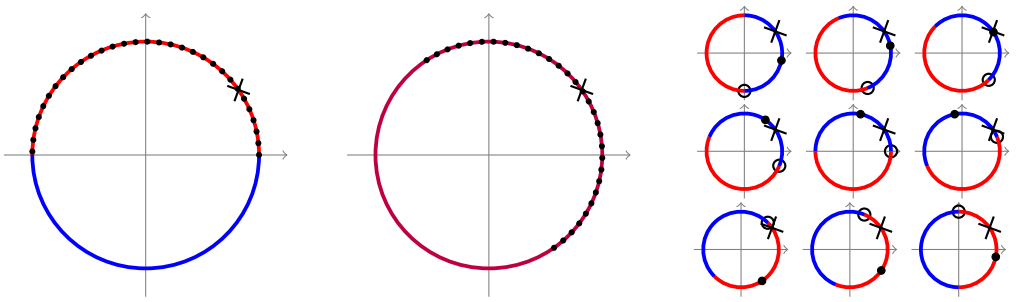
For one-shot coding (i.e., scalar coding), it is possible to achieve the following coding rate (Theis & Wagner, 2021) $R(D,P) + \log(R(D,P)+1) + 4$, using the sampling-based approach mentioned earlier, which is at a higher rate than the RDP function gives. The loss can be significant at the usual range of practical compression applications, e.g., at a target rate 4bits with a potential loss of more than 4bits. It is however 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 to bring the performance close to the RDP function. Dithered quantizer, with its common random dither, appears to be a natural match and can be utilized in this setting. However, note that the output of the original dithered quantizer has a distribution the same as $X + Z$, and therefore, there is a mismatch with the target RDP-optimal distribution. Particularly, for the perfect perceptual quality setting, the distribution of $X + Z$ 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 are often used to fulfill this role.

3 Analysis of quantization on the unit circle

As a starting step, let us 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 being 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 the data source was previously considered in Theis & Agustsson (2021) to illustrate the advantage of stochastic (dithered) encoders. Two types of quantizers were considered there:



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

Figure 1: 1-bit quantizers on the unit-circle with perfect perceptual quality: the "$\times$" indicates a sample realization of X; the dots "$\bullet$" 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 when the sample lies in the red region. It 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), "$\circ$" indicates realizations of common randomness Z, and the dithered quantization procedure is viewed as a mixture of uncountably many deterministic quantizers, each associated with a realization of Z.

- Quantizer with a deterministic encoder (no common randomness): 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:

$$f(\theta(x)) = \begin{cases} 1 & \theta(x) \in [0, \pi) \\ -1 & \text{otherwise} \end{cases}, \qquad g(i) = \frac{i \times \pi}{2} - \tilde{Z}, \tag{3}$$

  where $\tilde{Z}$ is a private random variable at the decoder side, independent of X, distributed uniformly on $[-\pi/2, \pi/2)$. Note here we view $g(i)$ as a random function, and therefore did not include $\tilde{Z}$ as part of the function input. This procedure gives a distortion $2 - 8/\pi^2$.

- Dithered quantizer (with common randomness): Let Z be distributed uniformly over $[-\pi/2, \pi/2)$ independent of X, dithered quantization operates as follows:

$$f(\theta(x) + Z) = \begin{cases} 1 & \theta(x) + Z \in [0, \pi) \mod 2\pi \\ -1 & \text{otherwise} \end{cases}, \qquad g(i) = \frac{i \times \pi}{2} - Z, \tag{4}$$

  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 exactly uniform on the unit circle, and thus naturally match the perception requirement; 2) if the perception consideration is not present, the quantizer without common randomness could choose a single reconstruction point to minimize the distortion, however it is forced to utilize private randomness at the decoder, over $1/2$ of the unit circle, to produce the desired distribution. In Fig. 1 (a) and (b), we illustrate this effect for the two procedures.

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 Theis & Agustsson (2021) 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 the perception consideration, while the latter is

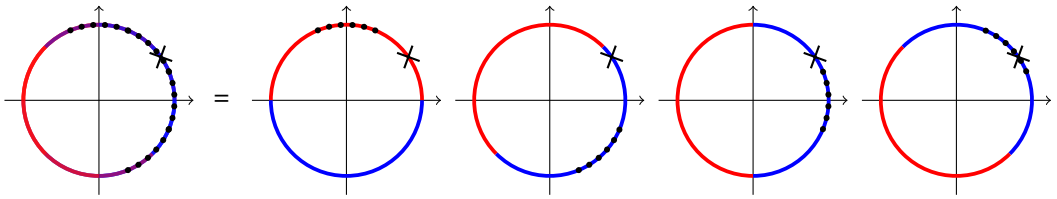


Figure 2: MultiQuan of 4 quantizers with 2 levels: 1 bit coding rate and 2 bits common randomness.

mixing (randomly selected using the common randomness) among an uncountably many quantizers each with a deterministic encoder that requires no decoder side randomness. In between the two extremes, we can consider mixing several 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 distortion, and this reduction on 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.

More generally, we can use L-quantization levels which uniformly partitions the unit circle. The N quantizers are obtained by offseting sequentially by an amount of $2\pi/(LN)$ in terms of the angle on the unit circle, one after another. We refer to this as the MultiQuan procedure.

Theorem 3.1. *In the unit-circle setting, at perfect perceptual quality MultiQuan of N quantizers with L levels achieves the rate-distortion pair*

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

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

Theorem 3.2. *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 $\lambda > 0$*

$$\left\{ (R, D) = \Big(\log(2\pi) - h(Z), \mathbb{E}[2 - 2\cos(Z)] \Big) : Z \sim p(z; \lambda) = \frac{e^{\lambda \cos(z)}}{\int_{-\pi}^{\pi} e^{\lambda \cos(z')} \mathrm{d}z'}, \ \lambda > 0 \right\}.$$

Note that this is a slice of the RDP function, which is the best possible allowing infinite large coding blocks of data samples, and it is in general not possible to achieve using scalar (single shot) coding.

Theorem 3.3. *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 introduced earlier.

As $N \to \infty$, $\frac{\sin(\pi/(LN))}{\pi/(LN)} \to 1$, therefore, the performance of the MultiQuan approaches that of dithered quantization in this setting. Due to the uniform data source distribution, dithered quantizers are optimal, and MultiQuan with N quantizers and L levels each does not offer any advantage over dithered quantizers. However, as we will show and discuss in more details in the next section, this is not the case in general, since the flexibility in entropy coding and optimizing the thresholds can lead to an additional edge. The proofs of Theorem 3.1-3.3 are given in the supplementary material.

Generalizing the deterministic encoder with only private randomness procedure, it can be shown (Theis & Agustsson, 2021) that without common randomness, the operating points $(R, D) = \left(\log L, 2 - 2\frac{\sin^2(\pi/L)}{(\pi/L)^2} \right)$ can be achieved, which can also be viewed as MultiQuan with $N = 1$.

In Fig. 3, we plot the rate and distortion pairs for different methods. The RDF function is that given in Theorem 3.2, which can only be approached by coding asymptotically long blocks of samples. As we increase the number of quantizers being used in MultiQuan, its performance improves and eventually approaches that of the optimal single-shot quantization. The optimal single-shot quantization is achieved by the dithered quantization procedure in this unit-circle setting. The procedure based on deterministic encoder with private randomness at the decoder performs the worst as we anticipated. More discussions on the unit-circle setting is postponed to the appendix.

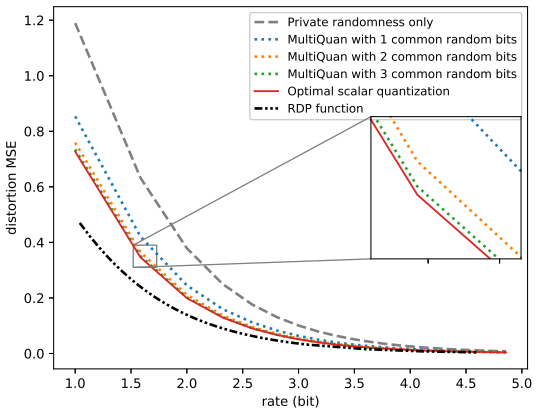


Figure 3: Performance comparison of different procedures in the unit-circle setting.

4 Design and Optimization of Offset Quantizers

The unit circle setting discussed in the previous section has a uniform source data distribution for which the dithered quantizer has a natural advantage since $\theta(X) + \tilde{Z}$ in fact matches the desired output distribution. This implies that there does not exist tension between distortion and perception quality in that particular setting. In this section, we consider more general distributions, where the tension between the distortion and the perception constraint does manifest.

In addition to providing an interpretation for the advantage of dithered quantization over deterministic encoder, potential coding advantages of MultiQuan are as follows: since the number of quantizers is small, it is possible to design and use tailored entropy code specifically for each, whereas for dithered quantizer, this becomes impractical (the noise realization z is in an uncoutable set); moreover, the explicit form of MultiQuan structure allows us to also optimize the quantization thresholds explicitly, in contrast to dithered approach for which the thresholds are always uniform.

4.1 Uniform quantizers with uniform offsets

We first introduce MultiQuan with uniform quantization stepsizes, before discussing optimization of the quantization thresholds in the next subsection. Assuming there are N uniform quantizers to be used jointly in the proposed method, the encoding function $f_n(x)$ for the n-th quantizer with stepsize Δ is

$$f_n(x) = \left\lfloor \frac{x}{\Delta} - \frac{n}{N} \right\rceil, \quad n = 0, 1, 2, \ldots, N-1 \tag{5}$$

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 non-uniformity of the distribution, it is more involved than simply subtracting certain random value. 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 generating 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 $q_{a,b}(x) \triangleq \frac{p_X(x)}{\int_{a,b} p_X(t)dt}$. A random variable that is 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 probability distribution which matches the data sample distribution in an interval. Now to specify the specific interval, we define a

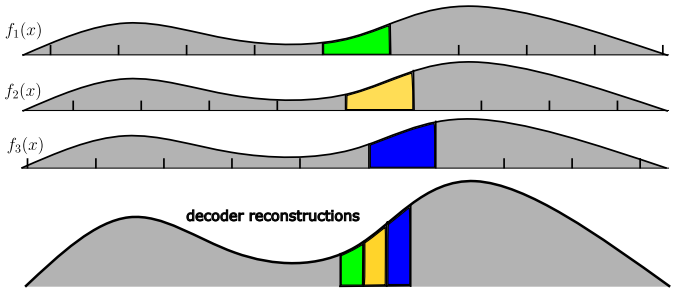


Figure 4: MultiQuan based on offset quantizers: each component quantization cell will be mapped to a reconstruction interval at the decoder, and the probability mass in the component quantizer cell is N times the mass in that reconstruction cell for the same probability distribution due to the perception constraint.

sequence of boundaries $(a(j), b(j))_{j\in\mathbb{Z}}$ as

$$a(j) \triangleq F_X^{-1}\left(\sum_{k=1}^{N} \frac{F_X(m_x^{-1}(j-k))}{N}\right), \qquad b(j-1) \triangleq a(j). \tag{6}$$

The encoding and reconstruction process can now be described as follows. Given data source X at the encoder side, the MultiQuan 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, MultiQuan decoder computes the index j using $f(x)$ and n by 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

$$g(i) = \tilde{Z}_{a(j),b(j)}, \quad \text{with } j = Ni + n, \tag{7}$$

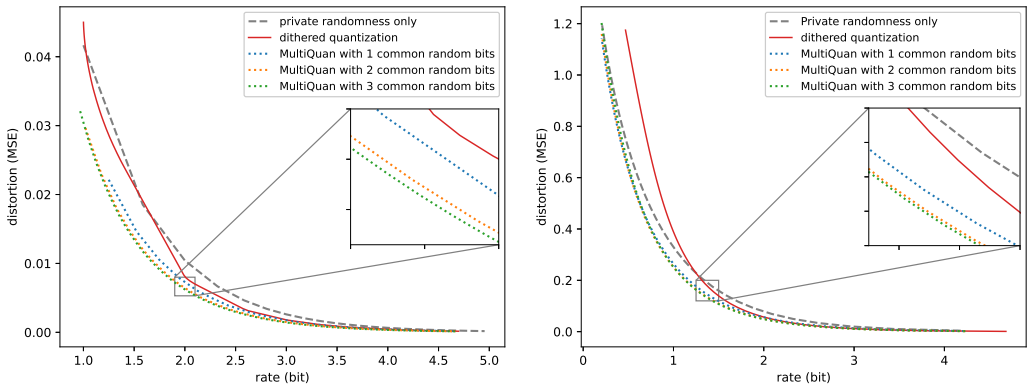
where n is the common randomness of the offset quantizer index.

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 an interval instead of a single value, unlike in classic deterministic quantizers or dithered quantizers.

4.2 Optimizing offset quantizers

Unlike dithered quantizers, where the encoding function is restricted to have uniform thresholds, MultiQuan does not have such constraint and we can further optimize the thresholds. However, such optimization is not expected to provide significant gain over uniform thresholds when entropy coding is allowed in the procedure. This type of effect was well-known at high rate in classic quantization theory (Gish & Pierce, 1968). The optimization of quantization thresholds turns out to be rather difficult in the RDP setting, and we consider two methods:

- Gradient descent-based method: At perfect perceptual quality, we can define the cost function as $D+\lambda R$, where λ is the Lagrange multiplier. Viewing the thresholds of the quantizers as the variables, we can directly apply gradient descent type of methods to minimize this cost function. Note that the algorithm may only converge to a local optimal, even if the stepsize is properly chosen.
- Pseudo-Llyod method: Generalized Lloyd method is a classic iterative method to train entropy-constraint quantizers that updates the thresholds, the reconstruction points, and the codeword lengths separately, assuming the other two are fixed. In the setting of MultiQuan, we can view the support of each $\tilde{Z}_{a,b}$ as a reconstruction interval, instead of a single reconstruction point, and apply the same iterative procedure. However, in the RDP setting the analogous condition for updating the thresholds does not guarantee optimal thresholds even with the other two components fixed, and therefore this step may not always reduce the cost, further implying that the algorithm may not always converge.



(a) Uniform distributed source

(b) Exponential distributed source

Figure 5: Comparing quantization procedures for simple sources with perfect perceptual quality

Details of these two approaches are given in the appendix together with some numerical results.

5 Experimental Results

We experimentally compare the performance of MultiQuan with that of quantization without common randomness and dithered quantization. For simple data sources, perfect perceptual quality is enforced, whereas for image data sources, a high perceptual quality constraint is enforced.

5.1 Simple source with perfect perceptual quality

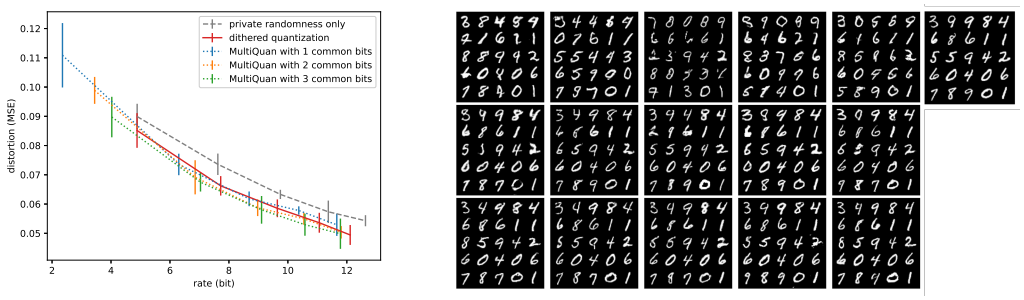
We first study scalar quantization for two simple sources – uniform distribution (source X distributes uniformly over the unit interval $[0, 1]$) and exponential distribution (source X has density $p_X(x) = e^{-x}$ for $x \geq 0$), where perfect perceptual quality can be analytically guaranteed. MultiQuan quantization procedures with 1-3 common random bits are compared to two reference quantization procedures: 1) quantization without common randomness (same as MultiQuan with $N = 1$); 2) dithered quantization with post processing $\phi(\cdot)$, which transforms the output of dithered quantizer $X + \tilde{Z}$ such that $\phi(X + \tilde{Z})$ has the same distribution as the source data X. This can be accomplished using the function $\phi(y) = F_X^{-1}(F_{X+\tilde{Z}}(y))$ (Li et al., 2010).

Although we mainly aim to develop better understanding on the connection between quantizers without randomness and dithered quantizers, as shown in Figure 7, the MultiQuan procedures can sometimes outperform both of them. Particularly, even just mixing 2 quantizers appears to provide very competitive performance, relative to dithered quantization. The performance of dithered quantization is superior to that of quantization without common randomness when rate is moderately high, but at low rate dithered quantization suffers, because entropy coding cannot be realistically and effectively performed conditioned on each Z realization, but rather is done on $f(X + Z)$.

In the appendix, we further improve the MultiQuan procedure by the gradient-descent and pseudo-Lloyd algorithms and compare it with dithered quantization with conditional entropy coding. Even though the latter improves significantly over the dithered quantization with marginal entropy coding, MultiQuan procedures (with optimization) is still able to outperform other procedures.

5.2 Neural network-based image compression

We next consider neural network-based image compression on the MNIST dataset, similar to that used in Blau & Michaeli (2019). We follows an auto encoder-decoder architecture under a Wasserstein GAN regularization, with three trainable neural network components $(f(\cdot; \omega_f), g(\cdot; \omega_g), c(\cdot; \omega_c))$ parameterized by $(\omega_f, \omega_g, \omega_c)$. A scalar quantization procedure Q is then used between the encoder and decoder, which is chosen to be either quantization without common randomness, dithered quantization, or MutliQuan quantization.



(a) (b)

Figure 6: Neural network based image compression. (a) Rate-distortion tradeoff of various procedures at high perception quality. (b) Reconstructed images of quantization without common randomness, dithered quantization, MultiQuan with $N = 2, 4, 6$ in the columns, respective; rate increases from top to bottom in each column.

More precisely, the neural network $f(\cdot;\omega_f)$ is a non-linear function that maps the input image X to a vector v in the low dimensional hidden space $[-1,1]^d$. The quantization procedure Q is applied to signal v coordinate-wise, which encodes v into bits at the encoder side and reconstructs it to $\hat{v}$ at the decoder side. Generator $g(\cdot;\omega_g)$ generates/reconstructs image $\hat{X}$ based on the reconstructed signal $\hat{v}$. The critic $c(\cdot;\omega_c)$ discriminates the images from the true image distribution and that of the generated image distribution to preserve the perceptual quality.

The training steps follow those in Blau & Michaeli (2019). Since perfect perceptual quality can not be analytically guaranteed, we take a sufficiently large Lagrange multiplier $\lambda = 0.08$ and minimize the distortion-perception Lagrangian

$$\mathcal{L} = \mathbb{E}[\|X - \hat{X}\|_2^2] + \lambda W_1(p_X, p_{\hat{X}}),$$

where W_1 is the 1-Wasserstein metric with Kanotorovich's dual formulation $W_1(p_X, p_{\hat{X}}) = \sup_{\text{1-Lipschitz } c:\mathcal{X}\to\mathbb{R}} \left(\mathbb{E}[c(X)] - \mathbb{E}[c(\hat{X})]\right)$, and is approximated by maximizing parameterized critic function $c(\cdot;\omega_c)$ with gradient regularization Gulrajani et al. (2017) during the training. The soft gradient estimator of Mentzer et al. (2018) is used to back-propagate through the quantizer. More details on the training procedure and related discussions are given in the appendix, and our code will be released upon paper's acceptance.

From Fig. 6, it is seen that similar to the simple data sources, there is a performance gap between dithered quantization and quantizer without common randomness. MultiQuan procedures can again outperform dithered quantization in some cases when neural network is used. On the plot we also indicate the error bar on the distortions in different training runs. It is seen that neural network based image compression induces significant variations. We train the neural network such that the perceptual quality is sufficiently high, however it should be noted that perfect perception quality cannot be enforced, and the discriminator can only estimate the true Wasserstein distance.

6 Conclusion

We consider RDP coding from a quantizer design perspective. By decomposing dithered quantization, we obtain MultiQuan as intermediates between the two extremes of dithered quantization and quantization without common randomness. This new perspective provides a new way to understand the advantage of coding procedures with common randomness. Interestingly, in some cases, the MultiQuan procedure can in fact outperform dithered quantization by mixing only a few quantizers, since its explicit structure allows effective entropy code of each individual quantizer, and further allows optimization of the quantization thresholds. We focus on the case of perfect perceptual quality in this work, and leave the study of more general perceptual quality to a future work.

REFERENCES

Eirikur Agustsson and Lucas Theis. Universally quantized neural compression. *Advances in neural information processing systems*, 33:12367–12376, 2020.

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

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pp. 675–685. PMLR, 2019.

Jun Chen, Lei Yu, Jia Wang, Wuxian Shi, Yiqun Ge, and Wen Tong. On the rate-distortion-perception function. *arXiv preprint arXiv:2204.06049*, 2022.

Philip A Chou, Tom Lookabaugh, and Robert M Gray. Entropy-constrained vector quantization. *IEEE Transactions on acoustics, speech, and signal processing*, 37(1):31–42, 1989.

Thomas Cover and A. Joy Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

Paul Cuff. Distributed channel synthesis. *IEEE Transactions on Information Theory*, 59(11):7071–7096, 2013.

Allen Gersho and Robert M Gray. *Vector quantization and signal compression*, volume 159. Springer, 1992.

Herbert Gish and John Pierce. Asymptotically efficient quantizing. *IEEE Transactions on Information Theory*, 14(5):676–683, 1968.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Cheuk Ting Li and Abbas El Gamal. Strong functional representation lemma and applications to coding theorems. *IEEE Transactions on Information Theory*, 64(11):6967–6978, 2018.

Minyue Li, Janusz Klejsa, and W Bastiaan Kleijn. Distribution preserving quantization with dithering and transformation. *IEEE Signal Processing Letters*, 17(12):1014–1017, 2010.

Huan Liu, George Zhang, Jun Chen, and Ashish J Khisti. Lossy compression with distribution shift as entropy constrained optimal transport. In *International Conference on Learning Representations*, 2022.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4394–4402, 2018.

Lawrence Roberts. Picture coding using pseudo-random noise. *IRE Transactions on Information Theory*, 8(2):145–154, 1962.

Naci Saldi, Tamás Linder, and Serdar Yüksel. Randomized quantization and optimal design with a marginal constraint. In *2013 IEEE International Symposium on Information Theory*, pp. 2349–2353. IEEE, 2013.

Naci Saldi, Tamás Linder, and Serdar Yüksel. Randomized quantization and source coding with constrained output distribution. *IEEE Transactions on Information Theory*, 61(1):91–106, 2014.

Lucas Theis and Eirikur Agustsson. On the advantages of stochastic encoders. In *Neural Compression: From Information Theory to Applications–Workshop@ ICLR 2021*, 2021.

Lucas Theis and Noureldin Y Ahmed. Algorithms for the communication of samples. In *International Conference on Machine Learning*, pp. 21308–21328. PMLR, 2022.

Lucas Theis and Aaron B Wagner. A coding theorem for the rate-distortion-perception function. In *Neural Compression: From Information Theory to Applications–Workshop@ ICLR 2021*, 2021.

Aaron B Wagner. The rate-distortion-perception tradeoff: The role of common randomness. *arXiv preprint arXiv:2202.04147*, 2022.

Zeyu Yan, Fei Wen, Rendong Ying, Chao Ma, and Peilin Liu. On perceptual lossy compression: The cost of perceptual reconstruction and an optimal training framework. In *International Conference on Machine Learning*, pp. 11682–11692. PMLR, 2021.

Yibo Yang, Stephan Mandt, and Lucas Theis. An introduction to neural data compression. *arXiv preprint arXiv:2202.06533*, 2022.

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

Ram Zamir and Meir Feder. On universal quantization by randomized uniform/lattice quantizers. *IEEE Transactions on Information Theory*, 38(2):428–436, 1992.

George Zhang, Jingjing Qian, Jun Chen, and Ashish Khisti. Universal rate-distortion-perception representations for lossy compression. *Advances in Neural Information Processing Systems*, 34: 11517–11529, 2021.

Jacob Ziv. On universal quantization. *IEEE Transactions on Information Theory*, 31(3):344–347, 1985.

A Related works

Rate-distortion (RD) theory provides a mathematical framework to study the fundamental limits of the rates needed for preserving the signal under a given distortion criteria (Cover & Thomas, 2006; Berger, 1971). Such studies are asymptotic (in term of the number of samples encoded together as a single block) in nature. Its more practical counterpart, quantization theory, provides analysis and design for efficient scalar or vector quantizers (Gersho & Gray, 1992). The distortion measure imposed in traditional studies does not adaquately consider the perceptual quality, and as a remedy, Blau and Michaeli (Blau & Michaeli, 2019) proposed the RDP framework, which has since attracted significant attention (Theis & Agustsson, 2021; Theis & Wagner, 2021; Yan et al., 2021; Zhang et al., 2021; Wagner, 2022; Chen et al., 2022). The perception constraint requires the recovered signal to be realistic by measuring the discrepancy between its distribution and the distribution of the original signal. The RDP problem was analyzed, by extending the rate-distortion theory approach (Theis & Wagner, 2021; Chen et al., 2022; Wagner, 2022). Theis & Wagner (2021) also presents one shot coding result, which is similar in spirit to scalar quantization.

A problem closely related to RDP coding is coding for a target probability distribution, which reduces to RDP coding when the target distribution is in fact RDP-optimizing. This more general setting had in fact been studied earlier (Saldi et al., 2013; 2014), and information theoretic results were given. The same problem was also considered under the moniker of channel synthesis (Cuff, 2013), and the work (Li & El Gamal, 2018; Theis & Ahmed, 2022) can also be viewed as treating this problem.

The sampling-based approach proposed in Li & El Gamal (2018) has often been invoked to prove rate-distortion type of results in the RDP setting, but it was well understood that such approach is fundamentally expensive in terms of computation (Agustsson & Theis, 2020). On the more practical front, dither-based approach is attractive and has been adopted in Agustsson & Theis (2020); Zhang et al. (2021); Yang et al. (2022). Dithered quantization with common randomness has a long history that traces back to Ziv (1985); see also Roberts (1962). Though it has some desirable properties (Zamir & Feder, 1992), which make it a convenient choice for some neural network based systems (Agustsson & Theis, 2020), its performance is in general inferior to deterministic quantizers when there is no perception consideration (Theis & Agustsson, 2021). Integrated with non-linear post mapping which preserves the perfect perception quality, dithered quantization was studied as a heuristic in Li et al. (2010). Dithered quantization provides a seemingly natural match for RDP coding, given the importance of utilizing common randomness in this setting (Chen et al., 2022; Wagner, 2022), and the illustrative example given in (Theis & Agustsson, 2021) appears to confirm the benefit of this approach in the RDP setting. However, the underlying mechanism of this advantage appears rather opaque.

B Proofs of Section 3

Proof of Theorem 3.1. Since each of N quantifiers are uniform with L levels, the rate for the corresponding MultiQuan procedure is $\log L$. Due to symmetry, we analyze the distortion with a fixd 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. The distortion can then be calculated as

$$\frac{L}{2\pi}\frac{LN}{2\pi}\int_{-\pi/L}^{\pi/L}\left(\int_{-\pi/(NL)}^{\pi/(NL)}\|(\cos(\theta),\sin(\theta))-(\cos(\alpha),\sin(\alpha))\|^2\mathrm{d}\alpha\right)\mathrm{d}\theta \tag{8}$$

$$=\frac{L}{2\pi}\frac{LN}{2\pi}\int_{-\pi/L}^{\pi/L}\left(\int_{-\pi/(NL)}^{\pi/(NL)}2(1-\cos(\theta-\alpha))\mathrm{d}\alpha\right)\mathrm{d}\theta \tag{9}$$

$$=2+\frac{L^2N}{2\pi^2}\int_{-\pi/L}^{\pi/L}\sin(\theta-\pi/(NL))-\sin(\theta+\pi/(NL))\mathrm{d}\theta \tag{10}$$

$$=2+\frac{L^2N}{\pi^2}\left(\cos(\frac{\pi}{L}\frac{N+1}{N})-\cos(\frac{\pi}{L}\frac{N-1}{N})\right) \tag{11}$$

$$= 2 - 2\frac{\sin(\pi/(NL))}{\pi/(NL)}\frac{\sin(\pi/L)}{\pi/L}, \tag{12}$$

which is the desired result. □

Proof of Theorem 3.2. We aim to minimize the rate distortion Lagrangian with perfect perceptual quality for any Lagrange multiplier $\lambda > 0$, i.e,

$$\min_{p_{\hat{X}|X}:\hat{X}\overset{d}{=}X} I(X;\hat{X}) + \lambda\mathbb{E}[\|X-\hat{X}\|^2]. \tag{13}$$

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 representation $\theta(\hat{X})$. The MSE distortion term can be written as

$$\|X-\hat{X}\|_2^2 = \Big\|(\cos(\theta(X)),\sin(\theta(X))) - (\cos(\theta(\hat{X})),\sin(\theta(\hat{X})))\Big\|_2^2 = 2 - 2\cos(\theta(X)-\theta(\hat{X})). \tag{14}$$

The mutual information can be lower bounded by

$$I(X;\hat{X}) = h(X) - h(X|\hat{X}) \geq h(X) - h(X-\hat{X}) = h(\theta(X)) - h(\theta(X)-\theta(\hat{X})). \tag{15}$$

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 following optimization problem, which is equivalent to lower-bounding (13)

$$\text{minimize}_{p(\beta)} \; -h(\beta) + 2\lambda\mathbb{E}[(1-\cos(\beta))] = 2\lambda + \int_{-\pi}^{\pi} p(\beta)[\log(p(\beta)) - 2\lambda\cos(\beta)]\mathrm{d}\beta. \tag{16}$$

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 an lower bound to (13).

To show that they are in fact equal, we only need to observe that in (15), the only inequality can be written as

$$\begin{aligned} I(X;\hat{X}) &= h(X) - h(X|\hat{X}) = h(\theta) - h(\theta|\hat{\theta}) = h(\theta) - h(\theta-\hat{\theta}|\hat{\theta}) \\ &= h(\theta) - h(\beta|\hat{\theta}) \geq h(\theta) - h(\beta). \end{aligned} \tag{17}$$

However observe that we have

$$p_{\beta|\hat{\theta}}(\beta|\hat{\theta}) = \frac{p_{\beta,\hat{\theta}}(\beta,\hat{\theta})}{p_{\hat{\theta}}(\hat{\theta})} = \frac{p_{\beta,\theta}(\beta,\hat{\theta}-\beta)}{p_{\hat{\theta}}(\hat{\theta})} = \frac{p_{\beta}(\beta)p_{\theta}(\hat{\theta}-\beta)}{p_{\hat{\theta}}(\hat{\theta})} = p_{\beta}(\beta), \tag{18}$$

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 (17) becomes an equality, which establishes overall equality. Thus the rate-distortion pairs are indeed characterized by

$$\left\{(R,D) = \Big(\log(2\pi) - h(Z), \mathbb{E}[2-2\cos(Z)]\Big) : Z \sim p(z;\lambda) = \frac{e^{\lambda\cos(z)}}{\int_{-\pi}^{\pi} e^{\lambda\cos(z')}\mathrm{d}z'},\ \lambda > 0\right\}.$$

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. □

Proof of Theorem 3.3. 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 an integer and then reconstructed by $\hat{X} = g(f(X,V),V)$, where V is the common randomness. This is the most general class of codec, and we will show the optimal rate and distortion tradeoff within this class.

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)

$$\begin{aligned} &H(f(X;V)|V) + \lambda \mathbb{E}_{X,V}[d(X, g(f(X;V);V))] \\ &= \mathbb{E}_V[\mathbb{E}_X[-\log(\mathbb{P}(f(X;V)|V)) + \lambda d(X, g(f(X;V);V))|V]] \end{aligned} \tag{19}$$

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.

To start with, first note that the optimal deterministic quantizer must have contiguous regions, i.e., the region in $\mathcal{X}$ of the same index $f(\cdot, v)$ should be contiguous. To see this, consider a partition F of the unit circle that has non-contigous cells; however, a different partition with only contiguous cells such that each index has an inverse image of the same measure as that in F, and reconstruction point in the center will strictly improve the distortion. Now, for such a quantizer with L levels, i.e., $|f(\cdot, v)| = L$, and we claim that it must be uniform quantizer on the unit circle. Consider two adjacent Voronoi cells. Suppose the two adjacent regions has 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 following function

$$l(\alpha; r) = (r - \alpha)\ln(r - \alpha) + \frac{\lambda}{\pi}\sin(\pi(r - \alpha)) + \alpha \ln(\alpha) + \frac{\lambda}{\pi}\sin(\pi\alpha). \tag{20}$$

Its derivative is

$$l'(\alpha; r) = -\ln(r - \alpha) - \lambda\cos(\pi(r - \alpha)) + \ln(\alpha) + \lambda\cos(\pi\alpha) \tag{21}$$

and its second derivative is

$$l''(\alpha; r) = \frac{1}{r - \alpha} + \frac{1}{\alpha} - \lambda\pi(\sin(\pi(r - \alpha)) + \sin(\pi\alpha)). \tag{22}$$

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 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 $\alpha \to r$.

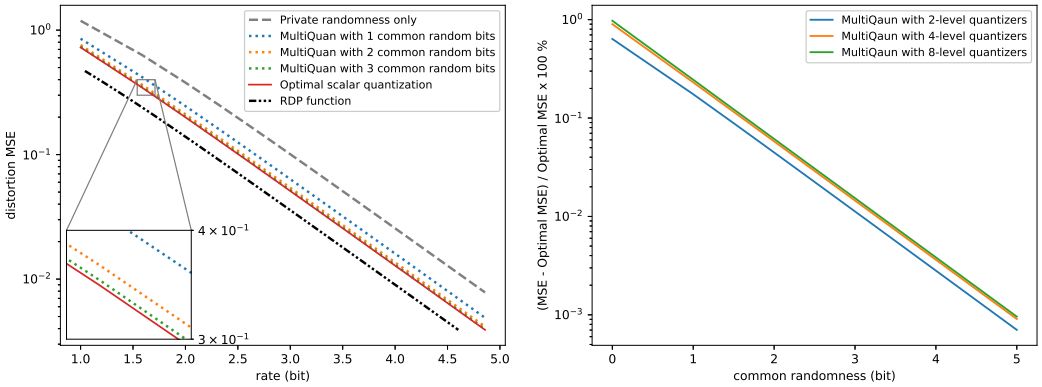
Therefore any two adjacent non-empty Voronoi cells have the same size. The optimal quantizer thus must have equal sized Voronoi cells, thus a uniform quantizer. The optimal scalar quantization (single shot coding) trade-off between the coding rate and the distortion is the piece-wise linear function with the following extreme points

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

The piece-wise linear function above is a lower bound, when considering perfect perceptual quality. However, it is straightforward to verify that dithered quantizations has 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) of dithered quantizers. □

C MORE DISCUSSIONS ON THE UNIT-CIRCLE SETTING

We observe a sharp difference between the information-theoretic rate-distortion function (optimal asymptotic rate-distortion tradeoff) and the performance of the optimal scalar quantization. Though



(a) Unit-circle rate-distortion log plot

(b) Common randomness v.s. distortion difference

Figure 7: Unit-circle setting: log plots

both distortions approach to zero as the rate increases, their ratio are roughly the same as shown in the Figure 7a. In other words, to obtain the same distortion as the joint coding of infinite number of i.i.d. data, scalar quantization requires 0.2-0.4 extra bits per dimension. Moreover, Figure 7a shows that MultiQuan converges uniformly to the optimal scalar quantization as the amount of common randomness increases. A linear convergence rate is also exhibited, i.e., the distortion difference between MutliQuan and the optimal scalar quantization converges to zero exponentially fast as the amount of common randomness increases.

D Optimzing MultiQuan Thresholds

We introduce two heuristic iterative algorithms to optimize the thresholds in the MultiQuan procedure. Recall that MultiQuan consists of N quantizers. Denote by $a_n(i)$ and $b_n(i)$ the lower boundary and upper boundary for quantization cell-i of quantizer-n, which are the decision variables to optimize. For the uniform offset quantizers introduced earlier, $a_n(i) = \left(i - \frac{1}{2} + \frac{n}{N}\right)\Delta$ and $b_n(i) = a_n(i) + \Delta$, which are used as initialization in the iterative algorithms. The boundaries of the reconstructed signal corresponding to cell-i of quantizer-n is $a(iN+n)$ and $b(iN+n)$, which can be calculated as in (6).

The source data X has a probability density $p(x)$. With a slight abuse of notation, denote $p(a,b) = \mathbb{P}(X \in [a,b))$. The objective function is the distortion-rate Lagrangian $L_\lambda = D + \lambda R$, i.e.,

$$\frac{1}{N}\sum_{n=0}^{N-1}\sum_{i}\left(\int_{a_n(i)}^{b_n(i)}\int_{a(iN+n)}^{b(iN+n)}(x-\tilde{z})^2 p(x)p(\tilde{z})\mathrm{d}\tilde{z}\mathrm{d}x + \lambda p(a_n(i),b_n(i))\log\frac{1}{p(a_n(i),b_n(i))}\right).$$

The Lagrangian L_λ is a function of the boundaries $(a_n(i),b_n(i))_{n\in[0:N-1],i\in\mathbb{Z}}$. Optimal tradeoff between the rate and the distortion can be achieved by minimizing the Lagrangian L_λ by varying $\lambda > 0$.

D.1 Gradient descent based approach

To apply gradient descent on the Lagrangian L_λ, it suffices to calculate the gradient. Since the cells of quantizer-n are adjacent to one another from left to right, the upper boundary for quantization cell-i is the same as the lower boundary for quantization cell-i, i.e., $b_n(i-1) = a_n(i)$. We can thus view $b_n(i-1)$ and $a_n(i)$ as one decision variable ρ.

Define $\bar{x}_{iN+n} = \bar{x}_{n,i} \triangleq \frac{\int_{a_n(i)}^{b_n(i)} p(x)\mathrm{d}x}{p(a_n(i),b_n(i))}$ and $\bar{z}_{iN+n} = \bar{z}_{n,i} \triangleq \frac{\int_{a(iN+n)}^{b(iN+n)} p(x)\mathrm{d}x}{p(a(iN+n),b(iN+n))}$, where index $iN+n$ matches the boundary functions $(a(\cdot),b(\cdot))$ of reconstruction function and (n,i) matches

the quantizer cells' boundary function $(a.(\cdot), b.(\cdot))$. Let $\ell_n(i)$ denote the codeword length for cell-i in quantizer-n.

For an interval $[a, b]$ denote by $\bar{x}_{a,b} \triangleq \frac{\int_a^b p(x)\mathrm{d}x}{p(a,b)}$ the expectation of X conditioned on that X lies in the interval. The partial derivative $\frac{\partial L_\lambda}{\partial \rho}$ can be calculated as

$$
\begin{aligned}
p(\rho)\Bigg(& \sum_{j=(i-1)N+n+1}^{iN+n-1} (\bar{z}_j - b(j))^2 - (\bar{z}_j - a(j)^2 - 2(b(j) - a(j))(\bar{x}_j - \bar{z}_j) \\
& + (\rho - \bar{x}_{n,i-1})^2 - (\bar{x}_{n,i-1} - \bar{z}_{n,i-1})^2 + (\bar{z}_{n,i-1} - b((i-1)N + n))^2 \\
& - (\rho - \bar{x}_{n,i})^2 + (\bar{x}_{n,i} - \bar{z}_{n,i})^2 - (\bar{z}_{n,i} - a(iN + n))^2 + \lambda \ell_n(i-1) - \lambda \ell_n(i) \Big).
\end{aligned}
$$

Note that valid boundaries of offset quantizers satisfy

$$
b_0(i-1) = a_0(i) \leq a_1(i) \leq \cdots \leq a_{N-1}(i) \leq b_0(i) = a_0(i+1).
$$

The thresholds after the descent update need to be valid, i.e., maintain the sequential order above. If any of the above inequalities is violated, we can either select a smaller stepsize using backtracking line search or project it to the set of valid boundaries.

D.2 Pseudo-Lloyd based approach

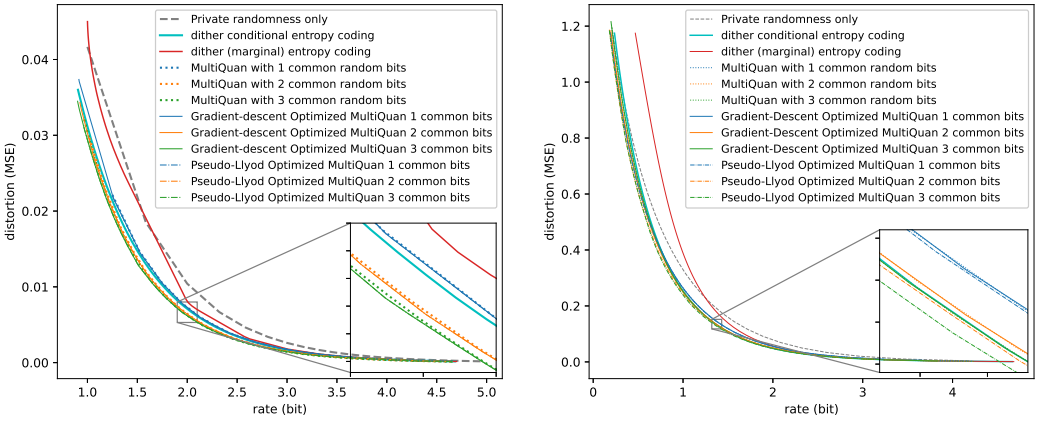
The classic Lloyd algorithm iterates among the optimization of three components: the thresholds, the reconstruction values, and the codeword length for each coding index. In our setting, a reconstruction interval takes of the role a reconstruction value in the classic setting. To be more concrete, we assume the probability density function exists and has a finite support on the interval $[L_0, U_0]$. Furthermore, assume the distortion is measure by mean squared error. We can adopt the following pseudo-Lloyd algorithm shown in Algorithm 1.

Algorithm 1 Pseudo-Lloyd for MultiQuan

Require: $N, \Delta, p(x)$
  Initialize $a_n(i) = \left(i - \frac{1}{2} + \frac{n}{N}\right)\Delta$ and $b_n(i) = a_n(i) + \Delta$, $\forall i, \forall n = 0, 1, 2, \ldots, N-1$;
    ▷ Quantization intervals
  Initialize $a(j), b(j), \forall j$, as given in (6); ▷ Reconstruction intervals
  Initialize $\ell_n(i) = -\log p(a_n(i), b_n(i))$, $\forall i, \forall n = 0, 1, 2, \ldots, N-1$;
    ▷ Codeword lengths for i-th cell in quantizer-n
  while Stopping criteria not satisfied **do**
    for i and $\forall n = 0, 1, 2, \ldots, N-1$ **do**
      $\underline{a} \leftarrow a_{(n-1) \pmod N}(i + \lfloor (n-1)/N \rfloor)$, $\bar{a} =\leftarrow a_{(n+1) \pmod N}(i + \lfloor (n+1)/N \rfloor)$
      Let

$$
f(t) = \frac{\int_{a((i-1)N+n)}^{b((i-1)N+n)} (x-t)^2 p(x) dx}{\int_{a((i-1)N+n)}^{b((i-1)N+n)} p(x) dx} + \lambda \ell_n(i-1) - \frac{\int_{a(iN+n)}^{b(iN+n)} (x-t)^2 p(x) dx}{\int_{a(iN+n)}^{b(iN+n)} p(x) dx} - \lambda \ell_n(i)
$$

      Solve for t such that $f(t) = 0$ in $[\bar{a}, \bar{b}]$
      if t exists **then**
        $a_n(i) \leftarrow t$; $b_n(i-1) = a_n(i)$;
      else if $f(\bar{a}) > 0$ **then**
        $a_n(i) \leftarrow \bar{a}$; $b_n(i-1) = a_n(i)$;
      else
        $a_n(i) \leftarrow \underline{a}$; $b_n(i-1) = a_n(i)$;
      end if
    end for
    Update $a(j), b(j), \forall j$, as given in (6);
    Update $\ell_n(i) = -\log p(a_n(i), b_n(i))$, $\forall i, \forall n = 0, 1, 2, \ldots, N-1$;
  end while



(a) Uniform distribution setting

(b) Exponential distribution setting

Figure 8: Simple source with perfect perceptual quality

In the pseudo code above, we have omitted some details, particularly regarding the updates of thresholds between two cells with measure zero, which clearly does not require any optimization. Here we adopt Lloyd's update in analogous manner to update the thresholds, however it does not guarantee optimality, unlike the classic setting. Nevertheless, experimental results show that the approach can lead to some performance improvements.

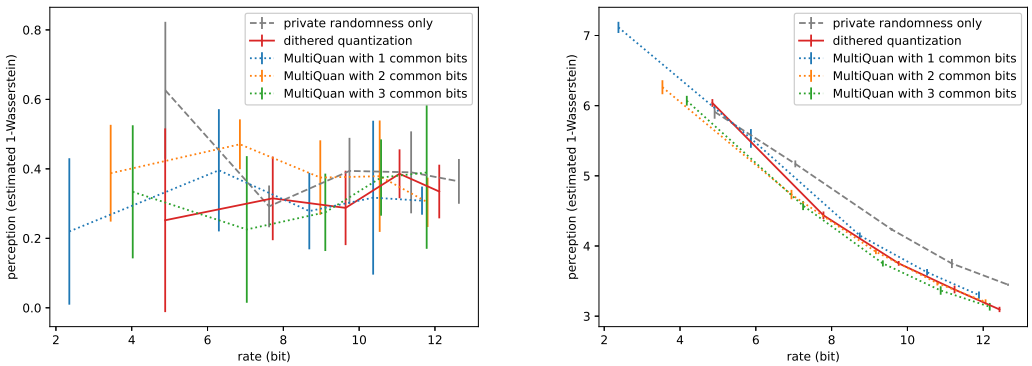
E Experimental Results and Details

E.1 Simple source with perfect perceptual quality

In Figure 8, we include further results with dithered quantization, assuming conditional entropy coding is allowed, i.e., the rate is the conditional entropy $H(f(X+Z)|Z)$, which can be achieved by infinitely many entropy codes. The performances of the MultiQuan quantizers optimized using the gradient-descent based method and pseudo-Llyod based method are also given. The Gradient-descent is updated for 20 iterates with step size 0.001 and stops early if the updated thresholds no longer satisfy the offset quantizers. The pseudo-Llyod algorithm iteratively optimizes the thresholds until the change of rate-distortion Lagrangian at one update step is within 1e-5.

Uniform distribution setting: As shown in Fig. 8a, dithered quantization with conditional entropy coding shows a significant rate improvement over the dithered quantization with marginal entropy coding. Unlike the latter, it is always superior to that of quantization with only private randomness at any rate. Its performance surpasses that of the MultiQuan procedure with 1 common random bit, however is inferior to MultiQuan with 2 or 3 common random bits. Gradient-based and Pseudo-Lloyd algorithms improve the performance of the MultiQuan procedure. These two algorithms appear to perform similarly, and their performances cannot be distinguished in the plot (the solid lines of Gradient-descent optimized MultiQuan overlap with the dashdot lines of Pseudo-Lloyd optimized MultiQuan). The best scalar quantization currently achieved is through MultiQuan with optimized thresholds, when there are at least 3 common random bits. Its performance is clearly better than that of the dithered quantization, even assuming unrealizable conditional entropy coding in the dithered approach.

Exponential distribution setting: Similar to the uniform distribution setting, in Fig. 8b, dithered quantization with conditional entropy coding shows a significant improvement over dithered quantization using marginal entropy coding. It also outperforms quantization without common randomness at most rates, but is still inferior in the low rate region. In certain rate region, as shown in the zoom box, its performance sometimes surpasses that of MultiQuan procedures without optimization. Gradient-descent does not make much improvement in this setting on MultiQuan, while



(a) $\lambda = 0.08$ (b) $\lambda = 0.005$

Figure 9: Estimated perception error with different lambda

pseudo-Lloyd algorithms is considerably more effective. The best rate-distortion tradeoff is achieved by pseudo-Lloyd optimized MultiQuan with 3 common bits. We believe, due to the non-convexity of the underlying optimization problem, the gradient-descent algorithm tends to stuck at a critical point near the uniform-offset-quantizer initialization point.

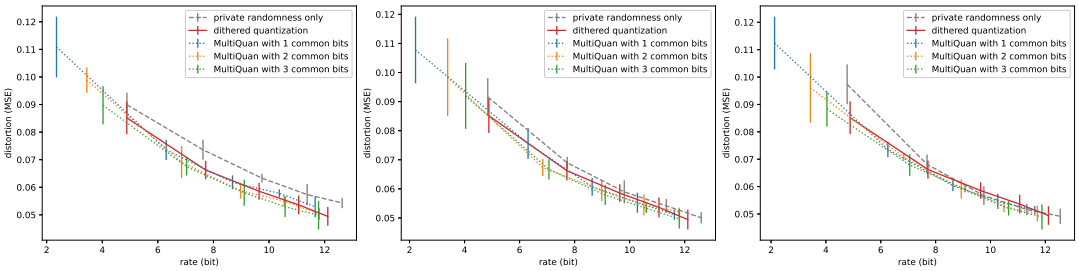
E.2 Neural network based image compression

We test different quantization procedures for neural network-based image quantization on the MNIST dataset (LeCun et al., 1998). The training follows auto encoder-decoder with Wasserstein GAN regularization, which is consisted of three trainable components $(f(\cdot;\omega_f), g(\cdot;\omega_g), c(\cdot;\omega_c))$ with parameters $(\omega_f, \omega_g, \omega_c)$ and a given scalar quantization procedure Q. The trainable components follows the same architecture as in Zhang et al. (2021). We take Lagrange multiplier $\lambda = 0.08$ for the distortion-perception Lagrangian, and perform adversarial training for a total of 30 epochs. The learning rate was decayed by a factor of 5 after 20 epochs. All models were trained with the Adam optimizer, and the batch size used was 64. The results are averaged over 5 repetitions with different random seeds, and the error bar is $\pm$ standard deviation.

The hidden space has dimension 5, and the value in each coordinate is within $[-1, 1]$. The scalar quantization procedure is applied to each coordinate. Specifically, three types of quantization procedures are compared. All the procedures rely on encoding with uniform thresholds. The step sizes are chosen as $\Delta = 2/L$, where $L = 2, 3, \ldots, 6$ for plotting the tradeoff.

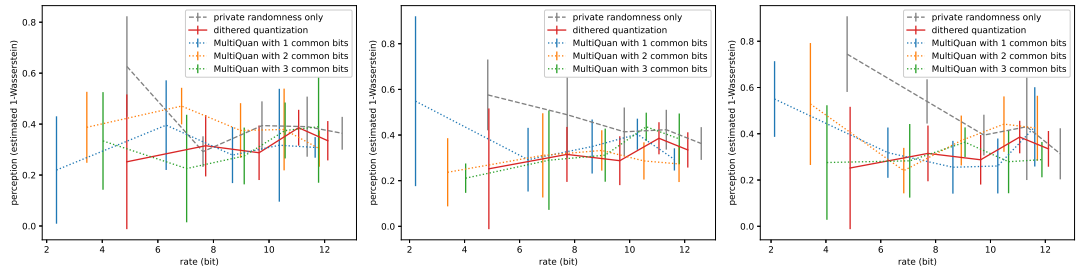
- Quantization without common randomness: As MultiQuan without common randomness, $N = 1$ and the encoding function is the deterministic f_0 as in (5). Since the distribution of the signal in the low dimensional space is not known, noise $\tilde{Z}$ uniformly distributed in the interval $[-\Delta/2, \Delta/2]$ is added to the reconstruction point so as to maintain perceptual quality.
- Dithered quantization: dithered quantization follows the same as scalar setting, where the generator function $g(\cdot;\omega_g)$ serves as a post-processing function that maintain perceptual quality.
- MultiQuan: For MultiQuan with N quantizers, each quantizer performs similarly as quantizer without common randomness, but with additive private noise $\tilde{Z}$ within $[-\Delta/(2N), -\Delta/(2N)]$.

Lagrange multiplier is sufficiently large to maintain high perceptual quality: The Lagrange multiplier chosen as 0.08 is five times larger than the largest Lagrange multiplier 0.015 chosen in Zhang et al. (2021), and also much larger than the Lagrange multiplier 0.0001 chosen in Liu et al. (2022), where perfect perceptual quality is required. Perceptually, the reconstructed images resemble the true images in the MNIST dataset. Moreover, the 1-Wasserstein metric estimated by



(a) Noise scale 1.0 (b) Noise scale 0.6 (c) Noise scale 0.2

Figure 10: Impact of inserted noise on the distortion



(a) Noise scale 1.0 (b) Noise scale 0.6 (c) Noise scale 0.2

Figure 11: Impact of inserted noise on the perception

the critic network $c(\cdot;\omega_c)$ is shown in Fig. 9. We observe that if the λ is not chosen large enough (e.g., $\lambda = 0.005$), the estimated 1-Wasserstein metric would decrease as the rate increases shown in Figure 9b. In contrast, under sufficiently large Lagrange multiplier ($\lambda = 0.08$), the estimated metric close to zero and does not decrease as the rate increases. These observations serve as evidence that convinces us the Lagrange multiplier has been chosen sufficiently large to maintain high perceptual quality.

Impact of the amount of inserted noise: Besides the common randomness, the inserted noise at the decoder side is another source of randomness during the coding process. It is generally believed that this noise should be inserted at the decoder side for quantization without common randomness to maintain high perceptual quality, but how large? We control the scale of noise and illustrate the impact of the inserted noise as well as its impact together with the common randomness. Note that dithered quantization, as a universal quantization, does not have the flexibility to reduce this noise, since the only source of randomness is the common randomness. However, we can control the size of the inserted noise in quantization without common randomness and the MultiQuan procedure, simply by introducing a multiplicative factor between $[0, 1]$ on the inserted noise. Let us call this factor noise scale. When the noise scale is 0.6, inserted noise $\tilde{Z}$ is uniformly distributed over $[-0.3\Delta, 0.3\Delta]$, which reduce the variance of the input to the generator $g(\cdot;\omega_g)$. Fig. 10 shows that, quantization without common randomness and MultiQuan procedures may achieve lower distortion due to reduced noise scale. The private noise only setting even catch up with the dithered quantization when the rate is high. One reason is that when the rate is high, the coded vector in the hidden space itself provide sufficient variations to maintain high perceptual quality without the need of inserted noise. On the other hand, when the rate is low, reducing the inserted noise may have negative impact on the distortion of quantization without common randomness.