ON THE ADVANTAGES OF STOCHASTIC ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic encoders have been used in rate-distortion theory and neural compression because they can be easier to handle. However, in performance comparisons with deterministic encoders they often do worse, suggesting that noise in the encoding process may generally be a bad idea. It is poorly understood if and when stochastic encoders can do better than deterministic encoders. In this paper we provide one illustrative example which shows that stochastic encoders can significantly outperform the best deterministic encoders. Our toy example suggests that stochastic encoders may be particularly useful in the regime of "perfect perceptual quality".

1 INTRODUCTION

In lossy compression, information is typically deterministically encoded through the following procedure. First, the input is transformed using a fixed analysis transform or a learned neural network. This is followed by a quantization step mapping continuous to discrete values. Finally, the remaining information is further compressed and encoded into bits using entropy encoding. Taken together, these steps implement a deterministic function from the input to the integers. In this paper we discuss whether it is ever advantageous to stochastically encode data by introducing noise into this process.

To make this discussion more precise, let

$$\mathcal{F} = \{f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{N}_0\}, \quad \mathcal{G} = \{g : \mathbb{N}_0 \times \mathbb{R} \to \mathbb{R}^n\} \tag{1}$$

be sets of *stochastic encoders* and *decoders*. The encoder expects n-dimensional data as well as an additional source of randomness as input. This source of randomness will be represented by a random variable distributed uniformly between zero and one, which we can think of as an infinite number of random bits when expressed in binary form ($U = 0.B_1B_2B_3...$). The decoder in turn expects the output of an encoder and a source of randomness as input. This could either be the same random input as for the encoder or a different source (Fig. 1). By deterministic encoders we mean

$$\tilde{\mathcal{F}} = \{f \in \mathcal{F} \mid \exists \tilde{f} : \forall \mathbf{x}, u : f(\mathbf{x}, u) = \tilde{f}(\mathbf{x})\}, \tag{2}$$

that is, encoders which ignore the random inputs. Defined in this way, it is clear that stochastic encoders can perform no worse than deterministic encoders since $\tilde{\mathcal{F}} \subset \mathcal{F}$. However, for a given loss and data distribution we can ask whether an optimal encoder can be found in $\tilde{\mathcal{F}}$ or whether we should hope to do better by considering the larger set $\mathcal{F}$.

Rate-distortion theory considers stochastic encoders because they can be more amenable to a theoretical analysis. For example, *universal quantization* is a randomized quantization procedure whose error distribution is independent of the source (Ziv, 1985), making it easier to analyze than deterministic quantization. In machine learning, stochastic encoders recently have been considered because they allow for the transmission of continuous signals and thus enable differentiable losses (Havasi et al., 2019). While this can make it easier to optimize stochastic encoders, the introduction of noise can also hurt performance (Agustsson & Theis, 2020).

For the Laplace and exponential distributions, optimal quantizers exist which are deterministic (Sullivan, 1996). Universal quantization is known to introduce redundancy compared to the best deterministic quantizer at the same squared error distortion level (Ziv, 1985). More generally we can say that the Lagrangian of a rate-distortion problem is always minimized by a deterministic encoder (Appendix A). This begs the question whether stochastic encoders are ever better than deterministic

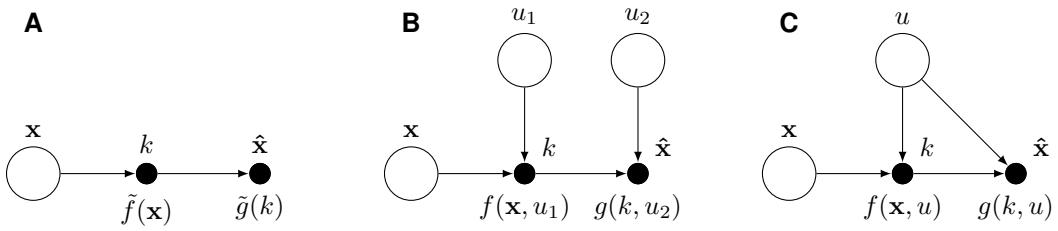

Figure 1: Visualizations of deterministic and stochastic encoders and decoders. Empty circles represent sources of randomness, filled circles represent variables which depend deterministically on their inputs. **A**: A deterministic encoder and a deterministic decoder. **B**: A stochastic encoder and a stochastic decoder with independent sources of randomness. **C**: A stochastic encoder and a stochastic decoder with a shared source of randomness. Setting B generalizes A and C generalizes B. (A single uniform random variable can be split into two independent variables. For example, the encoder can use the odd numbered bits and the decoder the even numbered bits.)

encoders. Wagner & Ballé (2021) provide one example where deterministic encoders are only optimal at certain levels of distortion. For levels in between, a better rate can be achieved by randomly choosing one of two deterministic codecs, proving that stochastic encoders are indeed sometimes needed for optimal performance. Here we consider another setting where stochastic encoders appear to provide further advantages.

2 Perfect perceptual quality

The setting we consider here is the setting of "perfect perceptual quality" (Blau & Michaeli, 2018). We still wish to optimize a rate-distortion trade-off but under the constraint that reconstructions follow the data distribution, that is, the reconstructions are guaranteed to be free of any artefacts. More formally, let $\hat{\mathbf{X}} = g(f(\mathbf{X}, U), U)$. We seek to

$$\text{minimize} \quad \mathbb{E}[d(\mathbf{X}, \hat{\mathbf{X}})] \quad \text{w.r.t.} \quad f \in \mathcal{F}_R,\, g \in \mathcal{G} \quad \text{s.t.} \quad \hat{\mathbf{X}} \sim \mathbf{X}. \tag{3}$$

For simplicity, we will only consider fixed-rate encoders

$$\mathcal{F}_R = \{f : \mathbb{R}^n \times \mathbb{R} \to \{0, \dots, 2^R - 1\}\}. \tag{4}$$

Tschannen et al. (2018) also called this problem "distribution-preserving lossy compression". Their approach combined deterministic encoders with stochastic decoders. In the next section we demonstrate how also making the encoder stochastic can be advantageous in this setting.

3 Circular uniform distribution

Consider a uniform distribution over a unit circle. It will be convenient to use polar coordinates (θ, r) in addition to Cartesian coordinates (x_1, x_2). In polar coordinates, we can write down the marginal density

$$p(\theta) = \frac{1}{2\pi}. \tag{5}$$

By our requirement of perfect perceptual quality, any reconstruction $\hat{\mathbf{x}}$ will have to live on the unit circle as well, that is, $\hat{r} = r = 1$. Any distortion can therefore be expressed as a function of θ and $\hat{\theta}$ and we choose

$$d(\theta, \hat{\theta}) = 1 - \cos(\hat{\theta} - \theta) = 1 - \hat{\mathbf{x}}^\top \mathbf{x} = \frac{1}{2}||\hat{\mathbf{x}} - \mathbf{x}||^2, \tag{6}$$

which is the cosine distance or equivalently the mean squared error. We further target a bit rate of 1, that is, $f(\mathbf{x}, u) \in \{0, 1\}$.

We will now compare two approaches for this example, vector quantization (which uses a deterministic encoder) and universal quantization (Ziv, 1985) (which uses a stochastic encoder).

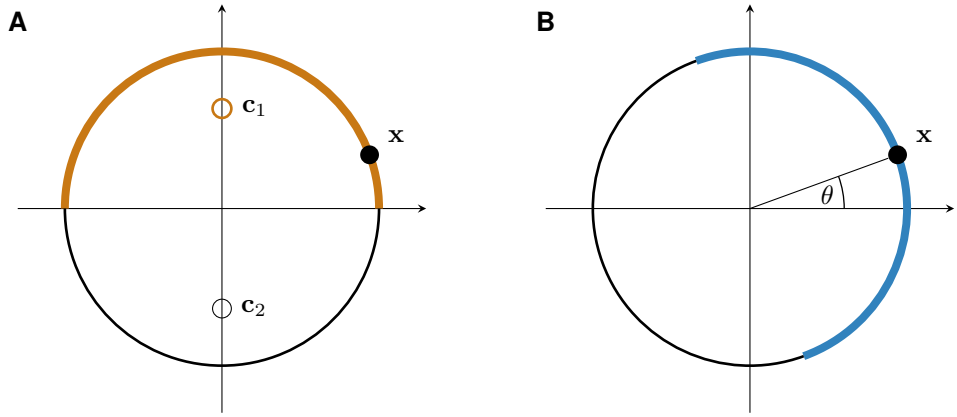

Figure 2: **A**: Visualization of deterministic 1-bit encoding of data drawn uniformly from the circle. Vector quantization maps data points of a Voronoi region (orange) to a codebook vector ($\mathbf{c}_1$). Alternatively, a stochastic decoder may sample uniformly from the Voronoi region to achieve *perfect perceptual quality*. Depending on the location of the data point, the decoder will sample from one of two regions. **B**: Universal quantization stochastically encodes data such that the reconstruction distribution (blue) is centered around the data point.

3.1 Vector quantization

Choose $\theta_0 = 0$ and $\theta_1 = \pi$. Let $f \in \tilde{\mathcal{F}}_1$ be the encoder with

$$f(\mathbf{x}, u) = \operatorname*{argmin}_{k \in \{0,1\}} d(\theta, \theta_k) \tag{7}$$

for any input $\mathbf{x}$ with polar coordinates $(\theta, 1)$. To satisfy the perceptual quality constraint, we decode by uniformly sampling from the Voronoi section corresponding to k,

$$g(k, u) = \begin{pmatrix} \cos\hat{\theta} \\ \sin\hat{\theta} \end{pmatrix} \quad \text{where} \quad \hat{\theta} = \theta_k + \pi(u - \frac{1}{2}). \tag{8}$$

For a fixed $\theta \in [-\pi/2, \pi/2)$, the error is

$$\mathbb{E}[d(\theta, \hat{\Theta}) \mid \theta] = 1 - \int_{\pi/2}^{-\pi/2} \frac{1}{\pi} \cos(\theta - \hat{\theta})\, d\hat{\theta} = 1 - \frac{2}{\pi} \cos(\theta), \tag{9}$$

and similarly for other angles. The expected distortion is

$$\mathbb{E}[d(\Theta, \hat{\Theta})] = 1 - \frac{4}{\pi^2}. \tag{10}$$

Appendix C shows that this is the best performance we can hope for with a deterministic encoder.

3.2 Universal quantization

Universal quantization (Ziv, 1985) is a simple trick which allows us to communicate a uniform sample using bits. Given $y \in \mathbb{R}$ and a random variable U uniformly distributed over any unit interval, the encoder computes

$$K = \lfloor y + U \rceil. \tag{11}$$

The decoder receives K and computes the reconstruction

$$\hat{Y} = K - U. \tag{12}$$

It can be shown that $\hat{Y} \sim y + U'$ where U' is uniformly distributed between -0.5 and 0.5. Note that universal quantization requires that the encoder and decoder both have access to U. We will adapt this trick to communicate a noisy version of the input angle θ. For the encoder, we choose

$$f(\mathbf{x}, u) = \left\lfloor \frac{\theta}{\pi} + u \right\rceil \bmod 2. \tag{13}$$

The decoder receives $k \in \{0, 1\}$ and computes

$$g(k, u) = \begin{pmatrix} \cos \hat{\theta} \\ \sin \hat{\theta} \end{pmatrix} \quad \text{where} \quad \hat{\theta} = \pi \left(k - u \right). \tag{14}$$

The modulo operation is only changing the decoder output by a multiple of 2. Hence, for a fixed input with angle θ, the reconstruction is distributed as follows:

$$\hat{\Theta} = \pi(k - U) = \pi \left(\left\lfloor \frac{\theta}{\pi} + U \right\rceil - 2J - U \right) \sim \pi \left(\frac{\theta}{\pi} + U' - 2J \right) = \theta + \pi U' - 2\pi J \tag{15}$$

for some integer valued random variable J which depends on θ and U. Note that we only care about θ up to multiples of 2π since $d(\theta, \hat{\theta} + 2\pi) = d(\theta, \hat{\theta})$. Figure 2C illustrates the conditional distribution of reconstructions when using universal quantization. Note that the distribution is centered around the data point, unlike for the deterministic encoder.

The expected distortion is independent of θ and we have

$$\mathbb{E}[d(\theta, \hat{\Theta}) \mid \theta] = 1 - \int_{\theta - \pi/2}^{\theta + \pi/2} \frac{1}{\pi} \cos(\theta - \hat{\theta}) \, d\hat{\theta} = 1 - \frac{2}{\pi} < 1 - \frac{4}{\pi^2}. \tag{16}$$

This distortion is about 38.9% smaller than with a deterministic encoder. Appendix D compares the two approaches further at other bit rates.

4 Discussion

Stochastic encoders can be much more convenient to optimize than deterministic encoders. It is thus important to understand the impact of noise on performance. When only optimizing for distortion, the potential for stochastic encoders to outperform deterministic encoders is limited (Appendix A). In practice, stochasticity can even degrade the performance of neural encoders (Agustsson & Theis, 2020).

On the other hand, here we showed that stochasticity can help significantly when taking into account perceptual quality (measured in terms of the reconstructions' distribution). This justifies using stochastic encoders beyond reasons of convenience and differentiability. In Appendix B we further show that stochasticity is not enough but that *shared* randomness is needed to achieve an improvement over deterministic encoders in our example.

Our stochastic encoder relied on universal quantization. Universal quantization is of particular interest because it is easy to implement and uniform noise is commonly used in neural compression. The setting of perfect perceptual quality is also practically relevant since it corresponds to artefact free codecs.

However, here we only considered a simple example in a one-shot setting (that is, encoding a single data point). From rate-distortion theory we know that a deterministic encoder (vector quantization) asymptotically achieves minimal distortion if we encode many data points at once. An important question for future research is whether a similar achievability result holds in the setting of perfect perceptual quality or whether we should also expect noisy encoders to perform significantly better in a wide variety of realistic settings.

Acknowledgments

References

E. Agustsson and L. Theis. Universally Quantized Neural Compression. In *Advances in Neural Information Processing Systems 33*, 2020.

J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici. Nonlinear transform coding. *IEEE Trans. on Special Topics in Signal Processing*, 15, 2021.

Y. Blau and T. Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018.

Y. Blau and T. Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, 2019.

M. Havasi, R. Peharz, and J. M. Hernández-Lobato. Minimal Random Code Learning: Getting Bits Back from Compressed Model Parameters. In *International Conference on Learning Representations*, 2019.

G. J. Sullivan. Efficient scalar quantization of exponential and Laplacian random variables. *IEEE Transactions on Information Theory*, 42(5):1365–1374, 1996.

M. Tschannen, E. Agustsson, and M. Lucic. Deep Generative Models for Distribution-Preserving Lossy Compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

A. B. Wagner and J. Ballé. Neural Networks Optimally Compress the Sawbridge. In *2021 Data Compression Conf. (DCC)*, 2021. URL `https://arxiv.org/pdf/2011.05065`.

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

APPENDIX A

Here we show that under mild assumptions deterministic encoders are optimal for a given rate-distortion trade-off. This generalizes a result by Ballé et al. (2021) showing that a random offset used in dithering can always be replaced by a fixed offset which performs at least as well.

Theorem 1. *For every stochastic codec $(f, g) \in \mathcal{F} \times \mathcal{G}$ and trade-off parameter $\lambda \in [0, \infty)$, we can derive a deterministic codec which performs at least as well in terms of the expected rate-distortion trade-off.*

Proof. Let $f \in \mathcal{F}$ be a stochastic encoder and $g \in \mathcal{G}$ be a stochastic decoder. Let $U \sim U([0, 1))$ be a uniform random variable representing a shared source of randomness. Further, let

$$r : \mathbb{N}_0 \times \mathbb{R} \to [0, \infty], \quad (\mathbf{x}, u) \mapsto r(\mathbf{x}, u) \tag{17}$$

represent a number of bits, rate, or other coding cost which in general may depend on the state of the noise. Now consider the rate-distortion trade-off for a given trade-off parameter $\lambda \in [0, \infty)$,

$$\ell_\lambda(\mathbf{x}, u) = r(f(\mathbf{x}, u), u) + \lambda d(\mathbf{x}, g(f(\mathbf{x}, u), u)). \tag{18}$$

Under mild conditions on ℓ_λ (see Tonelli's theorem), we can write for the expected trade-off

$$l_\lambda = \mathbb{E}[\ell_\lambda(\mathbf{X}, U)] = \mathbb{E}_U[\mathbb{E}_\mathbf{X}[\ell_\lambda(\mathbf{X}, U)]]. \tag{19}$$

We know there must exist a $u^* \in [0, 1]$ such that $\mathbb{E}_\mathbf{X}[\ell_\lambda(\mathbf{X}, u^*)] \leq l_\lambda$, otherwise the expected trade-off would be larger. We therefore have

$$\mathbb{E}_U[\mathbb{E}_\mathbf{X}[\ell_\lambda(\mathbf{X}, U)]] \geq \mathbb{E}_\mathbf{X}[\ell_\lambda(\mathbf{X}, u^*)] \tag{20}$$

$$= \mathbb{E}_\mathbf{X}[\tilde{r}(\tilde{f}(\mathbf{X})) + \lambda d(\mathbf{X}, \tilde{g}(\tilde{f}(\mathbf{X})))], \tag{21}$$

where

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}, u^*) \tag{22}$$

is a deterministic encoder and $\tilde{g}$ and $\tilde{r}$ are defined analogously. That is, for every stochastic codec we can derive a deterministic codec which performs at least as well in terms of the expected rate-distortion trade-off. □

Interestingly, the same cannot be said if we consider the constrained optimization problem to

$$\text{minimize}_{f,g} \quad \mathbb{E}[r(f(\mathbf{X}, U), U)] \quad \text{s.t.} \quad \mathbb{E}[d(\mathbf{X}, g(f(\mathbf{X}, U), U))] < D. \tag{23}$$

Here, stochastic encoders may already perform better as Wagner & Ballé (2021) showed in an example. This can happen when the Pareto front of deterministic encoders is non-convex such that its Lagrangian (that is, the trade-off in Eq. 21) can only address a subset of the constrained optimization problems. However, Wagner & Ballé (2021) also showed that in these cases a simple randomized interpolation of two deterministic codecs can achieve optimal performance.

APPENDIX B

The following theorem shows that if the encoder and decoder do not have a access to a shared source of randomness (Fig. 1C), then requiring perfect perceptual quality necessarily leads to a twofold increase in squared error relative to a codec which only minimizes distortion.

Theorem 2. *Let $\mathcal{H}_R \subseteq \mathcal{F}_R \times \mathcal{G}$ be the subset of encoders f and decoders g such that $\hat{\mathbf{X}} \sim \mathbf{X}$ where*

$$U_1, U_2 \sim U([0,1)), \tag{24}$$
$$K = f(\mathbf{X}, U_1), \tag{25}$$
$$\hat{\mathbf{X}} = g(K, U_2), \tag{26}$$

and U_1 and U_2 are assumed to be independent. Then

$$\inf_{f,g \in \mathcal{H}_R} \mathbb{E}[||\mathbf{X} - \hat{\mathbf{X}}||^2] \geq 2 \inf_{f \in \mathcal{F}_R} \mathbb{E}[||\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K]||^2]. \tag{27}$$

Proof. We can divide the error into two components,

$$\begin{aligned}
&\mathbb{E}[||\mathbf{X} - \hat{\mathbf{X}}||^2] && (28)\\
&= \mathbb{E}[||\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K] + \mathbb{E}[\mathbf{X} \mid K] - \hat{\mathbf{X}}||^2] && (29)\\
&= \mathbb{E}[||\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K]||^2 + ||\mathbb{E}[\mathbf{X} \mid K] - \hat{\mathbf{X}}||^2 + 2(X - \mathbb{E}[\mathbf{X} \mid K])^\top (\mathbb{E}[\mathbf{X} \mid K] - \hat{\mathbf{X}})] && (30)\\
&= \mathbb{E}[||\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K]||^2] + \mathbb{E}[||\mathbb{E}[\mathbf{X} \mid K] - \hat{\mathbf{X}}||^2], && (31)
\end{aligned}$$

since

$$\begin{aligned}
&\mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K])^\top (\mathbb{E}[\mathbf{X} \mid K] - \hat{\mathbf{X}})] && (32)\\
&= \mathbb{E}_K[\mathbb{E}_{\mathbf{X},\hat{\mathbf{X}}}[(\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K])^\top (\mathbb{E}[\mathbf{X} \mid K] - \hat{\mathbf{X}}) \mid K]] && (33)\\
&= \mathbb{E}_K[\mathbb{E}_{\mathbf{X}}[\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K] \mid K]^\top \mathbb{E}_{\hat{\mathbf{X}}}[\mathbb{E}[\mathbf{X} \mid K] - \hat{\mathbf{X}} \mid K]] && (34)\\
&= \mathbb{E}_K[(\mathbb{E}[\mathbf{X} \mid K] - \mathbb{E}[\mathbf{X} \mid K])^\top (\mathbb{E}[\mathbf{X} \mid K] - \mathbb{E}[\hat{\mathbf{X}} \mid K])] && (35)\\
&= 0. && (36)
\end{aligned}$$

The second equality is true because $\mathbf{X}$ and $\hat{\mathbf{X}}$ are independent conditioned on K. It follows that

$$\begin{aligned}
&\inf_{f,g \in \mathcal{H}_R} \mathbb{E}[||\mathbf{X} - \hat{\mathbf{X}}||^2] && (37)\\
&= \inf_{f,g \in \mathcal{H}_R} \mathbb{E}[||\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K]||^2] + \mathbb{E}[||\mathbb{E}[\mathbf{X} \mid K] - \hat{\mathbf{X}}||^2] && (38)\\
&\geq \inf_{f,g \in \mathcal{H}_R} \mathbb{E}[||\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K]||^2] + \inf_{f,g \in \mathcal{H}_R} \mathbb{E}[||\mathbb{E}[\mathbf{X} \mid K] - \hat{\mathbf{X}}||^2] && (39)\\
&\geq \inf_{f,g \in \mathcal{H}_R} \mathbb{E}[||\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K]||^2] + \inf_{f,g \in \mathcal{H}_R} \mathbb{E}[||\mathbb{E}[\hat{\mathbf{X}} \mid K] - \hat{\mathbf{X}}||^2] && (40)\\
&= 2 \inf_{f \in \mathcal{F}_R} \mathbb{E}[||\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K]||^2]. && (41)
\end{aligned}$$

Here, the second inequality is true because the conditional expectation minimizes the squared error. To understand the last equality, note that both terms in Eq. 40 represent the same minimization problem over the same set of joint distributions over $\mathbf{X}$ and K (or $\hat{\mathbf{X}}$ and K). In the first minimization problem we fix $p(\mathbf{x})$ and chose $p(k \mid \mathbf{x})$. In the second minimization problem we equivalently choose both $p(k)$ and $p(\hat{\mathbf{x}} \mid k)$ subject to a constraint on the marginal distribution of $p(\hat{\mathbf{x}})$. □

A result of Blau & Michaeli (2019, Theorem 2) could be adapted to prove the achievability of this bound (we assume a fixed rate here where they constrain the mutual information). While the minimum squared error increases at most by a factor of two by adding a *perfect perceptual quality* constraint (Blau & Michaeli, 2019), our result shows that we need a shared source of randomness if we want the error to be any lower.

APPENDIX C

Here we show that the deterministic encoder described in Section 3.1 is optimal among deterministic encoders. Let $f \in \mathcal{F}_1$ be any stochastic 1-bit encoder. Futher, let $K = f(\mathbf{X}, U)$. Given K, a decoder $g \in \mathcal{G}$ which minimizes the squared error is given by

$$g(k, u) = \mathbf{c}_k, \quad \text{where} \quad \mathbf{c}_k = \mathbb{E}[\mathbf{X} \mid K]. \tag{42}$$

Then the following deterministic encoder $f^* \in \mathcal{F}_1$ performs at least as well,

$$f^*(\mathbf{x}, u) = \underset{k \in \{0,1\}}{\operatorname{argmin}} ||\mathbf{x} - \mathbf{c}_k||^2. \tag{43}$$

In other words, when searching for an encoder which minimizes squared error we only need to consider deterministic encoders of the form of Equation 43.

The codebook vectors $\mathbf{c}_0$ and $\mathbf{c}_1$ create two Voronoi regions in $\mathbb{R}^2$ which split the circle into two sections. The squared error corresponding to a uniform distribution over a section of length t can be calculated to be

$$\mathbb{E}[||\mathbf{X} - \mathbf{c}^*||^2 \mid \theta \in [a, a+t)] = \frac{t^2 + 2\cos(t) - 2}{t^2} \tag{44}$$

where

$$\mathbf{c}^* = \mathbb{E}[\mathbf{X} \mid \theta \in [a, a+t)]. \tag{45}$$

Assume the lengths of the Voronoi sections corresponding to $\mathbf{c}_0$ and $\mathbf{c}_1$ are t and $2\pi - t$, respectively. Then the average error is given by

$$\mathbb{E}[||\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K]||^2] = \frac{t}{2\pi}\frac{t^2 + 2\cos(t) - 2}{t^2} + \frac{2\pi - t}{2\pi}\frac{(2\pi - t)^2 + 2\cos(2\pi - t) - 2}{(2\pi - t)^2} \tag{46}$$

$$= \frac{2 + 2\pi t - t^2 + 2\cos(t)}{2\pi t - t^2}. \tag{47}$$

This function is minimized at $t = \pi$ where

$$\mathbb{E}[||\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K]||^2] = 1 - \frac{4}{\pi^2}. \tag{48}$$

Using Theorem 2 (Appendix B), we conclude

$$\mathbb{E}[d(\Theta, \hat{\Theta})] = \frac{1}{2}\mathbb{E}[||\mathbf{X} - \hat{\mathbf{X}}||^2] \geq \inf_{f \in \mathcal{F}_1} \mathbb{E}[||\mathbf{X} - \mathbb{E}[\mathbf{X} \mid K]||^2] = 1 - \frac{4}{\pi^2}. \tag{49}$$

That is, there is no deterministic encoder which performs better than the one described in Section 3.1.

APPENDIX D

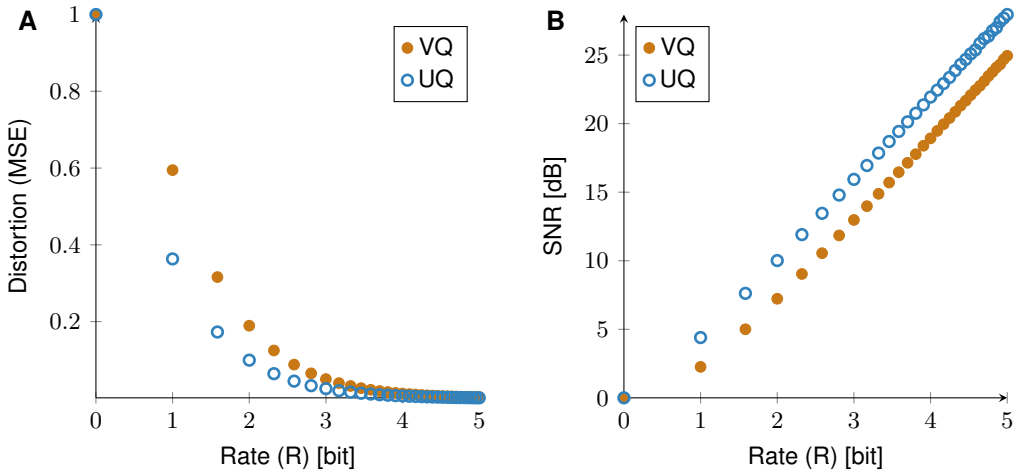

Figure 3: Comparison of a deterministic (VQ) and a stochastic fixed-rate encoder on the circular uniform distribution. **A**: The mean-squared error as a function of the bit rate. **B**: The signal-to-noise ratio as a function of the bit rate.

We showed that the optimal deterministic encoder at target bit rate of 1 is a vector quantizer (VQ) followed by stochastic decoding (Appendix C). Without proof, here we illustrate the performance of this approach and compare it to universal quantization (UQ) at other rates. If $R = \log_2 N$ is the target bit rate, then we have for the distortion:

$$\mathbb{E}_{\text{VQ}}[d(\Theta, \hat{\Theta})] = 1 - \left(\frac{N \sin(\pi/N)}{\pi}\right)^2, \tag{50}$$

$$\mathbb{E}_{\text{UQ}}[d(\Theta, \hat{\Theta})] = 1 - \frac{N \sin(\pi/N)}{\pi}. \tag{51}$$

Figure 3 visualizes both the average distortion D and the signal-to-noise ratio $-10 \log_{10} D$ (SNR). As we increase the bit rate, the distortion converges to zero in both cases. On the other hand, the gap between deterministic and stochastic encoders even increases if measured as a difference in SNR.