Neural Networks Optimally Compress the Sawbridge

Aaron B. Wagner* and Johannes Ballé†

*Cornell University
388 Frank H.T. Rhodes Hall
Ithaca, NY 14853 USA
wagner@cornell.edu

†Google Research
1600 Amphitheatre Pkwy
Mountain View, CA 94043 USA
jballe@google.com

Abstract

Neural-network-based compressors have proven to be remarkably effective at compressing sources, such as images, that are nominally high-dimensional but presumed to be concentrated on a low-dimensional manifold. We consider a continuous-time random process that models an extreme version of such a source, wherein the realizations fall along a one-dimensional "curve" in function space that has infinite-dimensional linear span. We precisely characterize the optimal entropy-distortion tradeoff for this source and show numerically that it is achieved by neural-network-based compressors trained via stochastic gradient descent. In contrast, we show both analytically and experimentally that compressors based on the classical Karhunen-Loève transform are highly suboptimal at high rates.

1 Introduction

Artificial Neural-Network (ANN)-based compressors have recently achieved notable successes on the task of lossy compression of multimedia, spanning an array of sources and in some cases outperforming compressors that have been extensively optimized (see, e.g., [1] and the references therein).

One explanation for the exemplary performance of ANNs is that it derives from their ability to approximate arbitrary functions (e.g., [2]), which in turn enables them to perform nonlinear dimensionality reduction [3]. To see this, first consider classical rate–distortion theory for Gaussian sources, which is based on linear dimensionality reduction [4, Sec. 4.5.2]. Specifically, one projects the source realization onto an orthogonal family of reconstructions obtained from the Karhunen-Loève Transform (KLT) of the source. One then quantizes the resulting coefficients, say with a uniform quantizer followed by entropy coding [5, Sec. 5.5]. At the decoder, the inverse transform is applied to the quantized coefficients. The size of the orthogonal family is generally less than the dimensionality of the source, which provides some amount of compression. The quantization process provides more. For Gaussian sources, this architecture is provably near-optimal at high rates [5, Sec. 5.6.2]. In particular, using a nonlinear transform in place of the KLT provides essentially no benefit.

For real-world multimedia sources, however, there is reason to believe that allowing for nonlinear transforms would be advantageous. The distribution of natural images is widely suspected to be supported by a low-dimensional manifold in pixel-space (e.g., [6]), for instance. That is, while the linear span of the manifold may be high, there

2021 Data Compression Conference (DCC) | 978-1-6654-0333-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/DCC50243.2021.00022

2375-0359/21/$31.00 ©2021 IEEE
DOI 10.1109/DCC50243.2021.00022

exists a continuous, presumably nonlinear, map with a continuous, presumably nonlinear, inverse, from the manifold to a low-dimensional Euclidean space. One could in principle use such a map in place of the linear projections in the classical architecture, with the reduced dimensionality of the output afforded by allowing for nonlinear transforms translating to a lower bit-rate. State-of-the-art ANN-based compressors indeed follow this architecture, with nonlinear analysis and synthesis transforms surrounding a conventional uniform quantizer with entropy coding [1]. As a consequence, one might hypothesize that ANNs are particularly adept at compressing sources for which there is a large discrepancy between the amount of dimensionality reduction afforded by linear versus nonlinear transforms.

We provide evidence for this hypothesis by showing that ANNs optimally compress a prototypical source of this type, and that their performance beats the linear-transform-based approach by a large margin. We consider a particular stochastic process over $[0, 1]$ that can be constructed from a continuous, nonlinear transformation of a single random variable. We learned of this process and its usefulness as a test-case for compression algorithms from the survey by Donoho *et al.* [7], who in turn credit Meyer [8]. Donoho *et al.* refer to this process as the *Ramp*; we shall call it the *sawbridge*. We focus on this process because it exposes the largest possible gap between linear and nonlinear dimensionality reduction. Donoho *et al.* point out that the sawbridge has the same autocorrelation, and thus the same KLT, as the Brownian bridge.[1] In particular, the KLT of the sawbridge is infinite-dimensional. We show that any linear transform requires infinitely many components to recover the source. In contrast, there is a simple nonlinear transform that can recover the source realization from a one-dimensional projection.

We show that this discrepancy in dimensionality reduction translates to a large gap in compression performance. We analytically characterize the performance of optimal one-shot compression for the sawbridge and numerically show that it is realized by a deep ANN trained via stochastic gradient descent. We also characterize the performance of KLT-based schemes and show, both numerically and mathematically, that they are exponentially suboptimal.

The next section introduces the sawbridge process and its properties. Section 3 shows how to optimally compress the sawbridge. Section 4 characterizes the performance of schemes based on the KLT and entropy coding. Section 5 numerically shows that the sawbridge is optimally compressed by existing neural network architectures and training methods, and that linear-based methods are highly suboptimal. All proofs have been omitted due to space constraints but are available in the extended version of the paper [9].

2 The Sawbridge

The focus of this paper is the following stochastic process.

[1] The two processes also share the property that they start and end at zero, which motivates our choice of the former's name.

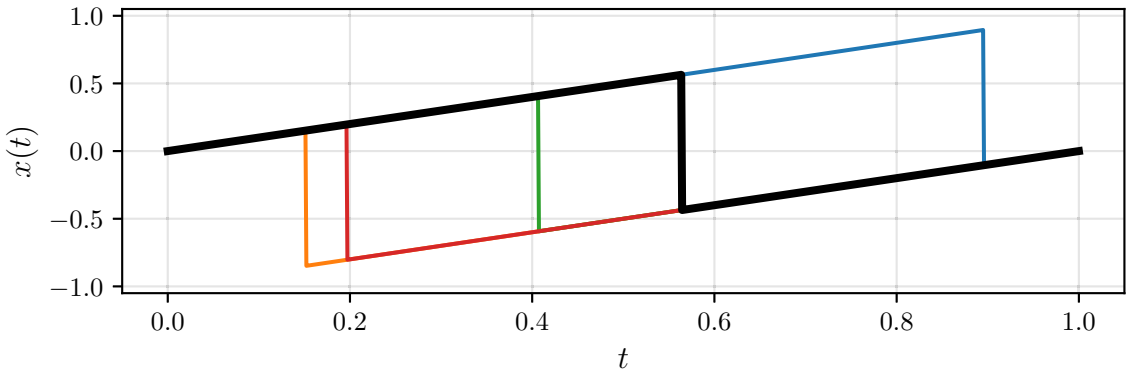

Figure 1: Realizations of the sawbridge. The bold line represents one full realization; others show additional samples.

Definition 1 (cf. [7]). *The* sawbridge *is the process*

$$X(t) = t - \mathbf{1}(t \geq U) \quad t \in [0,1], \tag{1}$$

where U is uniformly distributed over $[0,1]$. We use X or $X(\cdot)$ to refer to the entire process $\{X(t)\}_{t=0}^{1}$.

See Fig. 1 for sample realizations. In words, the process jumps from the "rail" t to the "rail" $t-1$ at the random time U. Alternatively, one can view $-X$ as the centered empirical cumulative distribution function of a single unif$[0,1]$ random variable.

As noted in the introduction, we are interested in the sawbridge because it exposes the largest possible gap between the performance of linear and nonlinear dimensionality reduction. Call a map $f : L^2[0,1] \mapsto \mathbb{R}^k$ a *transform (of dimension k)* if it is continuous and there exists a continuous function $g : \mathbb{R}^k \mapsto L^2[0,1]$ (the *inverse transform*) such that

$$g(f(X)) = X \quad \text{a.s.} \tag{2}$$

We call $\mathbb{R}^k$ the *latent space*.

If f and, especially, g are permitted to be nonlinear, then a transform of dimension one exists, which is clearly the lowest possible. Specifically, the choices

$$f(x) = \int_0^1 x(t)\, dt \tag{3}$$

and

$$(g(y))(t) = t - \mathbf{1}(y \leq t - 1/2) \tag{4}$$

suffice. This comports with the intuition that the process is completely described by U. In fact, the $f(\cdot)$ in (3) satisfies $f(X) = U - 1/2$ a.s. if X and U are related by (1). Note that this f happens to be a linear map.

On the other hand, if f and g are both required to be linear maps then (2) is impossible for any finite k. This follows from the following result on the Karhunen-Loève expansion of the process. Note that the sawbridge is zero mean and let

$$K(s,t) = E[X(s)X(t)] = \min(s,t) - st \tag{5}$$

denote its autocorrelation.

Theorem 1. *The functions*

$$\phi_k(t) = \sqrt{2} \cdot \sin(\pi k t) \quad t \in [0,1], \quad k = 1, 2, \ldots, \tag{6}$$

form an orthonormal basis for $L^2[0,1]$*. They are eigenfunctions of* $K(\cdot,\cdot)$ *with corresponding eigenvalues* $\lambda_k = \frac{1}{\pi^2 k^2}$*, meaning that*

$$\int_0^1 K(s,t)\phi_k(t)\, dt = \lambda_k \cdot \phi_k(s) \tag{7}$$

for all k *and* $s \in [0,1]$*. If we define*

$$Y_k := \int_0^1 X(t)\phi_k(t)\, dt \tag{8}$$

$$= -\sqrt{2\lambda_k}\cos(\pi k U), \tag{9}$$

then $\{Y_k/\sqrt{\lambda_k}\}_{k=1}^{\infty}$ *is a sequence of uncorrelated, zero-mean, unit-variance random variables such that the sawbridge can be written as*

$$X(t) = \sum_{k=1}^{\infty} Y_k \phi_k(t), \tag{10}$$

where the convergence in Eq. (10) *is in quadratic mean, uniformly over* t*.*

This implies that if the transform and inverse must be linear, then an infinite number of dimensions is required to represent the sawbridge, as shown in the following corollary.

Corollary 1. *For any finite* k*, if* $f : L^2[0,1] \mapsto \mathbb{R}^k$ *and* $g : \mathbb{R}^k \mapsto L^2[0,1]$ *are linear maps then they cannot satisfy (2).*

The large gap between linear and nonlinear dimensionality reduction for the sawbridge has consequences for compression, to which we turn next.

3 Optimal Compression

We consider a one-shot form of compression in which the goal is to minimize the entropy of the compressed representation for a given mean squared error.

Definition 2. *An* encoder *is a map* $f : L^2[0,1] \mapsto \mathbb{N}$*. Its* entropy *and* distortion *are*

$$H(f) = -\sum_{i \in \mathbb{N}} \Pr(f(X) = i) \log \Pr(f(X) = i)$$

$$D(f) = E\left[\int_0^1 (X(t) - E[X(t)|f(X)])^2\, dt\right],$$

respectively.[2]

[2]Throughout, all logarithms are base two.

Note that Definition 2 assumes that the reproduction $E[X(t)|f(X)]$ need not be a valid sawbridge realization. Also note that an encoder is distinct from a transform in that it maps the source realization to a discrete set and is therefore not invertible. In practice, compression involves mapping the source realization to a variable-length bit string, whose expected length one might wish to minimize. The minimum such expected length, $L^*(f)$, is known to satisfy $H(f) \leq L^*(f) < H(f) + 1$ if one requires that the codewords are prefix-free (e.g., [10, Theorem 5.4.1]) and

$$L^*(f) \leq H(f) \tag{11}$$
$$H(f) \leq L^*(f) + (1 + L^*(f)) \log(1 + L^*(f)) - L^*(f) \log L^*(f), \tag{12}$$

if one does not [11, Theorem 1]. As such, it is reasonable to focus on $H(f)$ as the figure-of-merit, especially at high rates.

Definition 3. *The* entropy-distortion function *of the sawbridge is*

$$H(\Delta) = \inf_f H(f), \tag{13}$$

where the infimum is over all encoders f such that $D(f) \leq \Delta$.

Theorem 2. *If $\Delta \geq 1/6$, then $H(\Delta) = 0$. For any $0 < \Delta < 1/6$, we have*

$$H(\Delta) = -\left\lfloor \frac{1}{p} \right\rfloor \cdot p \log p - q \log q, \tag{14}$$

where $q = \left(1 - \left\lfloor \frac{1}{p} \right\rfloor \cdot p\right)$ and p is the unique number in $(0, 1)$ such that

$$\left\lfloor \frac{1}{p} \right\rfloor \cdot p^2 + q^2 = 6\Delta. \tag{15}$$

A plot of $H(\Delta)$ is included in Fig. 2. Note that, unlike the rate-distortion function, the entropy-distortion function is not guaranteed to be convex and indeed it is not in this case. It reaches its lower convex envelope, which we denote by $\mathrm{LCE}(H(\cdot))$, at points of the form $(\log M, 1/(6M))$ [12, Corr. 6]. These points are achieved by encoders that quantize U to one of several equal-sized intervals. In particular, for any $\lambda > 0$, minimizers of the Lagrangian

$$\min_f \; H(f) + \lambda \cdot D(f) \tag{16}$$

are encoders of this type. Likewise, since $\mathrm{LCE}(H(\cdot))$ describes the entropy-distortion tradeoff of randomized encoders, the best randomized encoders are those that randomly select from among deterministic encoders that uniformly quantize U.

Theorem 2 also implies a simple high-rate characterization.

Corollary 2. *For the sawbridge,*

$$\lim_{\Delta \to 0} \left| H(\Delta) - \log \frac{1}{6\Delta} \right| = 0. \tag{17}$$

Note that this corollary implies that $H(\Delta)/\log \frac{1}{6\Delta} \to 1$ but is significantly stronger. Next we shall see that a compressor that follows the classical approach based on the KLT is far from meeting this optimal performance at high rates.

4 KLT-Based Compression

The classical approach to compressing a source such as the sawbridge is to uniformly quantize and separately entropy-code a subset of the KLT coefficients. Specifically, given a target distortion Δ, define the constants $D = \frac{\Delta^2 \pi^2}{4}$, $K = \left\lceil \frac{1}{\pi\sqrt{D}} \right\rceil$, and $\delta = \frac{\sqrt{12\gamma}}{K}$, where $\gamma > 0$ is the unique solution to the fixed-point equation $\tan^{-1}(\pi\sqrt{\gamma}) = \frac{1}{\pi\sqrt{\gamma}}$. Note that the dependence on Δ is suppressed in all three constants. Consider the stochastic encoder f_Δ that quantizes the first K coefficients of the KLT to resolution δ using random dither. That is,

$$f_\Delta(X) = \left(\left\lfloor \frac{Y_\ell}{\delta} + U_\ell \right\rceil, \ell \in \{1, \ldots, K\} \right), \tag{18}$$

where $U_1, \ldots, U_K$ are i.i.d. unif$[-1/2, 1/2]$ and $\lfloor \cdot \rceil$ denotes rounding to the nearest integer. The $U_1, \ldots, U_K$ represent side randomness that is independent of the source and available when decoding. We first show that f_Δ achieves distortion Δ.

Lemma 1. *The encoder f_Δ satisfies $D(f_\Delta) \leq \Delta$ for all $0 < \Delta < 1/6$.*

Since $f_\Delta(\cdot)$ is a stochastic encoder that achieves distortion Δ, it follows that

$$\mathrm{LCE}(H(\cdot))(\Delta) \leq H(f_\Delta | \{U_\ell\}) \leq \sum_{\ell=1}^{K} H\left(\left\lfloor \frac{Y_\ell}{\delta} + U_\ell \right\rceil \middle| U_\ell \right) =: \overline{H}(\Delta). \tag{19}$$

Note that $\overline{H}(\Delta)$ is the rate that is achievable if the quantized components are compressed separately, as is typically done in practice. When compressing a stationary Gaussian process over a long horizon the analogue of the $\overline{H}(\Delta)$ bound is provably near-optimal at high rates (combining [5, Sec. 5.6.2] and [4, Sec 4.5.3]). Comparing the following result with Corollary 2 shows that for the sawbridge, this bound is poor in the high-rate regime.

Theorem 3. *Let $s(\cdot)$ denote the arcsine density over $[-\sqrt{2}, \sqrt{2}]$, i.e.,*

$$s(x) = \frac{1}{\pi\sqrt{(\sqrt{2} - x)(\sqrt{2} + x)}}, \tag{20}$$

and let $u_x(\cdot)$ denote the uniform density over $[-x/2, x/2]$. Then

$$\overline{H}(\Delta) = \sum_{\ell=1}^{K} \left[h(s(\cdot) \star u_{\sqrt{12\gamma}\pi\ell/K}(\cdot)) - \log \frac{\pi\ell\sqrt{12\gamma}}{K} \right], \tag{21}$$

and as a result,

$$\lim_{\Delta\to 0} \Delta \cdot \overline{H}(\Delta) = \frac{2}{\pi^2} \cdot \left(\int_0^1 h(s(\cdot) \star u_{\pi x\sqrt{12\gamma}}(\cdot))\, dx - \log(\pi\sqrt{12\gamma}/e) \right), \tag{22}$$

where $h(\cdot)$ denotes differential entropy and $\star$ denotes convolution.

5 Neural-Network-Based Compression

We compare the performance of experimentally-trained ANNs against the optimal entropy-distortion tradeoff in Theorem 2 and various linear-transform-based schemes. We follow the approach summarized by Ballé *et al.* [1]. To represent the sawbridge digitally, we sample t at 1024 equidistant points between 0 and 1; thus, each realization is represented as a 1024-dimensional vector. Due to this discretization, only a finite number of realizations are possible, and we took care to keep this number high enough such that none of the transform codes were able to exploit it. We optimize over three sets of model parameters: the weights and biases of an analysis (f) and synthesis (g) transform, both of which are represented by ANNs, as well as the parameters of a non-parametric entropy model p. Note that we do not assume that the analysis and synthesis transforms form exact inverses of each other; not only would this be difficult to enforce with ANNs, it is also not necessarily optimal. We employ ANNs of three layers, with 100 units each (except for the last), and leaky ReLU as an activation function (except for the last), for each of the transforms. We perform uniform scalar quantization of the transform coefficients, and use an entropy model that assumes independence between each of the latent dimensions; i.e., each transform coefficient is assumed to be encoded separately. The objective is to minimize the Lagrangian in Eq. (16), as in Ballé *et al.* [1].

To make the comparison with linear transform coding schemes fair, we optimize linear transforms using the same methodology (linear transforms are special cases of ANNs, with just one layer). For a comparison with specified orthonormal transforms, we express the analysis transform as the composition of a fixed orthonormal matrix with a trainable diagonal scaling matrix (and analogously for the synthesis transform, in reversed order). The scaling enables uniform quantization with different effective step sizes in each latent dimension.

Empirical results are plotted in Fig. 2. Each point in the plots represents the outcome of one individual optimization of the Lagrangian in Eq. (16) with a particular predetermined value of λ, and with a predetermined constraint on the transforms (as described above). We spaced λ logarithmically in order to cover a wide range of possible trade-offs. Both H and D are computed as empirical averages over 10^7 source realizations, where H represents the cross-entropy between the fitted entropy

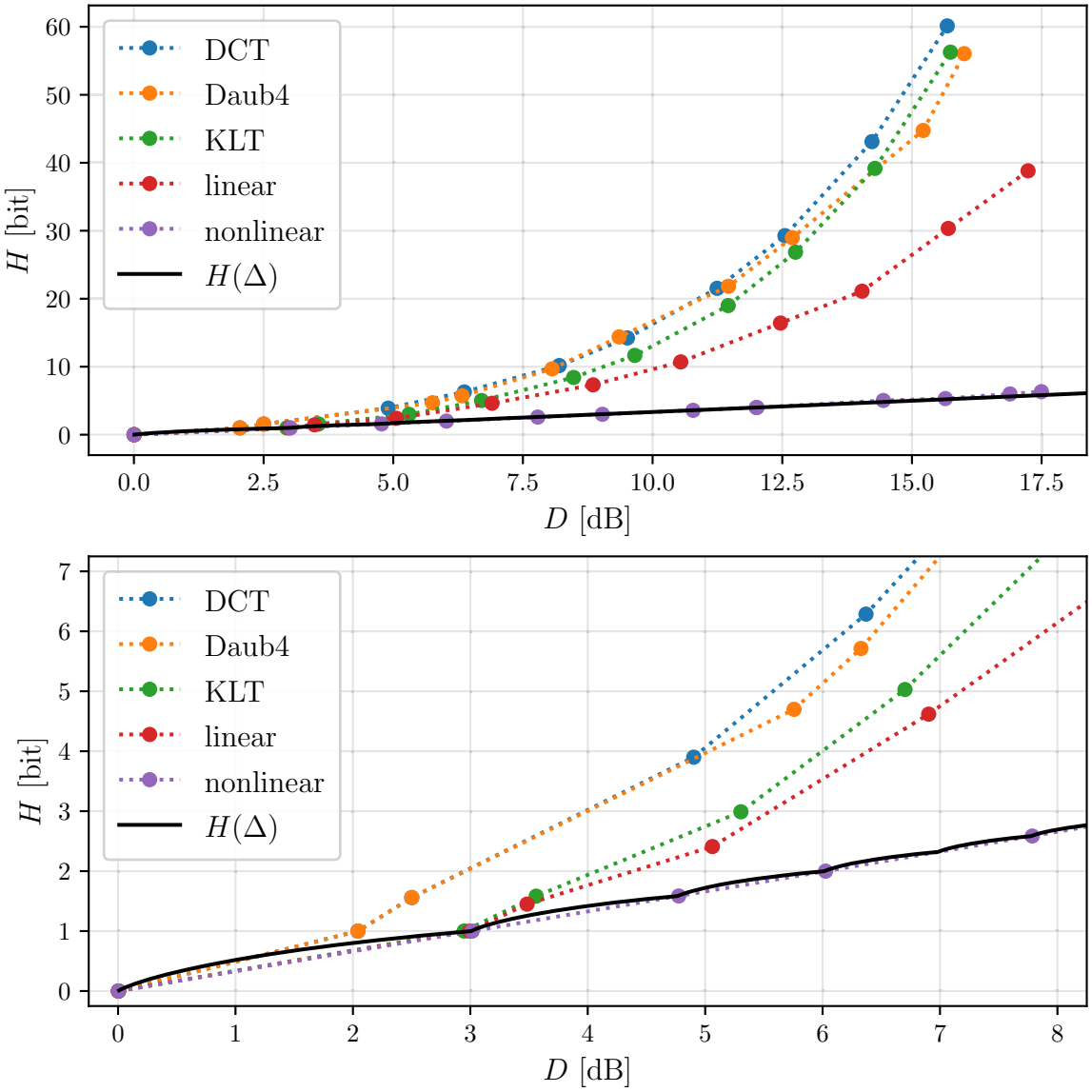

Figure 2: Empirical entropy–distortion plots for transform codes constrained to discrete cosine transform (DCT), Daubechies 4-tap wavelet (Daub4), Karhunen–Loève transform (KLT), arbitrary linear transforms, and nonlinear transforms implemented by ANNs. We also plot the entropy–distortion function of the source. The bottom panel shows the same data, zoomed in to the low-rate regime.

model p and the empirical distribution of the coefficients, and D is mean squared error across t. The top panel in the figure illustrates that the growth rate of the entropy of the linear transform codes is highly suboptimal, especially when the linear transform is further constrained to coincide with a fixed orthogonal transform such as the DCT, KLT, or a wavelet transform. In the bottom panel, we observe that the nonlinear transform code is essentially optimal. Since the transform codes are optimized for the Lagrangian, they settle into the "kinks" of the entropy–distortion function, as discussed in Section 3 (note that not all of the kinks are occupied; this is

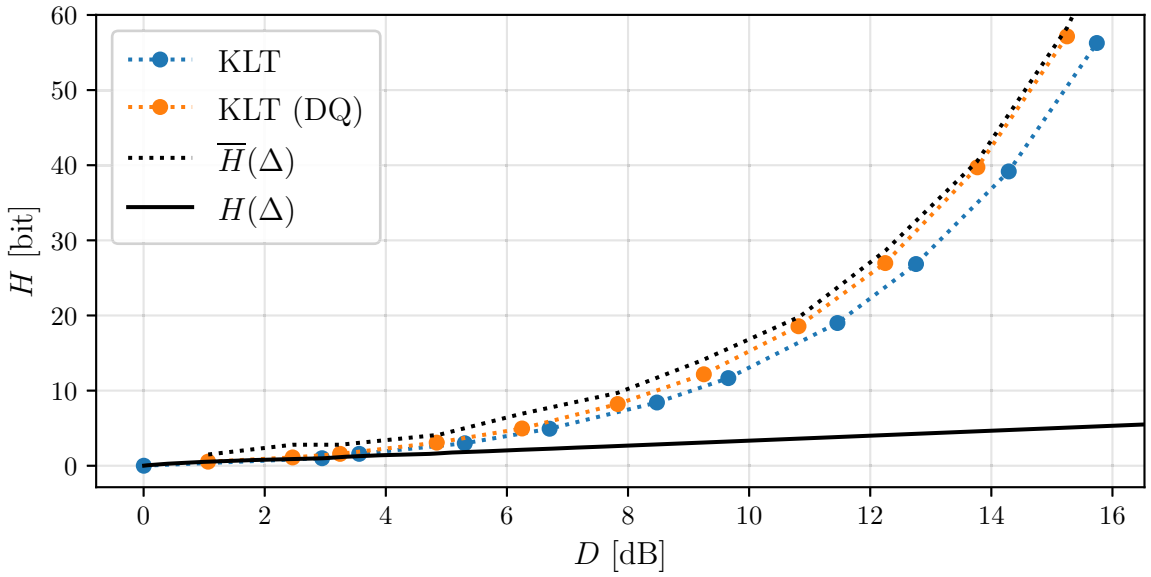

Figure 3: Empirical entropy–distortion plot for transform codes constrained to Karhunen–Loève transform (KLT) with/without dithered quantization (DQ). We also plot the entropy–distortion function of the source, and the bound developed in (22).

due to the pre-determined spacing of λ). As predicted by the theory, we find that a transform code with a linear analysis transform and a nonlinear synthesis transform, optimized with the same method, performs the same as the nonlinear code plotted in the figure (not shown).

Inspection of the latent-space activations as well as the entropy model reveal that linear transforms require successively larger numbers of latent dimensions as the rate increases. Nonlinear codes, on the other hand, use only a small number of latent dimensions (additional dimensions allowed for by the experimental setup are collapsed to zero-entropy distributions by the optimization procedure), implying that they successfully discover the low-dimensional structure of the source. In fact, we find that when constraining the latent space to a single dimension, the nonlinear codes do just as well. A simple explanation for why they may choose to use more dimensions in some cases is that an entropy model consisting of a single uniform distribution over a discrete set of states can be factorized into a product of multiple uniform distributions, as long as the number of states is not prime.

To illustrate the optimal KLT-constrained code from Section 4, and to verify that our optimization methodology is valid for linear codes, we plot $\overline{H}(\Delta)$ from Eq. (21) and the optimal tradeoff, $H(\Delta)$, along with empirical results for a KLT-constrained code with and without dithered quantization, in Fig. 3. Note that $\overline{H}(\Delta)$ and the curve for dithered quantization are quite close, especially at high rates.

Acknowledgment

The first author wishes to thank Jingsong Lin for his contributions to the early phases of this work. He was supported by the US National Science Foundation under

grants CCF-1617673 and CCF-2008266 and the US Army Research Office under grant W911NF-18-1-0426.

References

[1] 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 Proc.*, 2020, to appear. DOI: 10.1109/JSTSP.2020.3034501.

[2] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," *Neural Networks*, vol. 6, no. 6, 1993. DOI: 10.1016/S0893-6080(05)80131-5.

[3] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, 2006. DOI: 10.1126/science.1127647.

[4] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice Hall, 1971.

[5] W. A. Pearlman and A. Said, *Digital Signal Compression: Principles and Practice*. Cambridge University Press, 2011.

[6] O. Hénaff, J. Ballé, N. C. Rabinowitz, and E. P. Simoncelli, "The local low-dimensionality of natural images," in *3rd Int. Conf. on Learning Representations (ICLR)*, 2015. arXiv: 1412.6626.

[7] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, 1998. DOI: 10.1109/18.720544.

[8] Y. Meyer, "Wavelets and applications," in *Lecture at CIRM Luminy meeting*, Mar. 1992.

[9] A. B. Wagner and J. Ballé, "Neural networks optimally compress the sawbridge," arXiv:2011.05065.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd. Hoboken: John Wiley & Sons, 2006.

[11] W. Szpankowski and S. Verdú, "Minimum expected length of fixed-to-variable lossless compression without prefix constraints," *IEEE Trans. on Inf. Theory*, vol. 57, no. 7, 2011. DOI: 10.1109/TIT.2011.2145590.

[12] A. György and T. Linder, "Optimal entropy-constrained scalar quantization of a uniform source," *IEEE Trans. on Inf. Theory*, vol. 46, no. 7, pp. 2704–2711, 2000. DOI: 10.1109/18.887885.