Do Neural Networks Compress Manifolds Optimally?

Sourbh Bhadane and Aaron B. Wagner
School of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853 USA
snb62@cornell.edu, wagner@cornell.edu

Johannes Ballé
Google Research
Mountain View, CA 94043 USA
jballe@google.com

Abstract—Artificial Neural-Network-based (ANN-based) lossy compressors have recently obtained striking results on several sources. Their success may be ascribed to an ability to identify the structure of low-dimensional manifolds in high-dimensional ambient spaces. Indeed, prior work has shown that ANN-based compressors can achieve the optimal entropy-distortion curve for some such sources. In contrast, we determine the optimal entropy-distortion tradeoffs for two low-dimensional manifolds with circular structure and show that state-of-the-art ANN-based compressors fail to optimally compress them.

Index Terms—rate-distortion, neural networks, manifolds

I. INTRODUCTION

Stochastically-trained Artificial Neural-Network-based (ANN-based) lossy compressors are the state-of-the-art for various types of sources, most notably images [1]–[3]. One particularly successful framework for using ANNs for lossy compression, which is the focus of this paper, creates the compressed representation by quantizing the latent variables in an autoencoder-like architecture [4]. One can view this as similar to the transform-based coding approach that underpins JPEG and other standards, except that the transforms are implemented by multilayer perceptrons and can therefore be nonlinear. These ANN-based compressors outperform linear-transform-based methods, even under a mean-squared-error (MSE) distortion measure [1]. Since the latter is provably near-optimal for stationary Gaussian sources under an MSE distortion constraint, ANN-based compressors are evidently able to exploit non-Gaussianity in sources.

Given that ensembles of images have long been suspected to live on low-dimensional manifolds in pixel-space, it is natural to conjecture that ANN-based compressors are adept at compressing sources that exhibit low-dimensional manifold structure in a high-dimensional ambient space. Previous work [5] considered a particular random process (the “sawbridge”) over \([0, 1]\) exhibiting this structure and found that stochastically-trained ANNs indeed compress the process optimally.

We consider the random process obtained by applying a random cyclic shift to the function \(t \mapsto t - 1/2\) over \([0, 1]\). We call the resulting process the ramp. We characterize the entropy-distortion function for this process under an MSE distortion constraint. Despite the considerable similarities between the ramp and the sawbridge, we find that stochastically-trained ANN-based compressors fail to compress the ramp optimally at high-SNR. The difficulty stems from the fact that, unlike the sawbridge, the set of ramp realizations forms a closed loop in function space, which creates a topological challenge related to the impossibility of mapping a circle to a segment in a continuous and invertible way [6, Chap 2. Ex. 7]. To illustrate this issue in arguably its simplest form, we begin by considering the problem of compressing the unit circle in two-dimensions. We characterize the entropy-distortion function and again find that stochastically-trained ANNs are suboptimal at high rates.

We omit proofs of the theorems and supply proof sketches instead. Full proofs can be found in the extended version [7].

II. PRELIMINARIES

For a source \(X\) in a space \(\mathcal{M}\), we define an encoder and its entropy and distortion. In this paper \(\mathcal{M}\) will be either \(L^2[0, 1]\) or \(\mathbb{R}^2\). In both cases, conditional expectations and norms are well-defined.
**Definition 1.** An encoder is a mapping \( f : \mathcal{M} \mapsto \mathbb{N} \). Its entropy and distortion are given by
\[
H(f) = \sum_{i \in \mathbb{N}} - \Pr(f(X) = i) \log(\Pr(f(X) = i))
\]
\[
D(f) = \mathbb{E}\left[\|X - \mathbb{E}[X \mid f(X)]\|^2\right],
\]
respectively.

Note in particular that we consider mean squared error as the distortion measure. We shall characterize optimal compression performance via the entropy-distortion function.

**Definition 2.** The entropy-distortion function of \( X \) is
\[
E(D) = \inf_f H(f)
\]
\[
s.t. \ D(f) \leq D.
\]

We consider the entropy-distortion function instead of the more-conventional rate-distortion function [8, Theorem 10.2.1] because ANN-based compressors optimize entropy, which is known to be a lower bound to the expected codeword length under optimal, one-shot prefix-free encoding [8, Theorem 5.4.1]

### III. THE CIRCLE

The **circle** is a 2-D source with a 1-D latent variable. \( \theta \sim \text{Unif}[0, 2\pi], Z = (\cos \theta, \sin \theta) \).

We first derive its optimal entropy-distortion tradeoff and then analyze the performance of ANN-based compressors.

**A. The Optimal Tradeoff**

**Theorem 3.** For the circle, if \( D \geq 1 \), \( E_c(D) = 0 \). If \( 0 < D < 1 \), then
\[
E_c(D) = \inf_{(\theta_i)_{i=1}^\infty} - \sum_{i=1}^\infty \frac{\theta_i}{2\pi} \log\left(\frac{\theta_i}{2\pi}\right)
\]
\[
s.t. \ \sum_{i=1}^\infty \frac{\theta_i}{2\pi} \left(1 - \text{sinc}^2\left(\frac{\theta_i}{2}\right)\right) \leq D,
\]
\[
\sum_{i=1}^\infty \theta_i = 2\pi, \theta_i \geq 0 \text{ for all } i,
\]
where
\[
\text{sinc}(x) = \begin{cases} 
sin(x) & \text{if } x \neq 0 \\
1 & \text{otherwise}
\end{cases}
\]

**Proof Sketch.** For an encoder \( f_{\text{circ}} \) and for \( i \in \mathbb{N} \), we consider quantization cells on the unit circle, \( C_i \). The entropy can be written in terms of the Lebesgue measure of the \( C_i \), \( \mu(C_i) \). The main part of the proof involves bounding the distortion as
\[
D(f_{\text{circ}}) \geq \sum_{i} \frac{\mu(C_i)}{2\pi} \left(1 - \text{sinc}^2\left(\frac{\mu(C_i)}{2}\right)\right)
\]

**Theorem 4.**
\[
E_c(D) \geq \sup_{\lambda \geq 0} \inf_{0 < \theta < 2\pi} - \log\left(\frac{\theta}{2\pi}\right) + \lambda \left(1 - \text{sinc}^2\left(\frac{\theta}{2}\right)\right) - \lambda D.
\]

The lower bound in Theorem 4 is illustrated in Fig. 1 ("lower bound"). An upper bound can be obtained by partitioning the circle into arcs with a biuniform size distribution ("achievable for biuniform" intervals in Fig. 1). Note that these bounds essentially coincide at high-SNR.

**B. ANN Performance**

To compress a source of dimension \( d_s \) using a latent dimension of \( d_c \), a stochastically-trained ANN-based compressor consists of an analysis transform \( g_a : \mathbb{R}^{d_s} \mapsto \mathbb{R}^{d_c} \), a synthesis transform \( g_s : \mathbb{R}^{d_c} \mapsto \mathbb{R}^{d_s} \), a quantizer and an entropy model that is factorized across each dimension of the codeword. The analysis and synthesis transforms are fully-connected feedforward neural networks with 2 hidden layers containing 100 neurons each. The quantization operation is not differentiable, and therefore during training we replace it with a differentiable proxy that varies over the course of the training process, beginning with dithered quantization and ending...
with the hard quantizer that is used at test time [9]. During training, a $d_s$-dimensional vector is fed into the analysis transform to obtain a $d_c$-dimensional latent vector. The quantization-proxy is then applied to the latent vector, which is then fed to the synthesis transform to obtain the $d_r$-dimensional reconstruction. The entropy model is a feedforward neural network that computes the entropy, $E$ of the quantized latents. Distortion $D$ is the mean-squared error between the reconstruction and the input vector. The Lagrangian $E + \lambda D$ is stochastically minimized over the trainable parameters of the ANN-based compressor using Adam [10]. We sweep across different values of $\lambda$ to obtain points on the lower convex hull of the ANN-compressor’s entropy-distortion tradeoff. We remark that the neural-network architecture and entropy model are similar to the ANN-based compressor trained for the sawbridge process in [5].

Since the circle is described by the scalar random variable $\theta$, we take the latent dimension $d_c = 1$. The resulting performance is shown in Fig. 1. We see that the performance is suboptimal at high-SNR.

To see why, note that intuition suggests, and the proofs of Theorems 3 and 4 essentially confirm, that an optimal scheme for compressing the circle is to quantize the angle $\theta$ into contiguous cells, all but (at most) one of which are the same size. Thus we would like the analysis transform to extract the angle $\theta$ from the realization of the circle, i.e., to implement $\text{atan2}(\mathbf{Z})$ or some scaled and shifted version thereof. The issue is that the function $\mathbf{Z} \mapsto \theta$ is not continuous over the circle, while the analysis transform must be continuous (and indeed differentiable) by construction.

This is confirmed in Fig. 2a, which shows the quantized output of the trained analysis transform $g_a$ as a function of the angle $\theta$ at high SNR. We see that the analysis transform attempts to implement a discontinuity around 4.5 radians, but the function is insufficiently steep, so it passes through various intermediate quantization levels on its way from its minimum value to its maximum value. This creates an identifiability problem at the decoder, in that certain quantizer outputs can be caused by two values of $\theta$, one in the decreasing part of the function and one in the increasing part. Of course, the location of the discontinuity (4.5 radians in this case) is arbitrary and will be different if the network is retrained.

At low SNR, the distortion accruing from this lack of invertibility is negligible compared to the distortion arising from the quantization process. At high rates, it dominates, and the performance is off the optimal entropy-distortion curve. The extent of the suboptimality is determined by how steep the analysis transform can be made, which in turn is controlled by the training process. In order for the training process to make the “steep” portion of the analysis transform steeper, we require source realizations from the range over which the function is steep to be present in the batch. Smaller batch sizes are less likely to include such points, and thus reducing the batch size leads to a larger gap from optimality (Fig. 1). Fixing a training data set (instead of drawing fresh samples at each iteration) has a similar effect, because once the steep portion of the function falls entirely between two training points, the training process has no incentive to make it steeper (Fig. 1).

This issue is not unique to the circle in the Euclidean plane. Indeed, it can arise in more complex sources whose support has a circular structure, such as the ramp process to which we turn next.

IV. THE RAMP

Consider the following process.

$$J_t \overset{\text{def}}{=} ([t + V] \mod 1) - \frac{1}{2},$$

where $V \sim \text{Unif}[0,1]$. We call this the ramp process and $V$ the phase. We are interested in this process as a model of low-dimensional structure in high-dimensional spaces; on the one hand, the set of source realizations has infinite linear span; on the other hand, the realization is completely determined by the scalar random variable $V$. Note that $V = 1$ and $V = 0$ yield identical realizations of the ramp. Thus the set of realizations forms a circle in function space in some sense. This process is similar in some important respects to the “sawbridge” process considered in an earlier work [5]. But whereas the sawbridge is optimally compressed by the ANN-based architecture under study, we shall see that the ramp is not.

A. The Optimal Tradeoff

**Theorem 5.** For the ramp, if $D \geq \frac{1}{12}$, $E_r(D) = 0$. If $0 < D < \frac{1}{12}$, then

$$E_r(D) = \inf_{\{p_i\}_{i=1}^\infty} - \sum_{i=1}^\infty p_i \log p_i$$

subject to

$$\sum_{i=1}^\infty p_i^2 (2 - p_i) \leq \frac{D}{12},$$

and

$$\sum_{i=1}^\infty p_i = 1, p_i \geq 0 \text{ for all } i.$$

**Proof Sketch.** Each value of the phase variable defines a unique ramp realization which is mapped to a codeword by a given ramp encoder $f_{ramp}$. For $i \in \mathbb{N}$, let $S_i \subset [0,1]$
be the quantization cell of \( i \). The entropy can be written in terms of the Lebesgue measure of \( S_i \)'s. The main part of the proof involves bounding the distortion as

\[
D(f_{\text{ramp}}) \geq \sum_i \mu(S_i) \left( \frac{1}{12} - \frac{1}{3} \left( \frac{1 - \mu(S_i)}{2} \right)^2 \right)
\]

for an arbitrary measurable \( S_i \) with equality holding if each of the \( S_i \)'s are intervals.

The entropy-distortion function in (5) is formally an infinite-dimensional optimization problem. The following lower bound is easily computed.

**Theorem 6.**

\[
E_r(D) \geq \sup_{\lambda \geq 0} \inf_{0 < p_i < 1} -\log(p_i) + \lambda p_i (2 - p_i) \frac{1}{12}.
\]

This function is not continuous, however, and being a multilayer perceptron, \( g_s(\cdot) \) must be continuous (and indeed differentiable). In practice, the trained synthesis transform will implement a continuous approximation to (10) that incurs distortion because its discontinuity is insufficiently steep. The situation is thus analogous to the circle. Note that for the circle, however, the problem arose with the analysis transform, whereas here it arises in the synthesis transform. Indeed, evaluating the ramp at one fixed time \( t \) already has the desired form for the analysis transform,

\[
J_t = \left( \begin{array}{c}
\left( t + V \right) \mod 1 \\
1/2
\end{array} \right).
\]

Thus the analysis transform can simply transmit any of the input samples directly, which is clearly a continuous map. On the other hand, Fig. 4b shows the reconstruction, evaluated at a fixed time index, as a function of \( V \). We see that the discontinuity is insufficiently steep. As with the circle, at low rates, the approximation
Fig. 4: (a) Quantized encoder output vs. phase for $\lambda = 4096$ (away from optimal tradeoff). (b) Reconstruction at randomly chosen index vs. phase for $\lambda = 4096$. Here the synthesis transform is insufficiently steep. The error noted above is negligible compared with the quantization error, and thus this phenomenon does not lead to entropy-distortion suboptimality.

V. OVERPARAMETERIZING THE LATENT SPACE

One possible workaround to the suboptimality identified above is to overparametrize the latent space. In the above examples, the analysis transform could output a 2-D vector ($d = 2$) rather than a scalar, for instance. In the case of the circle, this evidently solves the identifiability problem, as the analysis transform can now simply be the identity map. This makes the quantization problem more difficult, however, because the latent variables are always quantized independently of each other, but if the analysis transform is indeed the identity map then the two latent variables are not statistically independent (though they are uncorrelated). With an identity analysis transform, it is possible to independently quantize the two latent variables so that the overall system is optimal if one uses quantizers that are highly asymmetrical between the two components. Specifically, one of the components is quantized to a single bit to indicate which hemisphere the source is in, and the other is quantized with high resolution. The resulting system will be optimal so long as the quantized values are independent and uniformly distributed over their respective supports. Remarkably, with enough training, the network is able to find this solution, as shown in Figs. 1 and 2b. Of course, this only works when the number of reconstructions is even.

For the ramp, on the other hand, overparametrizing the latent space appears to provide little benefit (Fig. 3). From the discussion in Section IV, this is unsurprising.

VI. DISCUSSION

We conclude that the argument that ANN-based compressors are adept at compressing low-dimensional manifolds should be applied with some care. While ANNs are essentially optimal compressors for some manifolds [5], they are suboptimal for the sources exhibiting a circular topology considered in this paper.

Low-level models of natural images have long included spatially-local image features that exist on one or more intrinsic dimensions and can be organized on circular topologies, such as spatial orientation [11]–[14]. It has also been noted that different edge or feature profiles can be determined by their phase, another dimension which has a circular topology, and their phase congruency, in complex wavelet representations [15], [16]. Empirically, it has been shown that local image statistics are inherently low-dimensional [17] and that these local statistics live on a topology equivalent to a Klein bottle [18], a topology which is also found in orientation-selective complex wavelet filters. Thus while the sources considered in this paper are obviously synthetic, the findings in this paper may have some applicability to image compression.

Beyond the findings about ANN-based compressors, we have also determined the entropy-distortion functions of the circle and ramp. Classical rate-distortion theory has favored the study of Gaussian sources over those with low-dimensional manifold structure, despite the relevance of the latter to the study of image compression. Prior work on compression of low-dimensional sets includes rate-distortion tradeoffs for fractal sources [19] and rate-distortion analysis of compressed sensing [20], [21]. More connected to the results of this paper, [22] provides a lower bound and [23] an upper bound on the rate-distortion function of the circle. We have provided a complete characterization of the entropy-distortion function.
REFERENCES

[1] J. Ballé, V. Laparra, and E. P. Simoncelli, “End-to-end optimized image compression,” in International Conference on Learning Representations, ICLR 2017, 2017.

[2] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” in International Conference on Learning Representations, ICLR 2018, 2018.

[3] F. Mentzer, G. Toderici, M. Tschannen, and E. Agustsson, “High fidelity generative image compression,” in Advances in Neural Information Processing Systems 34, 2020.

[4] J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici, “Nonlinear transform coding,” IEEE Journal of Selected Topics in Signal Processing, vol. 15, no. 2, pp. 339–353, 2020.

[5] A. B. Wagner and J. Ballé, “Neural networks optimally compress the sawbridge,” arXiv:2011.05065.

[6] J. R. Munkres, Topology. Prentice Hall Upper Saddle River, 2000, vol. 2.

[7] S. Bhadane, A. B. Wagner, and J. Ballé, “Do neural networks compress manifolds optimally?”

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

[9] E. Agustsson and L. Theis, “Universally quantized neural compression,” in Advances in Neural Information Processing Systems 33, 2020, pp. 12 367–12 376.

[10] D. Kingma, J. Ba, Y. Bengio, and Y. LeCun, “Adam: A method for stochastic optimization,” in 3rd International Conference on Learning Representations, 2015.

[11] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, “Shiftable multiscale transforms,” IEEE Trans. on Information Theory, vol. 38, no. 2, 1992. DOI: 10.1109/11.119725.

[12] B. Olshausen and D. J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” Nature, vol. 381, Jun. 1996. DOI: 10.1038/381607a0.

[13] A. J. Bell and T. J. Sejnowski, “The Independent Components’ of natural scenes are edge filters,” Vision Research, vol. 37, no. 23, 1997. DOI: 10.1016/S0040-0022(97)00121-1.

[14] J. H. van Hateren and A. van der Schaaf, “Independent component filters of natural images compared with simple cells in primary visual cortex,” Proc. of the Royal Society of London B: Biological Sciences, vol. 265, no. 1394, 1998. DOI: 10.1098/rspb.1998.0303.

[15] S. Venkatesh and R. A. Owens, “On the classification of image features,” Pattern Recognition Letters, vol. 11, no. 5, pp. 339–349, May 1990. DOI: 10.1016/0167-8655(90)90043-2.

[16] P. Kovesi, “Image features from phase congruency,” Videre: Journal of Computer Vision Research, vol. 1, no. 3, pp. 2–26, 1999. ISSN: 1089-2788.

[17] 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.

[18] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian, “On the local behavior of spaces of natural images,” Int. Journal of Computer Vision, vol. 76, 2008. DOI: 10.1007/s11263-007-0056-x.

[19] T. Kawabata and A. Dembo, “The rate-distortion dimension of sets and measures,” IEEE Transactions on Information Theory, vol. 40, no. 5, pp. 1564–1572, 1994.

[20] Y. Wu and S. Verdú, “Rényi information dimension: Fundamental limits of almost lossless analog compression,” IEEE Transactions on Information Theory, vol. 56, no. 8, pp. 3721–3748, 2010. DOI: 10.1109/TIT.2010.2050803.

[21] M. Leinonen, M. Codreanu, M. Jungti, and G. Kramer, “Rate-distortion performance of lossy compressed sensing of sparse sources,” IEEE Transactions on Communications, vol. 66, no. 10, pp. 4498–4512, 2018. DOI: 10.1109/TCOMM.2018.2834349.

[22] E. Riegler, H. Bölcskei, and G. Koliander, “Rate-distortion theory for general sets and measures,” in 2018 IEEE International Symposium on Information Theory (ISIT), IEEE, 2018, pp. 101–105.

[23] G. Koliander, G. Pichler, E. Riegler, and F. Hlawatsch, “Entropy and source coding for integer-dimensional singular random variables,” IEEE Transactions on Information Theory, vol. 62, no. 11, pp. 6124–6154, 2016.