Breaking Smoothness: The Struggles of Neural Compressors with Discontinuous Mappings

Ezgi Özyılkan[*] **Jona Ballé**[*] **Sourbh Bhadane**[+] **Aaron B. Wagner**[†] **Elza Erkip**[*]
[*]New York University [+]University of Amsterdam [†]Cornell University
{ezgi.ozyilkan, jona.balle, elza}@nyu.edu s.n.bhadane@uva.nl wagner@cornell.edu

Abstract

Artificial Neural Networks (ANNs) have revolutionized data compression by effectively learning nonlinear transformations directly from data. The Nonlinear Transform Coding (NTC) framework has demonstrated notable success in achieving favorable rate–distortion trade-offs, particularly for real-world multimedia such as image and video. Despite this progress, fundamental questions remain about whether NTC can compress various types of input sources *optimally*, and if not, where and why it falls short. To investigate these questions, we focus on simpler, closed-form sources where optimal compression strategies are well-characterized using tools from information theory. Reviewing key failure modes in NTC-based compressors from the literature points to a common underlying issue: their difficulty in learning high-frequency and discontinuous functions, leading to suboptimal compression performance compared to the information-theoretic optimum in certain setups. We also review several remedies that alleviate these failure modes, including a new one based on Fourier embeddings. By drawing a connection between these suboptimalities, our work provides a unified and fresh perspective on understanding them, thereby representing a step toward improving neural data compression.

1 Introduction

Artificial Neural Networks (ANNs) have demonstrated remarkable versatility across diverse tasks, revolutionizing many fields such as computer vision and natural language processing. Their ability to learn complex and nonlinear relationships makes them particularly effective for data compression, where they can represent high-dimensional data compactly while maintaining reasonable distortion (fidelity) values. Leveraging this capability, the Nonlinear Transform Coding (NTC) framework [1], which encompasses most published neural compressors [e.g., 2–4], has emerged as a highly competitive approach for lossy compression, particularly for image and video data [5, 6]. Unlike the Karhunen–Loève Transform or the Discrete Cosine Transform, which are limited to linear transformations, NTC leverages the *universal function approximation capability* of ANNs [7, 8] for nonlinear dimensionality reduction, learning all transforms directly from data and achieving superior rate–distortion trade-offs, especially for multimedia sources.

Despite the ongoing success story of NTC, fundamental questions remain: Is the compression performance of NTC optimal or near-optimal? Can the learning capabilities of NTC-based compressors be improved further? To address these questions, we turn our attention to simpler source models and consider the problem through an information-theoretic lens. This involves comparing the compression performance achieved by NTC against known theoretical limits. The advantage of focusing on these simpler synthetic sources is that, unlike real-world image and video data, their optimal compression strategies and/or rate–distortion trade-offs can be explicitly characterized using tools from information theory. By comparing NTC's performance against these theoretical benchmarks, its potential failure modes can be identified.

To understand how NTC handles different compression setups, we first examine compression of several stylized source models that reveal both its strengths and weaknesses; we refer to this first setting as the *point-to-point* (i.e., one encoder, one decoder) case. One such source is the *sawbridge* process [9], for which NTC compresses the process optimally. We then consider more complex sources, such as the *circle* and *ramp* processes [10], both of which introduce circular topologies, and for which naively using NTC-based compressors results in performance degradation. We examine modifications to NTC that enable it to compress these sources more effectively and, in some cases, near-optimally. The failures are apparently tied to the circular topologies, which necessitate sharp discontinuities in either the encoder or decoder transforms. This is in line with the observation that neural networks have an inherent smoothness bias [11–14].

Separately, a similar limitation of NTC has been observed in the context of *distributed compression* [15–17] (i.e., multiple encoders). In such scenarios, the encoders are able to send only partial information, because what is missing can be recovered at the decoder side due to known correlations between the various sources. Here, the failure modes of NTC [18, Sec. 2] appear to be connected with the technique of *binning* [19], which is an effective strategy to reduce the compression rate for typical correlation patterns. Binning groups together source realizations appearing in distant regions, which translates to discontinuities in the encoder transform, similar to the circular setups mentioned above. Likewise, we discuss remedies that enhance NTC's compression performance in distributed setups by replacing the Euclidean/continuous latent space with a categorical one [20, 18]. The two failure cases thus appear to have the same fundamental root cause, which we discuss further in the final section.

2 The Point-to-Point Case

The conceptual basis for classical compression algorithms such as linear transform coding (as in JPEG), is the rate–distortion theory of stationary Gaussian sources with mean-squared error (MSE) distortion [21, Sec. 4.5.2]. This theory asserts that nonlinear transforms are not needed; linear decorrelating transforms combined with entropy-coded quantization is close to optimal [22, Sec. 5.6.2]. The fact that, for images, NTC provides superior compression to methods that follow this architecture indicates that new source models are needed to explain the effectiveness of NTC and to identify its limitations.

Given that the distribution of natural images has long been suspected to be supported by a low-dimensional manifold in pixel-space [e.g., 23], it is reasonable to conjecture that ANN-based compressors excel at compressing sources with low-dimensional manifold structure in a high-dimensional ambient space. Hence, it is natural to consider simple source models that exhibit similar structure. To this end, Wagner and Ballé [9] propose considering the *sawbridge* process, which was previously studied in the survey by Donoho *et al.* [24]. This process is characterized by a jump from the "rail" t to the "rail" $t-1$ at a random time over $[0,1]$. More precisely, let U be uniformly distributed over $[0,1]$ and define the continuous-time process as:

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

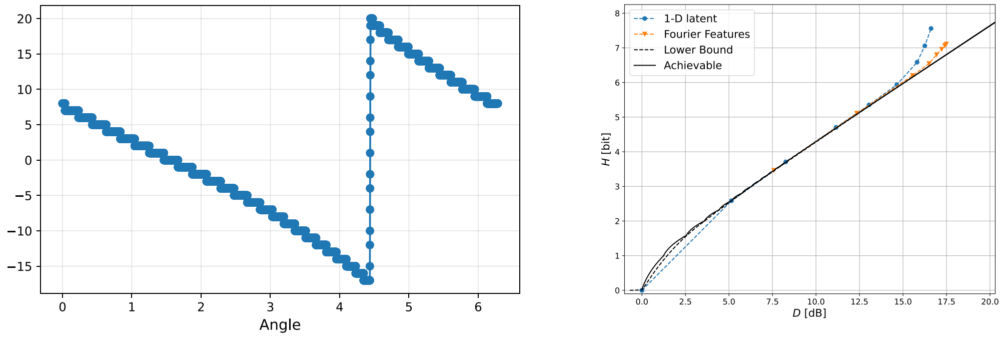
Since the process is completely determined by the realization of U, the set of realizations form a 1-D curved manifold in function space. Wagner and Ballé characterize the optimal one-shot compression performance for this source under an MSE distortion measure, and they find that an NTC-based compressor trained on sawbridge realizations empirically achieves this optimal performance [9].

Curiously, they find that this does not extend to other, similar sources. The *ramp* source, obtained by applying a random cyclic shift to the function $t \mapsto t - 1/2$ over the interval $[0,1]$, is defined as:

$$Y(t) = [(t+U) \mod 1] - 1/2 \quad t \in [0,1]. \tag{2}$$

Here, the manifold of realizations is again one-dimensional, but now it forms a circle. Bhadane *et al.* [10] characterize the optimal one-shot compression performance and show that the encoder should send a quantized version of U to the decoder, which should then output the minimum mean-squared error estimate of the ramp process $\{X(t)\}_{t=0}^{1}$ given this information. They find that NTC-based compressors with 1-D latent spaces are optimal at low rates but not at high rates (Fig. 1b).

Bhadane *et al.* note that the reason this source is difficult for NTC compressors to handle can be understood by considering an even simpler source, namely the *circle*, where data points are distributed



(a) circle

(b) ramp

Figure 1: Results of training NTC. (a) Quantized encoder output vs. angle U for the circle and 1-D latent, taken from [10, Fig. 4a]. Even ignoring the effects of quantization, the encoder is not invertible because its implementation of the step discontinuity is insufficiently steep. (b) Rate–distortion curves for the ramp. Dashed black: theoretical lower bound [10]. Solid black: theoretical upper bound on the optimal performance [10]. Blue: NTC without the aid of Fourier features [10, Fig. 3]. Orange: NTC with decoder provided random Fourier features of the 1-D latent (proposed, Sec. 2.2).

uniformly along a unit circle: $\vec{X} = [\cos(2\pi U) \;\; \sin(2\pi U)]$. This process exhibits "low-dimensional structure in high-dimensional space" in that the dimension of the manifold is one while the dimension of the ambient space is two. An optimal one-shot compressor for this source is to communicate a uniform quantization of U [10]. NTC-based compressors with a 1-D latent space are not optimal for this source at high rate, and the reason is illustrative. To approach optimality, at the encoder side, NTC must learn to implement the map $\vec{X} \mapsto h(U)$ for some invertible function $h(\cdot)$. The issue is that the map $\vec{X} \mapsto h(U)$ must be discontinuous if $h(\cdot)$ is invertible, but, as hypothesized by Bhadane *et al.*, the neural network at the encoder can only implement continuous maps. For simplicity, assuming $h(\cdot)$ to be identity, the function $\vec{X} \mapsto U$ has a step discontinuity around $\vec{X} = (1, 0)$. In practice, the training process learns a smoothed version of this function, which means the nonlinear transform at the encoder is not invertible (Fig. 1a). Therefore, the decoder cannot determine the source realization even if no quantization is performed. At low rate, this identifiability problem is dominated by the quantization error and has a negligible effect. At high rate, however, it prevents NTC from achieving the optimal performance [10, Fig. 1]. For the ramp, the problem turns out to be similar (Fig. 1b) but now with the roles of the encoder and decoder reversed. It is trivial for the encoder to learn a map $X \mapsto h(U)$ for some invertible function $h(\cdot)$: the coordinate function $X \mapsto X(t_0)$ serves this purpose for any fixed time t_0. The problem is that for any time t_1, the map $U \mapsto X(t_1)$ has a step discontinuity, so the decoder has to implement many discontinuous functions (one for each time sample at the output).

2.1 Adding Latent Dimensions and Increasing Batch Size

Bhadane *et al.* [10] show that for the circle, the problem of the discontinuity at the encoder can be alleviated by moving to a 2-D latent space. The idea is that one latent can convey how far the source realization is from a pole, and the other can be used to transmit a single bit indicating which hemisphere the point is in. This approach does not solve the problem for the ramp, however [10, Fig. 3]. Tancik *et al.* [11] argue via a neural tangent kernel analysis that stochastically trained ANNs are biased towards smoothness because they learn low-frequency functions quickly and high-frequency functions slowly. In line with this, Bhadane *et al.* also show that increasing the batch size helps close the gap to optimality [10, Fig. 1], as it reduces stochasticity of the training process, and hence the smoothness bias.

2.2 Expanding the Source Space with Fourier Embeddings

Following the remedy originally discussed by Tancik *et al.* [11], we can also expedite the learning of high-frequency functions by supplying random Fourier features of the input to the ANNs. The desired functions are then easier to synthesize from the inputs available to the network.

In the context of the ramp, we propose a new approach. Instead of supplying the ANN at the decoder with the (1-D) latent variable Y as in [10], we now input the vector:

$$(\cos(a_1 Y), \cos(a_2 Y), \ldots, \cos(a_m Y)), \tag{3}$$

where the parameters $a_1, \ldots, a_m$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, and m and $\sigma^2 > 0$ are hyperparameters. We find that the performance is sensitive to the choice of m and σ^2, but with appropriate choices of these hyperparameters the performance can be made quite close to the optimal rate–distortion trade-off for the ramp [10, Thm. 5] (Fig. 1b).

The two sources for which this phenomenon arises, the circle and the ramp, both take the form of circular manifolds. This raises the question of whether the phenomenon is peculiar to sources with circular symmetry. In the next section, we shall see that when one moves from the point-to-point setting to *distributed* scenarios, then similar issues arise even for Gaussian sources.

3 The Distributed Case: Side Information at the Decoder

Distributed compression considers the problem of efficiently compressing correlated sources from encoders that do not directly communicate with each other, but whose outputs are decoded jointly [15]. While the theory predicts substantial improvements in compression efficiency over point-to-point setups, practical implementations have lagged, primarily due to the challenge of handling complex correlations between and within the sources [17]. Here, we focus on a lossy compression setup characterized by Wyner and Ziv [16], which represents a simpler special case where the decoder has lossless access to a correlated source, the *side information*, yet still captures the core challenges of distributed compression. Note that the Wyner–Ziv (W-Z) rate–distortion (R-D) bound has been well characterized for Gaussian sources with linear correlation under the MSE distortion metric in the asymptotic regime [16], i.e., when compressing many source realizations at the same time.

To evaluate how NTC performs in a W-Z setup, Ozyilkan *et al.* [18] consider a simple test case involving a one-shot compression setting (i.e., compressing one source realization at a time) where the decoder has access to a version of the source corrupted by additive noise. Specifically, let X be the input to the encoder and Y the side information available at the decoder (Fig. 2), where both are zero-mean, stationary Gaussian memoryless sources. The correlation model is given by $X = Y + N$, having independent $Y \sim \mathrm{N}(0, 1.0)$ and $N \sim \mathrm{N}(0, 0.1)$.

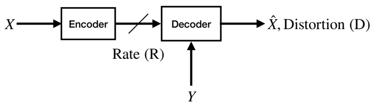


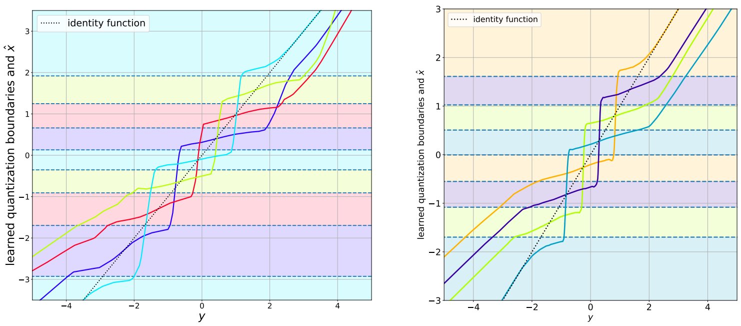
Figure 2: Lossy compression with decoder-only side information, the *Wyner–Ziv* setup, in the one-shot regime.

By evaluating various learned compressors operating under different trade-offs, Ozyilkan *et al.* [18] report that, at low rates, the NTC-based framework achieves slightly better performance than the R-D bound of the setup without side information (Fig. 4 in the Appendix). Analyzing the transform functions recovered by these NTC models reveals that this class of neural compressors learns exactly symmetric groupings around $(x = 0)$ in the source space at low rates, allowing it to make limited use of the side information (Fig. 5 in the Appendix). However, at higher rates, Ozyilkan *et al.* find no evidence of grouping in the source space (not shown) and as a result, the NTC-based compressor performs not much better the R-D bound without side information (Fig. 4).

Analogous to, but independent of the point-to-point case (Sec. 2), the authors attribute this suboptimal performance to NTC's inability to learn discontinuous many-to-one functions. However, in the present case, the necessity for discontinuities does not arise from source topology, but is likely linked to *binning*. While the W-Z achievability proof is non-constructive by nature, it involves *random binning* arguments, and existing constructive approaches to distributed compression are also based on binning [25]. Since binning is intrinsically discontinuous, this suggests that a practical distributed compressor may need to learn discontinuous functions to achieve efficient compression. The two solutions proposed by Ozyilkan *et al.* are to replace the Euclidean latent space of NTC with a *categorical* space.

3.1 Concrete Distribution for Continuous Relaxation

In [20], the encoder is a *structured* vector quantizer, parameterized by a neural network. This network outputs a vector $(\alpha_1, \ldots, \alpha_K)$, where K is the length of the codebook, and the quantization index



(a) Neural distributed compressor using Concrete distribution [26], taken from [20, Fig. 2a].

(b) Neural distributed compressor using variational ECVQ formulation, taken from [18, Fig. 5a].

Figure 3: Visualizations of the learned encoder output u and neural decoder $\hat{x} = g_{\boldsymbol{\phi}}(u, y)$ for the Gaussian Wyner–Ziv setup. The dashed lines are quantization boundaries, with colors representing unique values of u. The decoding function is shown for each value of u using the corresponding color. Both models outperform the rate–distortion function without side information (Fig. 4).

is chosen as $u = \arg\max_i \alpha_i$, i.e., the largest element of this vector. This enables the encoder to recover the equivalent of discontinuous maps in NTC: For the same index k to be chosen in disjoint regions of the source space (for example, yellow regions as in Fig. 3(a)), the map $x \mapsto \alpha_k$ (where x is a realization of X) can still be a smooth function. The decoder is given by another neural network $g_{\boldsymbol{\phi}}(u, y)$ parameters $\boldsymbol{\phi}$, outputting the reconstruction $\hat{x}$.

However, since the $\arg\max$ operation is not differentiable, training of this model using stochastic gradient descent (SGD) is a challenge. The authors leverage the Concrete distribution [26], a proxy for the $\arg\max$ which is both stochastic and *soft*, making it differentiable. Instead of yielding a single quantization index u, they obtain a distribution over the interior of the K-simplex, defined using a *softmax* function as:

$$U_k = \frac{\exp((\alpha_k + G_k) \,/\, t)}{\sum_{i=1}^{K} \exp((\alpha_i + G_i) \,/\, t)} \;, \tag{4}$$

where G_k are i.i.d. samples from a standard Gumbel distribution. Here, t is a temperature parameter that controls the degree of relaxation, allowing the sampling process to smoothly transition from soft to hard as $t \to 0^+$. However, even at $t = 0$, U_k is still a distribution over quantization indices, which represents a mismatch between training and inference.

As shown in the visualization of the learned compressor obtained with this formulation in Fig. 3(a), it enables the learned compressor to recover binning behavior, as evidenced by the periodic grouping in the source space, in contrast to the strictly symmetric grouping achieved by NTC (Fig.5). This leads to improved compression performance, as illustrated in Fig. 4, by making more efficient use of side information Y during the compression of X.

3.2 Variational Entropy-Constrained Vector Quantization (ECVQ)

In [18], the encoder is of the form of a classical entropy-coded vector quantizer (ECVQ) [27], choosing the best possible codeword by exhaustively evaluating all options, while the decoder is still implemented by a neural network. This clearly enables the encoder to implement arbitrary maps.

For a given $X = x$, Ozyilkan *et al.* [18] set the encoder function as:

$$u = \arg\min_{k \in K} \mathbb{E}_{p(y|x)} \Big[\underbrace{-\log p_{\boldsymbol{\xi}}(k)}_{\text{rate}} + \lambda \cdot \underbrace{d(x, g_{\boldsymbol{\phi}}(k, y))}_{\text{distortion}} \Big], \tag{5}$$

where $g_{\phi}(k, y)$ is the decoder, as above, and $p_{\boldsymbol{\xi}}(k) = \frac{\exp \xi_k}{\sum_{i=1}^{K} \exp \xi_i}$ represents the entropy model, where $\boldsymbol{\xi}$ denotes a vector of unnormalized probabilities. Here, $\lambda > 0$ controls the trade-off and $d(\cdot, \cdot)$ is a distortion measure, which is MSE for the quadratic-Gaussian WZ case that the authors considered. This solution explicitly enumerates all possible quantization indices k. Note that the encoder function in Eq. (5) does not have any learnable parameters of its own, but depends on the probability model and the decoding function, which makes it trainable using SGD.

Figs. 3(b) and 4 visualize the learned compressor obtained by optimizing an objective based on

Eq. (5), and show its corresponding R-D performance, respectively. As seen in Fig. 3(b), this learned compressor again exhibits binning behavior, which explains its improved R-D performance over NTC.

4 Discussion

In the previous sections, we summarized the difficulties arising from using NTC to compress sources with circular topology (Sec. 2), as well as using it in distributed scenarios which favor binning (Sec. 3). Here, we establish a link between the two observations: Namely, that within the standard NTC framework, both circular topologies and binning require either the encoder or the decoder transform to implement discontinuous functions. The inability to learn these functions is likely a consequence of what is called the *spectral bias* of ANNs [28, 12, 13] in the learning literature, which induces a learning bias towards low-frequency modes and thereby, favors learning smooth functions instead [11, 14]. In many applications, the bias towards smooth functions may actually be beneficial, as it helps prevent overfitting when training data is sparse. However, as observed, this learning bias can also negatively affect the compression performance of NTC. In fact, transform coding is not the only domain affected by this phenomenon: Similar issues have been observed in the use of autoencoders for anomaly detection in high-energy physics, where circular or rotational invariances arise naturally [29]. Esmaeili *et al.* [30] more generally discuss difficulties learning circular geometries in the context of variational autoencoders.

So, what can be done to improve NTC in this regard? The measures presented above can be effective, but represent only partial remedies. While adding a latent dimension is effective for the circle source (discontinuity in the encoder), it does not help with the ramp (discontinuity in the decoder) (Sec. 2.1). Increasing the batch size and Fourier embeddings help, in that they ameliorate spectral bias, but they still do not allow the network to implement true discontinuities (Sec. 2.2). They also come at the cost of increased computational complexity, during training or inference, respectively; may not be feasible when there is limited training data; and may also lead to more overfitting. In the context of image compression, however, the neural networks in question are typically relatively small while the data sets are relatively large, making overfitting less of a concern.

Switching from a continuous to a categorical latent space is equivalent to moving from transform coding towards *unstructured* vector quantization, with the promise that discontinuous maps between the source space and the code space are easily implemented. However, when the encoder is completely unstructured (Sec. 3.2), i.e., searches for the best codeword exhaustively, it inherits classical limitations of VQ: It fails to scale to higher bit rates, as its maximum rate is upper bounded by $\log_2 M$, where M is the (pre-determined) codebook size. The codebook size cannot be arbitrarily increased, as the computational complexity of the exhaustive search grows exponentially with M. Furthermore, if we try to mitigate this by *structuring* the encoder, for example using a neural network, we are still left with training difficulties: the loss function of the formulation using Concrete distribution (Sec. 3.1) is not a good approximation of the true objective (the rate–distortion Lagrangian). In order to mitigate the inevitable bias in the objective, tuning hyperparameters such as the temperature value t in Eq. (4) is essential [26]. Finally, all algorithms for fitting vector quantizers are sensitive to initialization [31], which includes classical ones like the Lloyd iteration [32], as well as the ones reviewed above [20, 18].

While our present work does introduce one additional way to mitigate the problem (Fourier features), the solutions presented here do not provide a universal fix to NTC that retains all the desirable properties of a learned compressor alluded to above. Yet, we believe that there is value in drawing these connections between the independently observed issues in the point-to-point case and the distributed case, in that it unifies the thinking about what may be the cause of the failures. As such, it may help in finding better solutions going forward.

On a higher level, it is also worth noting that analyzing suboptimalities of data compression of synthetic sources benefits from exact or near-exact converse bounding techniques, which, while not always being constructive, tell us how close practical methods are to optimality. This information-theoretic approach might be transferable to other machine learning techniques, where it otherwise may be unclear how to identify suboptimalities.

References

[1] Johannes Ballé, Philip A. Chou, David Minnen, Saurabh Singh, Nick Johnston, Eirikur Agustsson, Sung Jin Hwang, and George Toderici. Nonlinear transform coding. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):339–353, 2021.

[2] J. Ballé, V. Laparra, and E. P. Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2017.

[3] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.

[4] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, pages 10771–10780, 2018.

[5] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 1503–1511, New York, NY, USA, 2022. Association for Computing Machinery.

[6] Ties van Rozendaal, Tushar Singhal, Hoang Le, Guillaume Sautiere, Amir Said, Krishna Buska, Anjuman Raha, Dimitris Kalatzis, Hitarth Mehta, Frank Mayer, Liang Zhang, Markus Nagel, and Auke Wiggers. Mobilenvc: Real-time 1080p neural video compression on a mobile device. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4311–4321. IEEE, 2024.

[7] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, January 1993.

[8] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, Jul 1989.

[9] Aaron B. Wagner and Johannes Ballé. Neural networks optimally compress the sawbridge. In *2021 Data Compression Conference*, pages 143–152, 2021.

[10] Sourbh Bhadane, Aaron B. Wagner, and Johannes Ballé. Do neural networks compress manifolds optimally? In *2022 IEEE Information Theory Workshop*, pages 582–587, 2022.

[11] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[12] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5301–5310, 2019.

[13] Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. Frequency bias in neural networks for input of non-uniform density. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[14] Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[15] D. Slepian and J. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4):471 – 480, 1973.

[16] A. Wyner and J. Ziv. The rate–distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 22(1):1 – 10, 1976.

[17] Ezgi Özyılkan and Elza Erkip. Distributed compression in the era of machine learning: A review of recent advances. In *2024 58th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, 2024.

[18] Ezgi Ozyilkan, Johannes Ballé, and Elza Erkip. Neural distributed compressor discovers binning. *IEEE Journal on Selected Areas in Information Theory*, 5:246–260, 2024.

[19] T. Cover. A proof of the data compression theorem of slepian and wolf for ergodic sources (corresp.). *IEEE Transactions on Information Theory*, 21(2):226–228, 1975.

[20] Ezgi Özyılkan, Johannes Ballé, and Elza Erkip. Learned Wyner–Ziv compressors recover binning. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 701–706, 2023.

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

[22] William A. Pearlman and Amir Said. *Digital Signal Compression: Principles and Practice*. Cambridge University Press, 2011.

[23] Olivier J. Hénaff, Johannes Ballé, Neil C. Rabinowitz, and Eero P. Simoncelli. The local low-dimensionality of natural images. In *International Conference on Learning Representations (ICLR)*, 2015.

[24] D.L. Donoho, M. Vetterli, R.A. DeVore, and I. Daubechies. Data compression and harmonic analysis. *IEEE Transactions on Information Theory*, 44(6):2435–2476, 1998.

[25] S.S. Pradhan and K. Ramchandran. Distributed source coding using syndromes (DISCUS): Design and construction. *IEEE Transactions on Information Theory*, 49(3):626–643, 2003.

[26] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

[27] P.A. Chou, T. Lookabaugh, and R.M. Gray. Entropy-constrained vector quantization. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(1):31–42, 1989.

[28] Ronen Basri, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. In *Advances in Neural Information Processing Systems*, 2019.

[29] Joshua Batson, C. Grace Haaf, Yonatan Kahn, and Daniel A. Roberts. Topological obstructions to autoencoding. *J. High Energ. Phys.*, 280, 2021.

[30] Babak Esmaeili, Robin Walters, Heiko Zimmermann, and Jan-Willem van de Meent. Topological obstructions and how to avoid them. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 8865–8884. Curran Associates, Inc., 2023.

[31] Peric Zoran and Jelena Nikolic. An effective method for initialization of Lloyd–Max's algorithm of optimal scalar quantization for laplacian source. *Informatica, Lith. Acad. Sci.*, 18:279–288, 01 2007.

[32] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[33] R.M. Gray and D.L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998.

[34] S. Na and D.L. Neuhoff. Bennett's integral for vector quantizers. *IEEE Transactions on Information Theory*, 41(4):886–900, 1995.

A Appendix

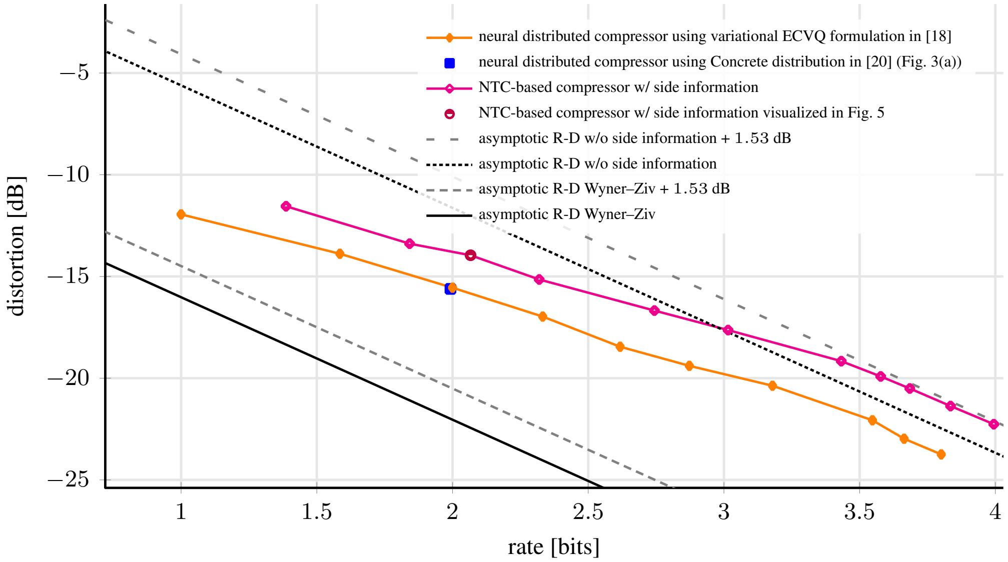


Figure 4: Rate–distortion (R-D) performances obtained with neural distributed compressors using Concrete distribution [20], variational entropy-constrained vector quantization (ECVQ) formulation [18], and an NTC-based compressor with side information [18][Sec. 2]. The quadratic-Gaussian Wyner–Ziv setup is considered, where $X = Y + N$ with independent $Y \sim \mathrm{N}(0, 1.0)$ and $N \sim \mathrm{N}(0, 0.1)$. The empirical results are plotted versus the lossy compression without side information and Wyner–Ziv R-D bounds, both of which assume that the blocklength of the sources is asymptotically large. 1.53 dB refers to the mean-squared error gap that the entropy-constrained scalar (one-shot) lattice quantizer is subjected to in a high-rate regime [33], due to space-filling loss (also known as cubic loss [34]). Figure is adopted from [18, Fig. 6a].

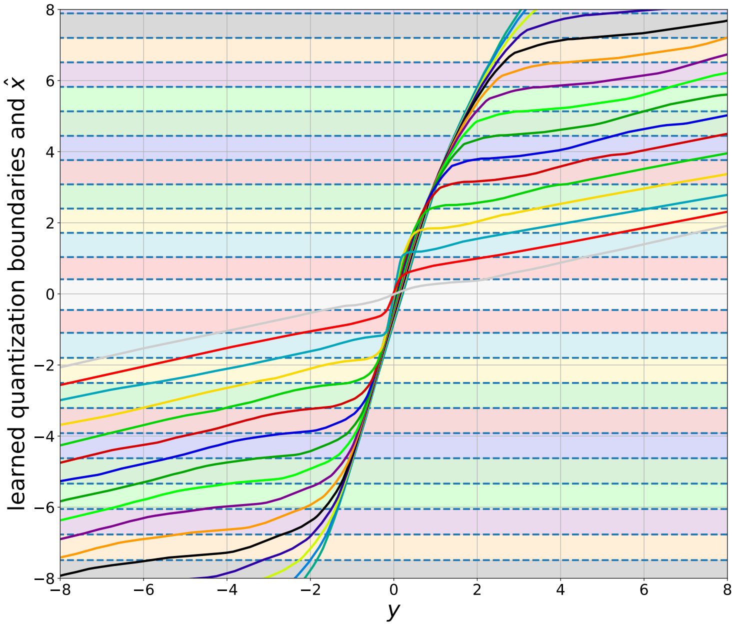


Figure 5: Visualization (best viewed in color) of an NTC-based compressor with side information, where the neural encoder output is $u = \lfloor f_{\boldsymbol{\theta}}(x) \rceil$ and the neural decoder output is $\hat{x} = g_{\boldsymbol{\phi}}(u, y)$ (see Eq. 1 in [18]), where $f_{\boldsymbol{\theta}}$ and $g_{\boldsymbol{\phi}}$ are learned encoder (*analysis*), $f_{\boldsymbol{\theta}}$, and decoder (*synthesis*) $g_{\boldsymbol{\phi}}$ functions, with parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, respectively. As in Fig. 4, The quadratic-Gaussian Wyner–Ziv setup is considered, where $X = Y + N$ with independent $Y \sim \mathrm{N}(0, 1.0)$ and $N \sim \mathrm{N}(0, 0.1)$. The dashed horizontal lines are quantization boundaries, and the colors between boundaries represent unique values of u. The decoding function is depicted as separate plots for each value of u, using the same color assignment. The visualized model achieves -13.96 dB at 2.07 bits, yielding better rate–distortion (R-D) performance than the R-D function without side information (Fig. 4). Figure is taken from [18, Fig. 9].