AN INFORMATION-THEORETIC APPROACH TO OPTIMAL NEURAL-NETWORK-BASED COMPRESSION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Sourbh Bhadane

August 2023

AN INFORMATION-THEORETIC APPROACH TO OPTIMAL NEURAL-NETWORK-BASED COMPRESSION

Sourbh Bhadane, Ph.D.

Cornell University 2023

Modern artificial-neural-network-based (ANN-based) compressors have recently achieved notable successes on compressing multimedia formats such as images. This is despite information-theoretic near-optimality results of the linear transform coding paradigm, which forms the basis of existing standard lossy compressors such as JPEG, AAC etc., for stationary Gaussian sources with respect to mean-squared error distortion (at high rate). This thesis attempts to fill in some of the gaps in our theoretical understanding of modern ANN-based compressors. We list our contributions below.

We propose a set of sources that obey the manifold hypothesis, i.e., that are high-dimensional in input space but lie on a low-dimensional manifold. We analytically derive optimal entropy-distortion tradeoffs for such sources and test the performance of ANN-based compressors on them. We find that for some sources that exhibit circular symmetry, ANN-based compressors are suboptimal. Our fix to this issue involves embedding Random Fourier Features (RFFs) before passing the input through either encoding or decoding nonlinear transforms.

As the set of manifold sources gets more sophisticated, exact characterization of entropy-distortion tradeoffs can get challenging. We focus on the low-rate regime and develop general methods for one-bit quantization of sources in an arbitrary Hilbert space. Using these methods, we derive optimal one-bit quantizers for several examples including elliptical distributions and a manifold source that we proposed. We also study the low-rate asymptotics for variable-rate dithered quantization for vector Gaussian sources.

We revisit the ubiquitous autoencoder architecture and analyze dimensionality-reducing linear autoencoders that are often used fo r general-purpose lossy compression. We propose an alternate autoencoder formulation that embraces the compression point of view by constraining the number of bits required to represent the output of the encoder. Our characterization of the optimal solution to this non-convex constrained linear autoencoder involves generalizing to any Schur-concave constraint on the variances of the encoder output. We provide experimental validation of our autoencoder-based variable-rate compressor.

BIOGRAPHICAL SKETCH

Sourbh Bhadane was born in Pune, India. He obtained a Bachelor of Technology and Master of Technology degree in Electrical Engineering from Indian Institute of Technology, Madras. Following that, he moved to the US to pursue a PhD in Electrical and Computer Engineering at Cornell University. His research interests include data compression, machine learning and more recently, causal data science.

To the village it took to get me here.

ACKNOWLEDGEMENTS

I would like to express my immense gratitude to my advisor Prof. Aaron Wagner. I would have probably quit long back if not for his kind words of support and encouragement. I consider myself fortunate for Aaron's generosity with his time and endless patience throughout my PhD journey. While there are many qualities of his I wish to imbibe in my professional career, the one that has had the most impact on me are his ability to clearly and concisely present even the most complicated of ideas.

I would also like to thank my advisor Prof. Jayadev Acharya for his enthusiastic encouragement especially during the initial days of my PhD. Jayadev pushed me to explore as widely as I could and always lent a patient ear to my half-baked research ramblings. I regret not making more of the orthogonal advising styles of Jayadev and Aaron.

I am thankful to the minor members of my committee, Prof. Kilian Weinberger and Prof. Ziv Goldfeld for their insightful comments and advice on my work during all stages of the program.

I have been fortunate to learn from multiple collaborators. Thanks to Johannes Ballé and Lucas Theis for being generous with their expertise and intuition of what works and what doesn't when it comes to training neural-network-based compressors. Thanks to Prof. Arnab Bhattacharyya and Saravanan Kandasamy for being excellent first teachers of causality. I am grateful to Ziteng Sun for being a very accessible collaborator with expertise on a range of topics. I would also like to thank my past and present office-mates and groupmates - Huanyu, Yuhan, Nirmal, Adeel, Yang and Sharang.

My stay at Cornell would have been very lonely without friends whose constant companionship gave me much needed support, which I am afraid has gone unacknowledged thus far. Thank you Aditya, Jashan, Karishni, Katherine, Kunal, Rewa, Shantanu, Shanthanu, Shriya, Vikram and many others. I am grateful to everyone who I have played ultimate with, and the folks at the Ithaca Area Ultimate Alliance for building a community so welcoming that someone new to the game,

like myself, felt at home. Special thanks to Emily, Dan and Ray for being generous with their time and much cherished company. Many thanks to Avinash, James, Sudeep and Thomas for the innumerable impromptu pickup games.

My academic journey would not have even started if not for my parents' sacrifices and unwavering commitment to providing the best educational environment for their kids at any cost. Thanks to my sister, Yuktha, for her love and roasts. Despite being away from home during the pandemic, my family's unconditional love and constant support has kept me going through my PhD.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In 2016, we ushered into the "zettabyte" era where global IP traffic exceeded one zettabyte.[1] Multimedia data formats such as video, constitute around 75% of global mobile data traffic [22]. With the number of internet users projected to only increase, the amount of data generated and consumed is slated to grow by 20% annually [22]. The increased data-demand can only be fulfilled by a corresponding upgradation of technology that facilitates this massive bit-transfer seamlessly. Data compression, broadly thought of as the suite of technologies that reduce data to the form that requires least storage, measured by bits, is a backbone of the data-dependent world.

Data compression techniques differ according to the data to be compressed. For the purposes of compression, data can be classified into two main classes; 1) symbolic, meaning a sequence of symbols from an alphabet, e.g. English text, and 2) numerical, meaning information obtained by measuring some physical quantity, e.g. audio, images, video etc. [53]. The key difference when it comes to compressing these different data sources is with regard to the requirement of fidelity to the original data. Symbolic data demands a perfect reconstruction of symbols primarily because the symbols could be arbitrary. On the other hand, compressing numerical data is possible if some error is allowed since one can craft some measure of distance to capture the phenomenon of a reconstruction being "close" to the original data. As a result compression techniques for compressing both forms of data can be quite different. "Lossless compression" refers to compressing data with no reconstruction loss, often applied to symbolic data, and "Lossy compression" refers to compression techniques that allow some reconstruction loss, often applied to numerical data. This thesis will only be concerned with the latter.

A rudimentary example of lossy compression is rounding a real number, say for simplicity,

[1] Zettabyte is a unit of digital storage equivalent to 10^{21} bytes.

whose absolute value is bounded by a large integer M. Rounding a real number x to its nearest integer can be represented as a quantization operation $Q(x)$ where $Q : \mathbb{R} \mapsto \mathbb{Z}$. The quantized "reconstructions" are integers on the real line $\{i : i \in \mathbb{Z}\}$, the quantization " intervals" are the intervals between consecutive half-integers $[i - 1/2, i + 1/2)$, and half-integers are the quantization "thresholds" $\{i + 1/2 : i \in \mathbb{Z}\}$. Therefore, the number of bits required to represent any real number is the number of bits required to succinctly represent the rounded real number, preferably in a lossless manner; necessitating lossless compression techniques. If all reconstructions are equally likely, we encode the set of $2M$ integers using $\log_2 2M$ bits. This quantization process incurs an error $Q(x) - x$. If the quantization is finer, i.e., say the reconstructions are half-integers instead of integers, the quantizer error, loosely the distortion, decreases while the number of bits required, loosely the rate, increases. The problem of optimal quantizer design then becomes a question of designing the best quantizer that minimizes rate for a given distortion or, put another way, minimizes distortion for a given rate. This tension between rate and distortion is fundamental to both the theoretical and practical study of lossy compression as we will see in the rest of this thesis and in the brief overview in Section 1.2.

In practice, an important consideration for quantizer design, that we didn't touch upon in the rounding example above, is the statistics of data. For example, if we know that the real numbers in the previous example are mostly clustered around 0, a good quantizer would cluster more reconstructions around 0 to reduce distortion. For high-dimensional data formats such as multimedia, it is not always feasible to estimate statistics. Instead, existing standardized lossy compression algorithms such as JPEG, MP3, MPEG etc. [1] are designed based on the "linear transform coding" paradigm that we review in Section 1.3. The information-theoretic justification for this paradigm is due to a result that proves its near-optimality for stationary Gaussian sources at low-distortions.(see Section 1.3 for details).

However, 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., [8] and the references therein). The exemplary empirical performance of nonlinear transforms to compress multimedia sources, and the aforementioned information-theoretic near-optimality results of linear transform coding for stationary Gaussian sources, exposes a gap in our current understanding of ANN-based compression. Bridging this gap calls for a "modern compression theory" that takes into account a) transform architectures, b) optimal compression performance, typically measured by rate-distortion trade-offs, of non-Gaussian sources that are tractable, yet faithful to multimedia sources, c) distortion measures that go beyond the traditional, mean-squared error, and take in to account perception-based losses. This thesis takes preliminary steps in addressing the first two factors.

1.1 Contributions

In **Chapter 2**, we present a set of manifold sources that adhere to the manifold hypothesis in the sense that there is a large discrepancy between the amount of dimensionality reduction afforded by linear versus nonlinear transforms. We derive optimal one-shot compression tradeoffs for these sources analytically and test whether off-the-shelf ANN-based compressors can achieve the optimal performances. We find that for a few sources that exhibit circular symmetry, off-the-shelf ANN-based compressors are suboptimal. We propose a fix for this phenomenon by embedding Random Fourier Features (RFFs) before the encoders or decoders depending on where the issue lies.

Even for simple sources, deriving optimal compression performance tradeoffs can be tricky at times. Therefore, it is common to characterize low and high-rate limits of performance tradeoffs. In **Chapter 3**, we focus on low-rate analysis of a specific source that we also considered in Chapter 2,

namely the stationary sawbridge. In the process, we derive much more general methods for one-bit quantization of sources in Hilbert spaces. We also analyze the asymptotic low-rate performance on vector Gaussian sources of variable-rate compressors that have an architecture similar to ANN-based compressors.

In **Chapter 4**, we take a closer look at autoencoders, a common architecture of choice for ANN-based compressors. While nonlinear autoencoders are commonly used for dimensionality reduction and representation learning, linear autoencoders such as Principal Components Analysis (PCA), are used for general-purpose lossy compression of datasets [15]. We propose an autoencoder formulation, called *Principal Bit Analysis*, that embraces the compression view of autoencoders by quantizing the encoder output and imposing a constraint on the number of bits required to represent the same. We characterize the optimal solution to this non-convex constrained linear autoencoder by generalizing it to any Schur-concave constraint on the variances of the encoder output. We find that, unlike PCA, this method recovers the principal directions of the dataset and empirically outperforms PCA and other variable-rate compressors such as JPEG for images and AAC for audio.

In the remaining part of this chapter, we will overview a few basics of rate-distortion theory, transform coding and neural compression.

1.2 Rate-Distortion Theory

Assume that we have a "data source", modeled by a probability distribution $\mathrm{P}_{\mathrm{data}}$ that outputs a stream of independent data symbols $X_1, X_2 \cdots$ that take values from an alphabet $\mathcal{X}$.

Definition 1 (Fixed-rate Source Code)**.** *[23] A lossy compression algorithm, or a source code, with fixed-rate R, is the pair* (f_n, g_n) *where an encoder* $f_n : \mathcal{X}^n \mapsto \left\{1, 2, \cdots, 2^{nR}\right\}$ *maps a block of n*

i.i.d (independent and identically distributed) symbols drawn from the source, to an index, and a decoder $g_n : \left\{1, 2, \cdots, 2^{nR}\right\} \mapsto \hat{\mathcal{X}}^n$*, maps an index to a reconstruction block.*

The reconstructions $g_n(1), g_n(2), \cdots g_n(2^{nR})$ are the codebook of the code and the rate can be interpreted as log of the codebook size per symbol. Often, the reconstruction alphabet is same as the source alphabet. The distortion of such a code is defined by a single-letter distortion measure $d(x, \hat{x})$, often for analytical tractability, taken to be the mean-squared error. Assuming continuous alphabets where the square function is defined,

$$d_{\mathrm{sl}}(x, \hat{x}) \stackrel{\mathrm{def}}{=} (x - \hat{x})^2. \tag{1.1}$$

The block distortion between two n-letter sequences is defined using the single-letter distortion measure as

$$d(x^n, \hat{x}^n) \stackrel{\mathrm{def}}{=} \frac{1}{n} \sum_{i=1}^{n} d_{\mathrm{sl}}(x_i, \hat{x}_i). \tag{1.2}$$

Therefore, for the fixed-rate source code mentioned above, the distortion is defined as the expected value of distortion when the blocks are drawn i.i.d. from the source.

$$D = \mathbb{E}[d(X^n, g_n(f_n(X^n)))].$$

Definition 2 (Achievable (R, D) pair)**.** *[23] A rate-distortion pair* (R, D) *is said to be achievable if there exists a sequence of fixed-rate source codes* (f_n, g_n) *such that* $\lim_{n\to\infty} \mathbb{E}[d(X^n, g_n(f_n(X^n)))] \leq D$*.*

Definition 3 (Rate-Distortion Function)**.** *[23] The rate-distortion function* $R(D)$ *is the infimum of rates* R *such that* (R, D) *is in the closure of the set of all achievable* (R, D)*-pairs for a given distortion* D*.*

Characterizing the rate-distortion function of a source information-theoretically is perhaps the main theorem of rate-distortion theory.

Theorem 4. *[23] For an i.i.d source with distribution* $\mathrm{P}_{\mathrm{data}}$ *and bounded single-letter distortion* $d(x, \hat{x})$*, the rate-distortion function is*

$$R(D) = \min_{p(\hat{x}|x):\sum \mathrm{P}_{\mathrm{data}}(\mathrm{x})\mathrm{p}(\hat{\mathrm{x}}|\mathrm{x})\mathrm{d}(\mathrm{x},\hat{\mathrm{x}})\leq \mathrm{D}} I\left(X, \hat{X}\right).$$

A few remarks about the rate-distortion function: a) Theorem 4 holds for stationary and ergodic sources in general. b) $R(D)$ is a nonincreasing, convex function in D. c) An alternate view of Theorem 4 is that for any $\varepsilon > 0$, there exists a large enough n such that there exists a code of rate R that achieves distortion D such that $R < R(D) + \varepsilon$.

While Theorem 4 characterizes the rate-distortion function of a source given a distortion measure, analytically a closed-form solution is unknown for most practical data sources such as multimedia. Computing the rate-distortion function numerically is possible with the Blahut-Arimoto algorithm [23].

Achieving the rate-distortion function might require codes with increasing blocklengths to be optimal. In practice, it might not be possible to implement codes that achieve the rate-distortion function for arbitrary sources. Multiple approaches can be taken instead. One approach is to analyze the performance of simpler lossy compression algorithms, say for example, uniform scalar quantization, and establish how far their performance is from the rate-distortion function. For scalar sources, Gish and Pierce [29] show that the rate-distortion tradeoff attained by uniform scalar quantization is at most 0.255 bits within the rate-distortion function in the low-distortion regime for mean-squared error distortion. Transform codes, as we shall see in the next section, are similar simpler lossy compression algorithms that are near-optimal for stationary Gaussian sources.

Another approach is to consider compression algorithms that satisfy practical constraints; for example, limited memory might not permit using arbitrarily large blocklengths. Therefore, in practice, optimal one-shot rate-distortion tradeoffs are more useful. However, in such cases, a fixed-rate formulation, as considered in classical rate-distortion theory, cannot be applied. In Section 1.4, we will see how ANN-based compressors operate under such a formulation.

1.3 Transform Coding

While rate-distortion theory helps us establish limits of performance of compression algorithms, designing good compression algorithms require separate considerations. For an i.i.d. scalar source, [29] showed that uniform quantization is near-optimal in the high-rate regime. However, for sources with redundancy, or otherwise termed memory in the information theory literature, preprocessing might be required to exploit redundancy. It should be noted that while even for memoryless sources preprocessing is optimal, in general removing redundancy results in better performing compression algorithms.

Transform coding is one technique to reduce redundancy in sources with memory. Consider a scheme that only scalar quantizes each component of a source vector. When applied to sources with memory, a strict improvement would be to apply an orthogonal transformation such that the transformed source vector has uncorrelated components and further scalar quantize each uncorrelated component. This is precisely the intuition behind transform coding; blocks of source samples are projected onto an orthogonal basis, which is a linear transformation, to obtain the resulting transform coefficients. We will address the issue of quantizing transform coefficients after the following brief discussion about optimal transforms.

Let X be a zero-mean, wide-sense stationary source with autocorrelation function $R_X(\tau)$. For

blocks of length n, the source vector is $\boldsymbol{X} = (X_1, \cdots X_n)$. The covariance matrix can then be written as $\Phi_n \stackrel{\text{def}}{=} \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$, whose $(i, j)^{\text{th}}$ element is given by $R_X(|i - j|)$.The best transform codes are likely to leverage the statistics of the source.

Definition 5 (Karhunen-Loéve Transform (KLT))**.** *The Karhunen-Loéve Transform (KLT) coefficients are obtained by the dot product of* $\boldsymbol{X}$ *with the n eigenvectors of* Φ_n*.*

For arbitrary sources, the KLT is decorrelating, in the sense that the resultant transform coefficients are decorrelated. For Gaussian sources, they can be shown to be independent. In machine learning parlance, the KLT is also known as principal components analysis (PCA). Transform codes are usually compared using a ratio called the coding gain [1, p. 7.16], an approximation for the fractional reduction in distortion obtained by transform coding as opposed to scalar quantizing the source directly. It can be shown that the KLT achieves the highest coding gain among all possible transforms. While the KLT is considered optimal, it might not be feasible to obtain the statistics thereby making it less practical. Instead, other transforms are common place in practice; for e.g., the Discrete Cosine Transform (DCT) is ubiquitous in image and video compression.

Once a decorrelating transform is applied, the only question that remains is that of quantizing the decorrelated transform coefficients. One possible approach is vector quantization of the transform coefficients but this is almost as hard as compressing the original source. Another approach is to quantize each of the transform coefficients using an optimal scalar fixed-rate quantizer. Finding the optimal scalar fixed-rate quantizer for each transform coefficient might be computationally expensive but this approach can be made practical with uniform scalar quantization which is only slightly worse than the optimal fixed-rate scalar quantizer. The quantized transform coefficients can be further losslessly compressed by variable-length lossless coding techniques. Note that this differs from the fixed-rate lossy source codes in two ways; namely by being variable-rate and by being lossless codes. Since, the rate is no longer log of the size of the codebook as before, a use-

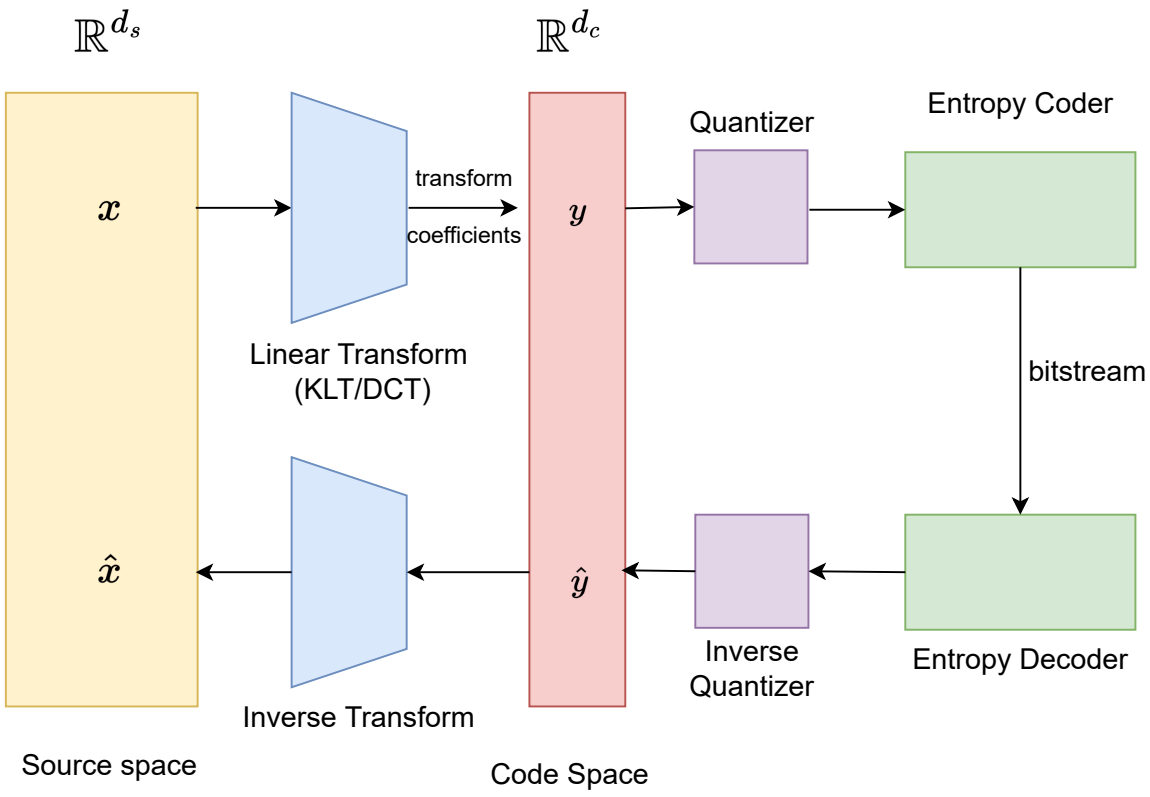


Figure 1.1: Linear Transform Coding

ful proxy, suggested by Shannon's limits of optimal lossless compression [59], is the entropy of quantized transform coefficients.

Combining the results of Gish and Pierce [29] and the optimality of the KLT for stationary Gaussian sources in the high-rate regime by Huang and Schultheiss [35], it can be shown that for stationary Gaussian sources, applying the KLT followed by entropy coding of the uniform scalar-quantized transform coefficients, is near-optimal at high-rates. Figure 1.1 shows a block diagram of this linear transform coding paradigm.

1.4 Neural Compression

Classical rate–distortion theory for Gaussian sources, is based on linear dimensionality reduction [12, 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 [1, 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 [1, 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., [32]), for instance. That is, while the linear span of the manifold may be high, there 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 paradigm called "nonlinear transform coding", with nonlinear analysis and synthesis transforms surrounding a conventional uniform quantizer with entropy coding [8]. Figure 1.2 illustrates the nonlinear transform coding paradigm.

A natural architecture choice for a nonlinear transform code is an autoencoder. An autoencoder (with *latent dimension* $k \leq d$) consists of an *encoder* $f : \mathbb{R}^{d_s} \mapsto \mathbb{R}^{d_c}$ and a *decoder* $g : \mathbb{R}^k \mapsto \mathbb{R}^d$. The goal is to select f and g from prespecified classes C_f and C_g respectively such that

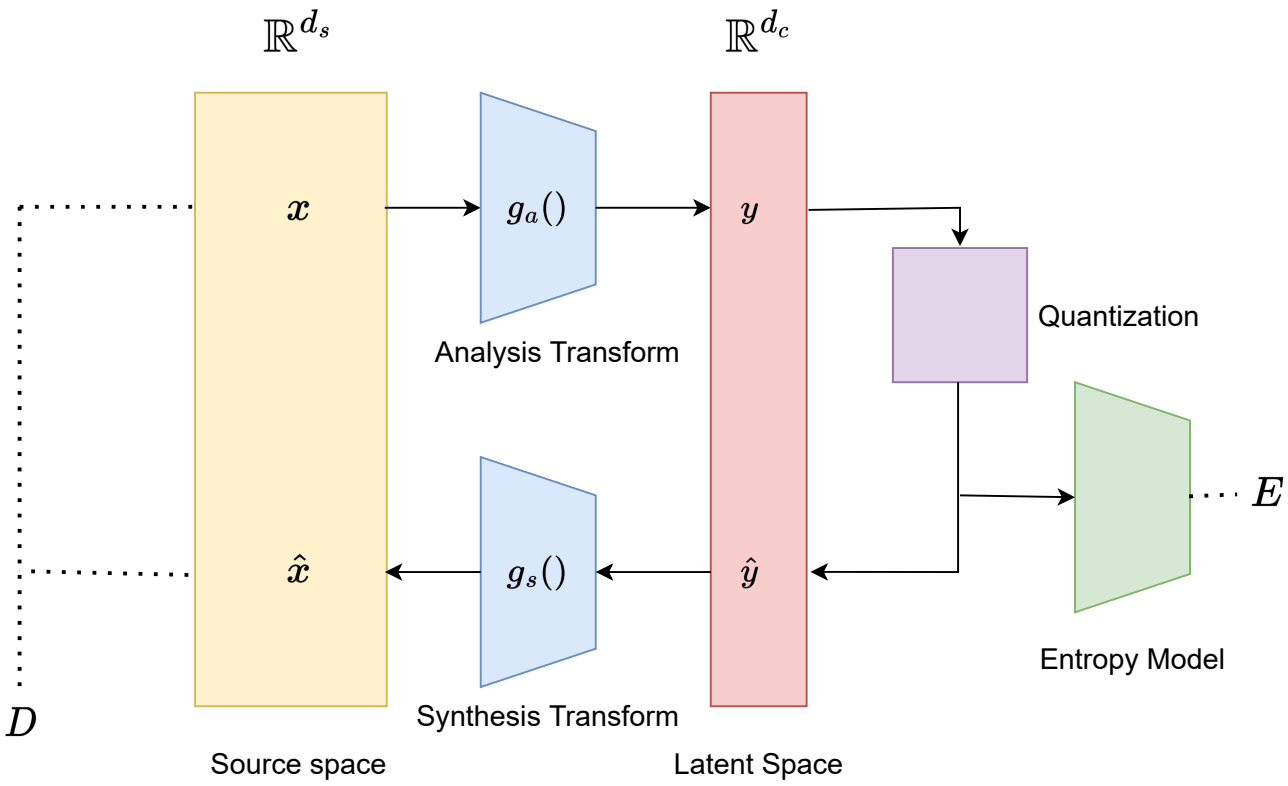


Figure 1.2: Nonlinear Transform Coding

if a random point $\boldsymbol{x}$ is picked from the data set then $g(f(\boldsymbol{x}))$ is close to $\boldsymbol{x}$ in some sense, for example in mean squared error. They have achieved striking successes in representation learning and dimensionality reduction when f and g are selected through training from the class of functions realized by multilayer perceptrons of a given architecture [33]. To compress a source of dimension d_s using a latent dimension of d_c, 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 a neural-network-based entropy model that is factorized across each dimension of the codeword. The encoder passes an input source vector $X \in \mathbb{R}^{d_s}$ to the analysis transform to obtain $g_a(X)$ which is rounded to its nearest integer (uniform scalar quantization) and represented by $\lfloor g_a(X) \rceil$. These quantized latents can then be entropy-coded using an ANN-based entropy model that learn a parametric probability density function P of the quantized latents. The decoder outputs the

reconstruction $g_s(\lfloor g_a(X) \rceil)$. The analysis transform, synthesis transform and the entropy model are trained on data by stochastically optimizing a rate(more precisely, entropy)-distortion Lagrangian that can be written as

$$\mathcal{L}(g_a, g_s, P) = \mathbb{E}_{X \sim \mathrm{P_{data}}}\Big[-\log\left(P(\lfloor g_a(X) \rceil)\right) + \lambda \left\| X - g_s(\lfloor g_s(X) \rceil) \right\|^2\Big].$$

The objective as written above is not amenable to stochastic optimization. The reason being that stochastic optimization relies on backpropagating gradients and the non-differentiable rounding operation nullifies any incoming gradients. Multiple workarounds for this issue have been suggested in literature such as using straight-through gradient estimates while training, dithered quantization, reverse channel coding-based annealing. The latter is closest to hard quantization by the end of the training process and is also used in most of our numerical experiments unless otherwise stated.

CHAPTER 2

OPTIMAL NEURAL-NETWORK-BASED COMPRESSION AND THE MANIFOLD HYPOTHESIS

We study a suite of sources that are uniformly distributed on compact Lie groups G. Previous work [69] considered a random process (the "sawbridge") over $[0, 1]$, i.e., $G = \mathbb{R}$, that can be constructed from a continuous, nonlinear transformation of a single random variable, thus exhibiting low-dimensional manifold structure in a high-dimensional ambient space. They found that stochastically-trained Artificial Neural Networks (ANNs) compress the process optimally. We consider sources that lie on circular manifolds, i.e, $G = S^1$ and $G = \mathbb{R} \times S^1$. We first 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 the stochastically-trained ANN-based compressors that succeeded at compressing the sawbridge fail to compress the ramp optimally at high-SNR. At first glance, 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 [51, 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 the stochastically-trained ANN-based compressors that were optimal on the sawbridge are suboptimal at high rates on the circle. We also consider a high-dimensional analogue of the circle, which we call the sinusoid process. We characterize the optimal entropy-distortion tradeoff for this source and like the circle, show that ANN-based compressors fail to achieve the optimal entropy-distortion tradeoff at high rates. Inspired from the literature on the neural tangent kernel (NTK) theory that attempts to explain the inductive bias of neural networks, we propose

a fix that involves embedding random Fourier features (RFFs). For the ramp and the circle, we demonstrate that this helps achieve the optimal-entropy distortion tradeoffs. Finally, we consider a source with latent dimension 2 only one of which is circular. We study the stationary sawbridge process which is the sawbridge shifted in time by a random phase variable in $[0, 1]$. Since the optimal entropy-distortion tradeoff for this source is complicated to derive, we resort to analytical high-rate bounds.

2.1 Preliminaries

Let $\mathcal{M}$ denote the ambient space of the source. 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. For a source X, we first define a transform and then an encoder and its entropy and distortion.

Definition 6. *A mapping $f : \mathcal{M} \mapsto \mathbb{R}^k$ is a* transform (of dimension k) *if it is continuous and there exists a continuous function $g : \mathbb{R}^k \mapsto \mathcal{M}$ (the* inverse transform*) such that*

$$g(f(X)) = X \quad a.s. \tag{2.1}$$

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

Definition 7. *An* encoder *is a (deterministic) mapping $f : \mathcal{M} \mapsto \mathbb{N}$. Its* entropy *and* distortion *for a given source X are given by*

$$H(f) = \sum_{i \in \mathbb{N}} -\Pr(f(X) = i) \log(\Pr(f(X) = i))$$
$$D(f) = \mathbb{E}\left[\left\|X - \mathbb{E}[X \mid f(X)]\right\|^2\right],$$

respectively, where throughout $\log(\cdot)$ *is base-2.*

Note that Definition 7 does not require that the reproduction be close to a valid source realization. That is, we do no impose the "realism" constraint that gives rise to the rate-distortion percep-

tion function (cf. [16]). 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 wishes to minimize. The minimum such expected length, $L^*(f)$, is known to satisfy $H(f) \leq L^*(f) \leq H(f)+1$ if one requires prefix-free encoding [24, Theorem 5.4.1] and

$$L^*(f) \leq H(f) \tag{2.2}$$

$$H(f) \leq L^*(f) + (1 + L^*(f)) \log(1 + L^*(f)) - L^*(f) \log L^*(f), \tag{2.3}$$

if one does not [61, Theorem 1]. As such, it is reasonable to focus on $H(f)$ as the figure-of-merit, especially at high rates. Practical ANN-based compressors aim to minimize $H(f)$, which further motivates this choice. 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 8. *The entropy-distortion function of X is*

$$H(D) = \inf_{f} H(f)$$
$$\text{s.t. } D(f) \leq D.$$

2.1.1 Sawbridge

Wagner and Ballé [69] proposed a manifold source, that they termed the "sawbridge" as it exposes the largest gap between the performance of linear and nonlinear dimensionality reduction

Definition 9 (cf. [26])**.** *The* sawbridge *is the process*

$$X(t) = t - \mathbf{1}(t \geq U) \quad t \in [0, 1], \tag{2.4}$$

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

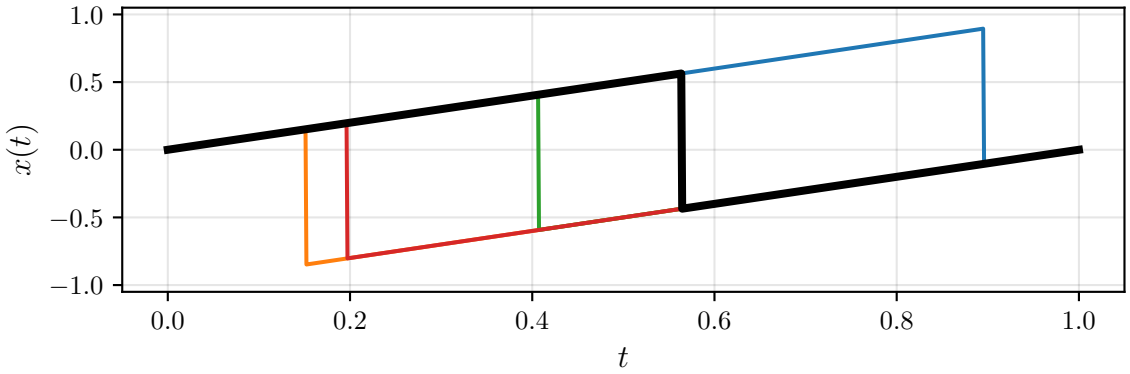


Figure 2.1: Realizations of the sawbridge. The bold line represents one full realization; others show additional samples.

The optimal entropy-distortion tradeoff for the sawbridge was derived in [69].

Theorem 10 ([69])**.** *If* $\Delta \geq 1/6$*, then* $H(\Delta) = 0$*. For any* $0 < \Delta < 1/6$*, we have*

$$H(\Delta) = -\left\lfloor \frac{1}{p} \right\rfloor \cdot p \log p - q \log q, \tag{2.5}$$

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*

$$\left\lfloor \frac{1}{p} \right\rfloor \cdot p^2 + q^2 = 6\Delta. \tag{2.6}$$

A plot of $H(D)$ is included in Fig. 2.2.

Corollaries to the entropy-distortion tradeoff characterization involve a characterization of the best fixed-rate encoders and a high-rate characterization.

Corollary 11 (Fixed-rate [69])**.** *Define an* M-encode *for the sawbridge as an encoder with the property that the support of* $f(X)$ *has cardinality* M *or less. The minimum distortion among all* M*-codes is* $\frac{1}{6M}$*, which is achieved by an encoder that quantizes* U *uniformly to* M *different values.*

Corollary 12 (High-rate [69])**.** *For the sawbridge,*

$$\lim_{\Delta \to 0} \left| H(\Delta) - \log \frac{1}{6\Delta} \right| = 0. \tag{2.7}$$

2.1.2 ANN-based compressors

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. We outline the approach summarized by Ballé et al. [8] using sawbridge as a working example throughout. This is also the approach that we follow to train ANN-based compressors for the sources we study in this chapter.

To represent a process digitally, we sample t at d_s equidistant points between 0 and 1; thus, each realization is represented as a d_s-dimensional vector. For the sawbridge, $d_s = 1024$ is set to be high enough that ANN-based compressors don't exploit it by memorizing realizations. The analysis and synthesis transforms are fully-connected feedforward neural 3 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 [4]. 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_s-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 ANNbased compressor using Adam. We sweep across different values of λ to obtain points on the lower convex hull of the ANN-compressor's entropy-distortion tradeoff.

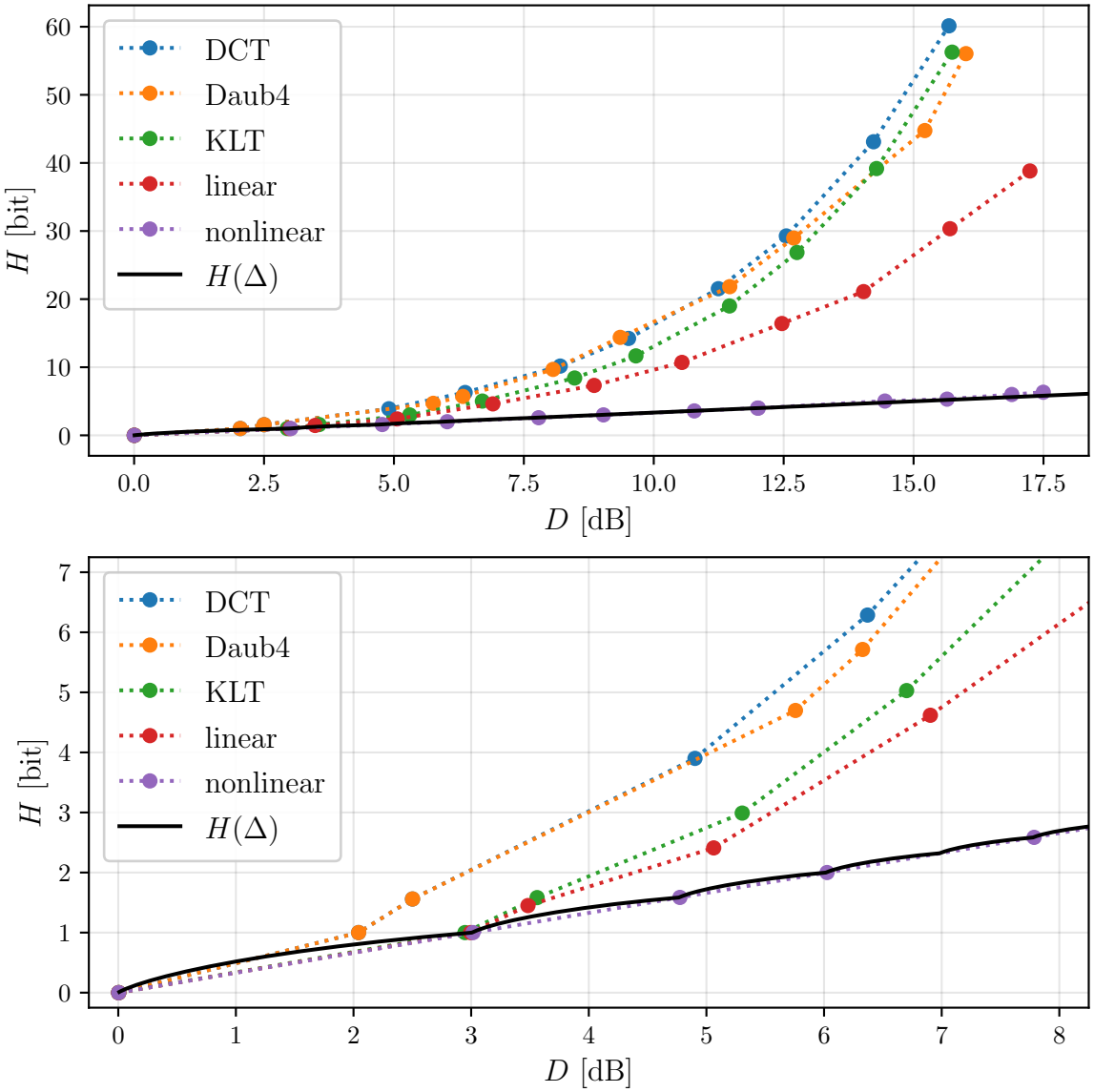


Figure 2.2: Empirical entropy–distortion plots for transform codes constrained to discrete cosine transform (DCT), Daubechies 4-tap wavelet (Daub4), Karhunen–Loève transform (KLT), arbitrary linear transforms, and nonlinear transforms implemented by ANNs. We also plot the entropy–distortion function of the source. The bottom panel shows the same data, zoomed in to the low-rate regime.

For the sawbridge, empirical results from [69] are plotted in Fig. 2.2. As can be seen from the figure, linear transform codes are exponentially suboptimal.

2.2 Optimal Entropy-Distortion Tradeoffs of Manifold Sources

We outline the sources that we will consider in this paper and derive their optimal entropy-distortion tradeoffs.

2.2.1 Circle

The *circle* is a 2-D source with a 1-D latent variable.

$$\theta \sim \text{Unif}\left[0, 2\pi\right), \overline{Z} = (\cos\theta, \sin\theta)\,. \tag{2.8}$$

Alternatively, it can be viewed as being uniformly distributed over the Lie group $G = S^1$ with an identity map.

The optimal entropy-distortion tradeoff is given by the following theorem.

Theorem 13. *For the circle, if $D \geq 1$, $H_c(D) = 0$. If $0 < D < 1$, then*

$$\begin{aligned} H_c(D) = \inf_{\{\theta_i\}_{i=1}^{\infty}} \quad & -\sum_{i}^{\infty} \frac{\theta_i}{2\pi} \log\left(\frac{\theta_i}{2\pi}\right) \\ s.t. \quad & \sum_{i=1}^{\infty} \frac{\theta_i}{2\pi}\left(1 - sinc^2\left(\frac{\theta_i}{2}\right)\right) \leq D, \\ & \sum_{i=1}^{\infty} \theta_i = 2\pi, \theta_i \geq 0 \text{ for all } i, \end{aligned}$$

where

$$sinc(x) = \begin{cases} \frac{\sin(x)}{x} & \text{if } x \neq 0 \\ 1 & \text{otherwise.} \end{cases} \tag{2.9}$$

Proof. Let $f : \mathbb{R}^2 \mapsto \mathbb{N}$ be an encoder. For the circle source, an encoder f can be alternatively represented as a function of the angle $\theta \in [0, 2\pi)$ as $h(\theta) = f([\cos\theta, \sin\theta])$. Let C denote the unit circle and define $C_i \overset{\text{def}}{=} C \cap f^{-1}(i)$ and $\Theta_i \overset{\text{def}}{=} [0, 2\pi) \cap h^{-1}(i)$. C_i is said to be contiguous if Θ_i is an interval or if it is of the form $[0, \theta_1] \cup [2\pi - \theta_2, 2\pi)$. Let $\mu(C_i)$ be the Lebesgue measure of C_i within the unit circle. The entropy and distortion are given by

$$\begin{aligned} H(f) &= \sum_{i=1}^{\infty} -\frac{\mu(C_i)}{2\pi} \log\left(\frac{\mu(C_i)}{2\pi}\right) \\ D(f) &= \sum_{i=1}^{\infty} \frac{\mu(C_i)}{2\pi} \mathbb{E}\left[\left\|\overline{Z} - \mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C_i\right]\right\|^2 \mid \overline{Z} \in C_i\right] \end{aligned}$$

respectively. Note that we can write

$$\begin{aligned} \mathbb{E}\left[\left\|\overline{Z} - \mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C\right]\right\|^2 \mid \overline{Z} \in C\right] &= \mathbb{E}\left[\left\|\overline{Z}\right\|^2 \mid \overline{Z} \in C\right] - \left\|\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C\right]\right\|^2 \\ &= 1 - \left\|\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C\right]\right\|^2. \end{aligned}$$

We first show that for a subset of the unit circle C,

$$\left\|\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C\right]\right\|^2 \leq \text{sinc}^2\left(\frac{\mu(C)}{2}\right). \tag{2.10}$$

First suppose that C is contiguous with corresponding $\Theta = [\theta_l, \theta_r]$ where $\theta_l < \theta_r$, without loss of

genarality. Then we have

$$\begin{aligned}
\left\|\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C\right]\right\|^2 &= (\mathbb{E}[\cos\theta \mid \theta \in [\theta_l, \theta_r]])^2 + (\mathbb{E}[\sin\theta \mid \theta \in [\theta_l, \theta_r]])^2 \\
&= \frac{1}{(\theta_r - \theta_l)^2}\Big((\sin\theta_r - \sin\theta_l)^2 + (\cos\theta_l - \cos\theta_r)^2\Big) \\
&= \frac{1}{(\theta_r - \theta_l)^2}(2 - 2\cos(\theta_r - \theta_l)) \\
&= \frac{4}{(\theta_r - \theta_l)^2}\sin^2\left(\frac{\theta_r - \theta_l}{2}\right) \\
&= \operatorname{sinc}^2\left(\frac{\theta_r - \theta_l}{2}\right) \\
&= \operatorname{sinc}^2\left(\frac{\mu(C)}{2}\right). \qquad (2.11)
\end{aligned}$$

More generally, suppose C is a finite, disjoint union of closed intervals $C_0, C_1, \cdots, C_{n-1}$. Due to rotational invariance, assume $\Theta_0 = [0, \theta]$ for some $\theta > 0$ and $\Theta_i = [\theta_{\ell_i}, \theta_{u_i}]$ where $\theta_{\ell_i} < \theta_{u_i}$. Denote the union of all closed intervals except the i^{th} as $C_{\not i}$. We have

$$\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C\right] = \frac{\mu(C_i)}{\mu(C_i) + \mu(C_{\not i})}\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C_i\right] + \frac{\mu(C_{\not i})}{\mu(C_i) + \mu(C_{\not i})}\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C_{\not i}\right]. \qquad (2.12)$$

and therefore,

$$\begin{aligned}
&\left\|\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C\right]\right\|^2 \\
&= \left(\frac{\mu(C_i)}{\mu(C_i) + \mu(C_{\not i})}\right)^2 \left\|\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C_i\right]\right\|^2 \qquad (2.13) \\
&\quad + \left(\frac{\mu(C_{\not i})}{\mu(C_i) + \mu(C_{\not i})}\right)^2 \left\|\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C_{\not i}\right]\right\|^2 \\
&\quad + 2\left(\frac{\mu(C_i)}{\mu(C_i) + \mu(C_{\not i})}\right)\left(\frac{\mu(C_{\not i})}{\mu(C_i) + \mu(C_{\not i})}\right)\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C_i\right]^\top \mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C_{\not i}\right].
\end{aligned}$$

We show that there exists an I such that rotating C_i so that it abuts one of its neighbors increases the norm in (2.13). Observe that rotating C_i affects $\left\|\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C\right]\right\|^2$ only through

$$\mathbb{E}\left[\overline{Z}|\overline{Z} \in C_i\right]^\top \mathbb{E}\left[\overline{Z}|\overline{Z} \in C_{\not i}\right]. \qquad (2.14)$$

Thus the norm can be increased by rotating C_i so long as $\mathbb{E}\left[\overline{Z}|\overline{Z} \in C_i\right]$ and $\mathbb{E}\left[\overline{Z}|\overline{Z} \in C_{i'}\right]$ are not aligned. But since we have

$$\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C\right] = \frac{\mu(C_i)}{\mu(C)} \cdot \mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C_i\right] + \frac{\mu(C_i)}{\mu(C_{i'})} \cdot \mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C_{i'}\right]. \tag{2.15}$$

and the left-hand side is independent of i, at most one i can be such that $\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C_i\right]$ and $\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C_{i'}\right]$ are aligned. It follows that we can increase the norm in (2.13) by rotating one of the C_i until it abuts with a neighboring interval. Repeating this procedure results in a contiguous C, for which (2.11) holds. This establishes (2.10) for C that are unions of disjoint intervals.

Finally, consider an arbitrary measurable C. For all $\epsilon > 0$, there exists a subset of the circle, C' that is a finite, disjoint union of closed intervals such that [28, Theorem 1.20]

$$E[|1(\overline{Z} \in C) - 1(\overline{Z} \in C')|] \leq \epsilon. \tag{2.16}$$

Then we have

$$\left\|\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C\right] - \mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C'\right]\right\|^2 = \left\|\mathbb{E}\left[\frac{\overline{Z}1\left(\overline{Z} \in C\right)}{\mu(C)} - \frac{\overline{Z}1\left(\overline{Z} \in C'\right)}{\mu(C')}\right]\right\|^2 \tag{2.17}$$

$$\leq \mathbb{E}\left[\left(\frac{1(Z \in C)}{\mu(C)} - \frac{1(Z \in C')}{\mu(C')}\right)^2\right] \tag{2.18}$$

$$\leq \frac{\epsilon}{\min^2(\mu(C), \mu(C'))} + \frac{(\mu(C) - \mu(C'))^2}{\mu(C)^2\mu(C)^2} \tag{2.19}$$

$$\leq \frac{\epsilon(1+\epsilon)}{(\mu(C) - \epsilon)^2}. \tag{2.20}$$

Thus, since C' satisfies the triangle inequality,

$$\left\|\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C\right]\right\| \leq \left\|\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C'\right]\right\| + \left\|\mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C\right] - \mathbb{E}\left[\overline{Z} \mid \overline{Z} \in C'\right]\right\| \tag{2.21}$$

$$\leq \left|\operatorname{sinc}\left(\frac{\mu(C')}{2}\right)\right| + \frac{\sqrt{\epsilon(1-\epsilon)}}{(\mu(C) - \epsilon)^2} \tag{2.22}$$

$$\leq \left|\operatorname{sinc}\left(\frac{\mu(C) - \epsilon}{2}\right)\right| + \frac{\sqrt{\epsilon(1-\epsilon)}}{(\mu(C) - \epsilon)^2}. \tag{2.23}$$

Since $\epsilon > 0$ was arbitrary, (2.10) and the theorem follows.

□

Theorem 14.

$$H_c(D) \geq \sup_{\lambda \geq 0} \inf_{0<\theta<2\pi} -\log\left(\frac{\theta}{2\pi}\right) + \lambda\left(1 - sinc^2\left(\frac{\theta}{2}\right)\right) - \lambda D. \tag{2.24}$$

Proof. By weak duality,

$$\begin{aligned} H_{\mathrm{c}}(D) &\geq \sup_{\lambda \geq 0} \inf_{\substack{\{\theta_i\}_{i=1}^{\infty}: \\ \sum_i \theta_i = 2\pi, \\ \theta_i \geq 0 \text{ for all } i}} \sum_i^{\infty} -\frac{\theta_i}{2\pi} \log\left(\frac{\theta_i}{2\pi}\right) + \lambda\left(\sum_{i=1}^{\infty} \frac{\theta_i}{2\pi}\left(1 - \mathrm{sinc}^2\left(\frac{\theta_i}{2}\right)\right) - D\right) \\ &= \sup_{\lambda \geq 0} \inf_{\substack{\{\theta_i\}_{i=1}^{\infty}: \\ \sum_i \theta_i = 2\pi, \\ \theta_i \geq 0 \text{ for all } i}} \sum_i^{\infty} \frac{\theta_i}{2\pi}\left[-\log\left(\frac{\theta_i}{2\pi}\right) + \lambda\left(1 - \mathrm{sinc}^2\left(\frac{\theta_i}{2}\right)\right) - \lambda D\right] \\ &\geq \sup_{\lambda \geq 0} \inf_{0<\theta<2\pi} -\log\left(\frac{\theta}{2\pi}\right) + \lambda\left(1 - \mathrm{sinc}^2\left(\frac{\theta}{2}\right)\right) - \lambda D. \end{aligned}$$

□

The lower bound in Theorem 14 is illustrated in Fig. 2.3 ("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. 2.3). Note that these bounds essentially coincide at high-SNR.

2.2.2 Ramp

$$X_t \stackrel{\text{def}}{=} [(t + V) \mod 1] - \frac{1}{2}, \tag{2.25}$$

where $V \sim \mathrm{Unif}[0, 1]$. 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. In other words, the source is supported on $G = S^1$. This process is similar in some important respects to the sawbridge process. Both are continuous functions in the interval $[0, 1]$ that rise with slopes of 1 with a single drop. In fact, the ramp is obtained by subtracting from the sawbridge its DC.

Since realizations of the ramp process are in one-to-one correspondence with realizations of V, it is natural to compress the ramp process by quantizing the V. The next theorem confirms that this is indeed optimal.

Theorem 15. *For the ramp, if $D \geq \frac{1}{12}$, $H_r(D) = 0$. If $0 < D < \frac{1}{12}$, then*

$$H_r(D) = \inf_{\{p_i\}_{i=1}^{\infty}} \quad -\sum_{i=1}^{\infty} p_i \log p_i \tag{2.26a}$$

$$s.t. \quad \sum_{i=1}^{\infty} \frac{p_i^2(2-p_i)}{12} \leq D, \tag{2.26b}$$

$$\sum_{i=1}^{\infty} p_i = 1, p_i \geq 0 \text{ for all } i. \tag{2.26c}$$

Proof. Let $f_{\text{ramp}} : L^2[0,1] \mapsto \mathbb{N}$ be an encoder. Each value of the phase variable defines a unique realization. Therefore, define $S_i = f_{\text{ramp}}^{-1}(i)$ where

$$f_{\text{ramp}}^{-1}(i) \stackrel{\text{def}}{=} \left\{ v \in [0,1] : f_{\text{ramp}}([(t+v) \bmod 1] - 0.5) = i \right\}. \tag{2.27}$$

Let μ be the Lebesgue measure. The optimal reconstruction is $t \mapsto \mathbb{E}\left[X_t \mid f_{\text{ramp}}(X_t)\right]$. Thus the entropy and distortion of the encoder-decoder pair is given by

$$H\left(f_{\text{ramp}}\right) = -\sum_i \mu(S_i) \log\left(\mu(S_i)\right),$$

$$D\left(f_{\text{ramp}}\right) = \mathbb{E}\left[\int_0^1 \left(X_t - \mathbb{E}\left[X_t \mid f_{\text{ramp}}(X_t)\right]\right)^2 dt\right].$$

The distortion can be expressed as

$$D\left(f_{\text{ramp}}\right) = \mathbb{E}\left[\int_0^1 \left(X_t - \mathbb{E}\left[X_t \mid f_{\text{ramp}}(X_t)\right]\right)^2 dt\right] \tag{2.28}$$

$$= \sum_i \mu(S_i)\, \mathbb{E}\left[\int_0^1 (X_t - \mathbb{E}[X_t \mid V \in S_i])^2\, dt \middle| V \in S_i\right] \tag{2.29}$$

$$= \sum_i \mu(S_i)\left(\mathbb{E}\left[\int_0^1 X_t^2 dt \mid V \in S_i\right] - \int_0^1 (\mathbb{E}[X_t | V \in S_i])^2\ dt\right) \tag{2.30}$$

$$= \sum_i \mu(S_i)\left(\frac{1}{12} - \int_0^1 (\mathbb{E}[X_t \mid V \in S_i])^2\, dt\right). \tag{2.31}$$

where the last line follows because, irrespective of the phase, we have $\int_0^1 X_t^2\ dt = \frac{1}{12}$.

We will show that, for any measurable $S \subset [0,1]$, we have

$$\int_0^1 (\mathbb{E}[X_t \mid V \in S])^2\, dt \le \frac{1}{3}\left(\frac{1-\mu(S)}{2}\right)^2. \tag{2.32}$$

Write $y_t = E[X_t | V \in S]$. Then we have

$$\mathbb{E}\left[(t+V \mod 1) - \frac{1}{2}\middle| V \in S\right] \tag{2.33}$$

$$= \mathbb{E}\left[\left(t+V-\frac{3}{2}\right)1(t+V>1) + \left(t+V-\frac{1}{2}\right)1(t+V\le 1)\middle| V \in S\right]$$

$$= t + \mathbb{E}[V|V\in S] - \mathbb{P}(t+V>1 \mid V\in S) - \frac{1}{2}$$

$$= t + \mathbb{E}[V|V\in S] - \frac{\mu(S\cap[1-t,1])}{\mu(S)} - \frac{1}{2}. \tag{2.34}$$

From (2.33) we see that the left-hand side of (2.32) is invariant with respect to cyclic rotation of S of the form $S \mapsto s + S \mod 1$. Evidently the right-hand side of (2.32) is also invariant with respect to such a shift. Thus to show (2.32) we may cyclically rotate S as convenient.

First, suppose that S can be cyclicly rotated to obtain an interval of the form $[0, s]$. In this case

(2.34) reads

$$y_t = \begin{cases} t\left(1 - \frac{1}{s}\right) + \frac{s}{2} - \frac{3}{2} + \frac{1}{s} & \text{if } t > 1 - s \\ t - \frac{1-\mu(S)}{2} & \text{if } t \leq 1 - s, \end{cases} \tag{2.35}$$

and thus we can explicitly compute

$$\int_0^1 y_t^2 dt = \frac{1}{3}\left(\frac{1-\mu(S)}{2}\right)^2. \tag{2.36}$$

and equality in (2.32) holds in this case. More generally, suppose that S is a finite union of closed, disjoint intervals, $S_0, \ldots, S_{n-1}$ such that all points in S_0 are less than those in S_1, etc. Due to the rotational invariance of (2.32), we can assume that $\min S_0 = 0$ and $\max S_{n-1} < 1$. Define

$$s_i = \min S_i \quad i = 0, \ldots, n-1 \tag{2.37}$$

$$s_n = 1 \tag{2.38}$$

$$t_i = s - s_{n-i} \quad i = 0, \ldots, n. \tag{2.39}$$

Note that for each i, from (2.34), over the interval $[t_i, t_{i+1}]$, y_t is linearly increasing and then decreasing. Thus for $t \in [t_i, t_{i+1}]$,

$$y_t \leq y_{t_i} + \frac{t - t_i}{t_{i+1} - t_i}(y_{t_{i+1}} - y_{t_i}) \tag{2.40}$$

and as a result

$$\int_{t_i}^{t_{i+1}} y_t \, dt \leq \frac{y_{t_i} + y_{t_{i+1}}}{2}[t_{i+1} - t_i]. \tag{2.41}$$

Since every realization of J_t integrates to zero, we have

$$0 = \mathbb{E}\left[\int_0^1 J_t dt \middle| V \in S\right] = \int_0^1 \mathbb{E}[J_t | V \in S] dt = \int_0^1 y_t dt = \sum_{i=0}^{n-1} \int_{s_i}^{s_{i+1}} y_t dt \leq \sum_{i=1}^{n} \frac{y_{t_i} + y_{t_{i+1}}}{2} [t_{i+1} - t_i]. \tag{2.42}$$

It follows that there exists i such that $y_{t_i} + y_{t_{i+1}} > 0$. By rotational invariance, we can assume that $i = 0$. Consider the modified set $\tilde{S} = \tilde{S}_0 \cup \cdots \cup \tilde{S}_{n-1}$ wherein $\tilde{S}_i = S_i$ for $i = 0, \ldots, n-2$ and

$\tilde{S}_{n-1}$ is a closed interval satisfying $\mu(\tilde{S}_{n-1}) = \mu(S_{n-1})$ and $\max \tilde{S}_{n-1} = 1$. In words, we shift S_{n-1} to the right so that its right end-point is 1. Define

$$\tilde{y}_t = \mathbb{E}[J_t | V \in \tilde{S}] \tag{2.43}$$

and note that, from (2.34), we have

$$\tilde{y}_t = \begin{cases} y_0 + y_{t_1} - y_{t_1 - t} & \text{if } t \in [0, t_1] \\ y_t & \text{if } t \in (t_1, 1]. \end{cases} \tag{2.44}$$

From this it follows that the norm of $\tilde{y}_t$ dominates that of y_t:

$$\int_0^1 \tilde{y}_t^2 dt = \int_0^{t_1} (y_0 + y_{t_1} - y_{t_1 - t})^2 dt + \int_{t_1}^1 y_t^2 dt \tag{2.45}$$

$$= \int_0^1 y_t^2 dt + (y_0 + y_{t_1})^2 t_1 - 2(y_0 + y_{t_1}) \int_0^{t_1} y_t dt \tag{2.46}$$

$$\geq \int_0^1 y_t^2 dt, \tag{2.47}$$

where the final equality follows from (2.40) and the assumption that $y_{t_i} + y_{t_{i+1}} > 0$. Now $\tilde{S}$ can be written as a union of $n - 1$ disjoint closed intervals, since $\tilde{S}_0$ and $\tilde{S}_{n-1}$ are effectively conjoined. Repeating this process until we arrive at a single interval and then applying (2.36) gives

$$\int_0^1 \mathbb{E}[J_t | V \in S]^2 dt \leq \frac{1}{3} \left(\frac{1 - \mu(S)}{2} \right)^2 \tag{2.48}$$

for any S that is a disjoint union of closed intervals. Finally, consider an arbitrary (measurable) $S \subset [0, 1]$. Then for any $\epsilon > 0$, there exists a $\tilde{S} \subset [0, 1]$ that is finite union of disjoint closed intervals such that

$$\mu\Big((S \setminus \tilde{S}) \cup (\tilde{S} \setminus S) \Big) \tag{2.49}$$

can be made arbitrarily small [28, Theorem 1.20], and in particular for any ϵ we can have

$$\left| \frac{\mathbb{E}[V \cdot 1(V \in S)]}{\mathbb{P}(V \in S)} - \frac{\mathbb{E}[V \cdot 1(V \in \tilde{S})]}{\mathbb{P}(V \in \tilde{S})} \right| \leq \epsilon \tag{2.50}$$

$$\left| \frac{\mu(S \cap [1 - t, 1])}{\mu(S)} - \frac{\mu(\tilde{S} \cap [1 - t, 1])}{\mu(\tilde{S})} \right| \leq \epsilon. \tag{2.51}$$

Define the two conditional means

$$y_t = \mathbb{E}[J_t | V \in S]$$

$$\tilde{y}_t = \mathbb{E}[J_t | V \in \tilde{S}].$$

Then from (2.50) and (2.51) we have, for each t, which, since

$$\int_0^1 \tilde{y}_t dt = 0. \tag{2.52}$$

implies that

$$\int_0^1 y_t^2 dt \leq \frac{1}{3}\left(\frac{1-\mu(\tilde{S})}{2}\right)^2 + 4\epsilon^2 \tag{2.53}$$

$$\leq \frac{1}{3}\left(\frac{1-\mu(S)+\epsilon}{2}\right)^2 + 4\epsilon^2, \tag{2.54}$$

from (2.51). Taking $\epsilon \to 0$ establishes (2.32).

Now given any code, f_{ramp}, for the ramp, define S_i as in (2.27). Then we have from (2.31)

$$D\left(f_{\text{ramp}}\right) = \sum_i \mu(S_i)\left(\frac{1}{12} - \int_0^1 \mathbb{E}[J_t | V \in S_i]^2 dt\right) \tag{2.55}$$

$$\geq \sum_i \mu(S_i)\left(\frac{1}{12} - \frac{1}{3}\left(\frac{1-\mu(S_i)}{2}\right)^2\right) \tag{2.56}$$

$$= \sum_i \mu(S_i)\left(\frac{1}{6}\mu(S_i) - \frac{1}{12}\mu^2(S_i)\right). \tag{2.57}$$

Setting $p_i = \mu(S_i)$ establishes the converse of the theorem. Achievability follows by noting that, for any $\{p_i\}$ satisfying the constraint in (2.26b), we can create a code with entropy $-\sum_i p_i \log p_i$ satisfying the distortion constraint by setting $S_0, S_1, \ldots, S_{n-1}$ to be intervals with $\mu(S_i) = p_i$ and noting equality in (2.32). □

The entropy-distortion function in (15) is formally an infinite-dimensional optimization problem. The following provides two lower bounds, one of which is closed form and the other of which is easily computed.

Corollary 16.

$$H_r(D) \geq \sup_{\lambda \geq 0} \inf_{0<p<1/12} -\log(p) - \lambda\left(D - \frac{p(2-p)}{12}\right) \tag{2.58}$$

$$\geq \log\left(\frac{2-\sqrt{12D}}{12D}\right). \tag{2.59}$$

Proof. Note that from (2.26b), we have that for all i,

$$D \geq \frac{p_i^2(2-p_i)}{12} \geq \frac{p_i^2}{12}. \tag{2.60}$$

Thus we can restrict attention to p_i such that $p_i \leq \sqrt{12D}$. Then by weak duality,

$$\begin{aligned} H_{\mathrm{r}}(D) &\geq \sup_{\lambda \geq 0} \inf_{\substack{\{p_i\}_{i=1}^{\infty}: \\ \sum_i p_i = 1, \\ 0 \leq p_i \leq 1/12 \text{ for all } i}} \sum_{i=1}^{\infty} -p_i \log(p_i) - \lambda\left(D - \sum_{i=1}^{\infty} \frac{p_i^2(2-p_i)}{12}\right) \\ &= \sup_{\lambda \geq 0} \inf_{\substack{\{p_i\}_{i=1}^{\infty}: \\ \sum_i p_i = 1, \\ 0 \leq p_i \leq 1/12 \text{ for all } i}} \sum_{i=1}^{\infty} p_i\left[-\log(p_i) - \lambda\left(D - \frac{p_i(2-p_i)}{12}\right)\right] \\ &\geq \sup_{\lambda \geq 0} \inf_{0<p<1/12} -\log(p) - \lambda\left(D - \frac{p(2-p)}{12}\right), \end{aligned}$$

which establishes (2.58). This can be weakened to

$$H_{\mathrm{r}}(D) \geq \sup_{\lambda \geq 0} \inf_{0<p<1/12} -\log(p) - \lambda\left(\Delta - \frac{p(2-\sqrt{12D})}{12}\right). \tag{2.61}$$

Choosing $\lambda = (\log e)/D$, we have that the infimum over p is achieved by $p = 12D/(2-\sqrt{12D}) < \sqrt{12D}$. Substituting this choice of λ and p into (2.61) yields (2.59). □

The bound in Theorem 16 is illustrated in Fig. 2.7 ("lower bound"). For rates equal to $\log K$ where $K \geq 2$ is an integer, an achievable scheme uniformly quantizes the phase. The encoder assigns a ramp realization J_t to the midpoint of the interval that contains $J_0 + 0.5$. The decoder outputs the conditional mean of the ramp restricted to the phase lying in the encoded interval. For

rates between $\log K$ and $\log K + 1$, an upper bound on the optimal tradeoff is obtained by dividing the ramp into biuniform intervals, where one interval has length ε and the other K intervals have length $\frac{1-\varepsilon}{K}$.

2.2.3 Sinusoid

The *sinusoid*, like the ramp, is a model of a high-dimensional process with a 1-D circular latent variable. Unlike the ramp however, a given realization is continuous in time. The sinusoid can be viewed as a high-dimensional analogue of the circle and therefore, $G = S^1$. For $\phi \in [0, 2\pi]$, let S_ϕ denote the function

$$S_\Phi(t) = \sqrt{2}\sin(2\pi t + \Phi) \quad t \in [0, 1]. \tag{2.62}$$

Consider the source process $Y_t = S_U(t)$, where U is Unif[0, 1]. Let $H_{\sin}(\cdot)$ denote the entropy-distortion function of this source.

Theorem 17.

$$H_{sin}(D) = H_c(D).$$

Proof. We show the equivalence between the sinusoid and circle at an operational level. Note that we can deterministically couple the sources via $Y_t = S_{\angle\overline{Z}}(t)$. Let $f_{\sin} : L^2[0, 1] \mapsto \mathbb{N}$ be an encoder for the sinusoid. Then

$$f_{\text{circ}}(\overline{Z}) \stackrel{\text{def}}{=} f_{\sin}(S_{\angle\overline{Z}}) \tag{2.63}$$

is an encoder for the circle. Evidently we have $H(f_{\text{circ}}) = H(f_{\sin})$. We can similarly equate the distortions. We have

$$D(f_{\text{circ}}) = \sum_{i\in\mathbb{N}} \mathbb{E}\left[|\overline{Z} - \mathbb{E}\left[\overline{Z}|f_{\text{circ}}(\overline{Z}) = i\right]|^2 | f_{\text{circ}}(\overline{Z}) = i\right] \Pr(f_{\text{circ}}(\overline{Z}) = i) \tag{2.64}$$

$$= 1 - \sum_{i\in\mathbb{N}} |\mathbb{E}\left[\overline{Z}|f_{\text{circ}}(\overline{Z}) = i\right]|^2 \Pr(f_{\text{circ}}(\overline{Z}) = i) \tag{2.65}$$

and similarly

$$D(f_{\sin}) = 1 - \sum_{i\in\mathbb{N}} \int_0^1 |\mathbb{E}\left[Y(t)|f_{\sin}(Y) = i\right]|^2 \Pr(f_{\sin}(Y) = i) \tag{2.66}$$

Now from standard trigonometric considerations, we have that if

$$\mathbb{E}\left[\overline{Z}|f_{\text{circ}}(\overline{Z}) = i\right] = ae^{j\phi} \tag{2.67}$$

then

$$\mathbb{E}\left[Y(t)|f_{\sin}(Y) = i\right] = a\sqrt{2}\sin(2\pi t + \phi) \tag{2.68}$$

and thus

$$|E[\overline{Z}|f_{\text{circ}}(\overline{Z}) = i]|^2 = \int_0^1 \mathbb{E}^2\left[Y|f_{\sin}(Y) = i\right] dt. \tag{2.69}$$

It follows that $D(f_{\text{circ}}) = D(f_{\sin})$. Similarly, given an encoder for the circle, f_{circ}, one can create an encoder for the sinusoid, $f_{\sin}$, such that $H(f_{\sin}) = H(f_{\text{circ}})$ and $D(f_{\sin}) = D(f_{\text{circ}})$. The conclusion follows. □

2.2.4 Stationary Sawbridge

In Section 2.1.1, we saw that to provide the sawbridge as an input to ANN-based compressors, it has to be discretized which therefore forces the number of realizations to be finite. This training issue can be mitigated by rotating the sawbridge with a random *phase*. The *stationary sawbridge* is the process

$$Y_t \overset{\text{def}}{=} X_{(t+V)\bmod 1} \tag{2.70}$$

$$= t + V \bmod 1 - \mathbf{1}(t + V \bmod 1 \geq U), \tag{2.71}$$

for $t \in [0, 1]$, where U and V are i.i.d. Unif$[0, 1]$. We denote the entire process $\{Y_t\}_{t=0}^1$ as Y. The variables U and V are termed the *drop* and the *phase*, respectively, of the stationary sawbridge,

respectively. An alternative parametrization of the drop is the average value or DC, $\int_0^1 Y_t dt = U - \frac{1}{2} \stackrel{\text{def}}{=} U_{DC}$. Since the stationary sawbridge is just a rotation of the sawbridge in time, the DC value and the variance is energy are unaffected by the realization of V.

The stationary sawbridge can be rewritten as the ramp process with an added DC. To do so first reparametrize the sawbridge as follows

$$\begin{aligned} X_t &= t - \mathbf{1}(t \geq U) \quad t \in [0,1] \\ &= (t-U) \bmod 1 + U - 1. \end{aligned}$$

Therefore,

$$\begin{aligned} Y_t &= (t + V - U) \bmod 1 + U - 1 \\ &= \underbrace{(t + (V-U) \bmod 1) \bmod 1 - \frac{1}{2}}_{\text{ramp}} + U_{DC}. \end{aligned} \tag{2.72}$$

$$\tag{2.73}$$

It can be shown that the random variables $(V - U) \bmod 1$ and U_{DC} are independent.

The stationary sawbridge has a 2-D latent space and therefore deriving the exact optimal entropy-distortion tradeoff, $H_{\text{stat}}(D)$, is challenging. We following bounds are within .255 bits at all rates, which provides for a tight characterization in the high-rate regime. See [13] for a low-rate (one-bit) analysis.

Theorem 18. *For any* $0 < D < 1/12$*,*

$$\frac{3}{2}\log\left(\frac{1}{D}\right) - \frac{1}{2}\log\left(\frac{64\pi e}{3(2-\sqrt{12D})}\right) \leq H_{stat}(D) \leq \frac{3}{2}\log\left(\frac{1}{D}\right) - 3. \tag{2.74}$$

Proof. We will use the reparametrization of the stationary sawbridge as given in (2.72). Let $V' = (V - U) \bmod 1$ and $X'_t \stackrel{\text{def}}{=} (t + V') \bmod 1$. For the upper bound consider the uniform quantizer

$$f(Y) = \left(Q\left(2^{R_U} U_{DC}\right), Q\left(2^{R_V} V'\right)\right)$$

where R_U, R_V are real numbers and Q is a function that rounds a real number to its nearest half-integer. Therefore, the entropy $H(f)$ is upper bounded by $R_U + R_V$ and the distortion is

$$\begin{aligned} D(f) &= \mathbb{E}\left[\int_0^1 (Y_t - \mathbb{E}[Y_t \mid f(Y)])^2 dt\right] \\ &= \mathbb{E}\left[\int_0^1 (X'_t - \mathbb{E}[X'_t \mid f(Y)] + U_{DC} - \mathbb{E}[U_{DC} \mid f(Y)])^2\right] \\ &= \mathbb{E}\left[\int_0^1 (X'_t - \mathbb{E}[X'_t \mid f(Y)])^2 dt\right] + \mathbb{E}\Big[(U_{DC} - \mathbb{E}[U_{DC} \mid f(Y)])^2\Big]. \end{aligned}$$

We bound the two quantization errors separately:

$$\begin{aligned} \mathbb{E}\Big[(U_{DC} - \mathbb{E}[U_{DC} \mid f(Y)])^2\Big] &\le \frac{2^{-2R_U}}{12}, \\ \mathbb{E}\left[\int_0^1 (X'_t - \mathbb{E}[X'_t \mid f(Y)])^2 dt\right] &\le \frac{2^{-R_V}}{6}\left(1 - \frac{2^{-R_V}}{2}\right), \end{aligned}$$

where the final inequality follows from (2.57). Then,

$$\begin{aligned} H_{\text{stat}}(D) &\le \min R_U + R_V \\ &\qquad \text{s.t. } \frac{2^{-2R_U}}{12} + \frac{2^{-R_V}}{6}\left(1 - \frac{2^{-R_V}}{2}\right) \le D \\ &\le \min R_U + R_V \\ &\qquad \text{s.t. } \frac{2^{-2R_U}}{12} + \frac{2^{-R_V}}{6} \le D. \end{aligned}$$

Solving the above gives us

$$H_{\text{stat}}(D) \le \frac{3}{2}\log\left(\frac{1}{D}\right) - 3$$

For the lower bound, note that for any encoder f,

$$H(f) = I\left(U_{DC}, V'; f(Y)\right) = I(U_{DC}; f(Y)) + I(V'; f(Y) \mid U_{DC}).$$

Define

$$H_u(D_u) \stackrel{\text{def}}{=} \min_f I(U_{DC}; f(Y)) \tag{2.75}$$

$$\text{s.t. } \mathbb{E}\Big[(U_{DC} - \mathbb{E}[U_{DC} \mid f(Y)])^2\Big] \le D_u, \tag{2.76}$$

and

$$H_v(D_v) \stackrel{\text{def}}{=} \min_f I(V'; f(Y) \mid U_{DC}) \tag{2.77}$$

$$\text{s.t. } \mathbb{E}\left[\int_0^1 (X'_t - \mathbb{E}[X'_t \mid f(Y)])^2 dt\right] \le D_v. \tag{2.78}$$

Then we have

$$H_{\text{stat}}(D) \ge \min H_u(D_u) + H_v(D_v) \tag{2.79}$$

$$\text{s.t. } D_u + D_v \le D. \tag{2.80}$$

We bound each of the above terms separately.

$$\begin{aligned} H_u(D_u) &\ge \min_{U'} h(U_{DC}) - h(U_{DC} \mid U') \\ &\qquad \text{s.t. } \mathbb{E}\left[(U_{DC} - U')^2\right] \le D_u \\ &\ge -\frac{1}{2}\log(2\pi e D_u). \end{aligned} \tag{2.81}$$

For the ramp component, $I(V'; f(Y) \mid U_{DC}) = H(f(Y) \mid U_{DC})$:

$$\begin{aligned} H_v &\ge \min_{f'} H(f'(V')|U_{DC}) \\ &\qquad \text{s.t. } \mathbb{E}\left[\int_0^1 (X'_t - \mathbb{E}[X'_t \mid f'(V'), U_{DC}])^2 dt\right] \le D_v. \end{aligned}$$

If we define

$$\delta_u = \mathbb{E}\left[\int_0^1 (X'_t - \mathbb{E}[X'_t \mid f'(V'), U_{DC} = u])^2 dt \mid U_{DC} = u\right], \tag{2.82}$$

Then from Corollary 16, we have

$$H(f'(V')|U_{DC} = u) \ge \frac{1}{2}\log\left(\frac{2 - \sqrt{12\delta_u}}{12\delta_u}\right). \tag{2.83}$$

Since $\log((2 - \sqrt{12D})/(12D))$ is convex over $(0, 1/12)$, this implies

$$H(f'(V')|U_{DC} = u) \ge \log\left(\frac{2 - \sqrt{12D_v}}{12D_v}\right) \tag{2.84}$$

$$\ge \log\left(\frac{2 - \sqrt{12D}}{12D}\right). \tag{2.85}$$

Substituting (2.81) and (2.85) into (2.79) gives

$$H_{\text{stat}}(D) \geq \min_{D_v, D_u} -\frac{1}{2}\log(2\pi e D_u) + \log\left(\frac{2-\sqrt{12D}}{12D_v}\right) \tag{2.86}$$

$$\text{s.t. } D_u + D_v \leq D, \tag{2.87}$$

which is solved by the choices $D_u = D/3$ and $D_v = 2D/3$, yielding

$$H_{\text{stat}}(D) \geq \frac{3}{2}\log\frac{1}{D} - \frac{1}{2}\log\left(\frac{64\pi e}{2-\sqrt{12D}}\right). \tag{2.88}$$

□

2.3 ANN Performance

In this section, we analyze the performance of ANN-based compressors on sources whose optimal entropy-distortion tradeoffs we derived in the previous section.

2.3.1 Circle

Since the circle is described by the scalar random variable θ, we take the latent dimension $d_c = 1$. In Fig. 2.3, the curve labeled "1-D latent" shows the performance of the ANN-based compressor as described in Section 2.1.2. We see that this performance is suboptimal at high-SNR.

We now highlight the fundamental difficulty for the suboptimal performance of ANN-based compressors on the circle source. Intuition suggests, and the proofs of Theorems 13 and 14 essentially confirm, that an optimal scheme for compressing the circle is to quantize the angle θ into contiguous cells, all but (at most) one of which are the same size. Thus we would like the analysis

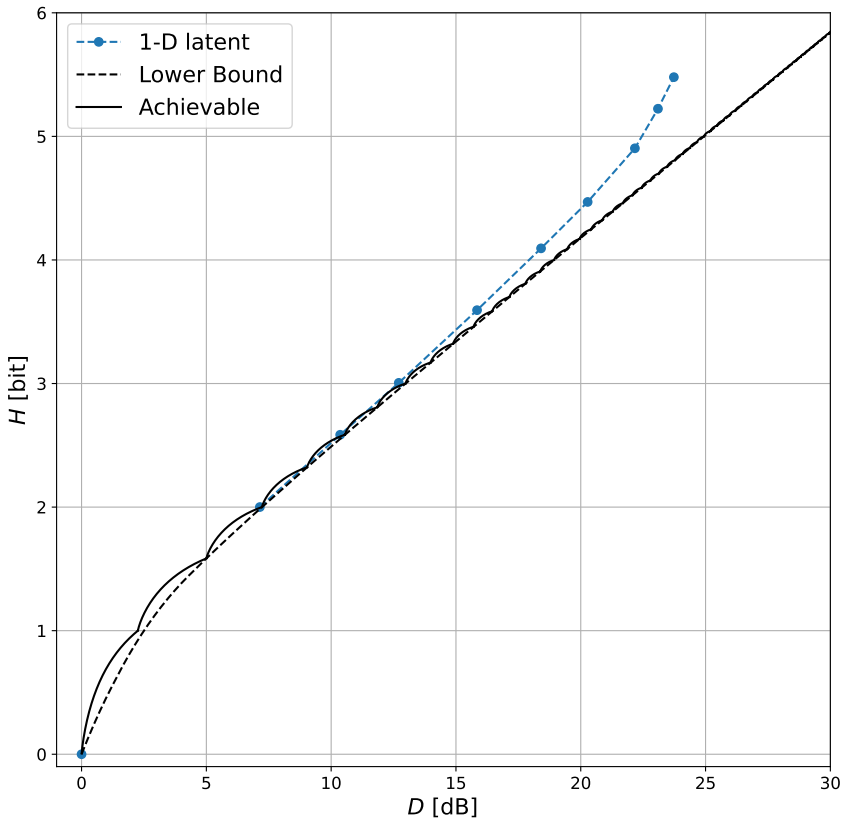


Figure 2.3: Circle entropy-distortion tradeoff of existing off-the-shelf ANN-based compressors with latent dimension 1 along with lower bound and upper bound on optimal tradeoff.

transform to extract the angle θ from the realization of the circle, i.e., to implement atan2($\overline{Z}$) or some scaled and shifted version thereof.

While the function $\theta \mapsto \overline{Z} = (\cos(2\pi\theta), \sin(2\pi\theta))$ is continuous, the issue is that the inverse function $\overline{Z} \mapsto \theta$ is not continuous over the circle since the unit circle and the line segment $[0, 1)$ are not homeomorphic [51, Section 18, Ex. 6]. However the analysis transform, being a composition of alternating linear and nonlinear continuous and differentiable functions, must be continuous (and indeed differentiable).

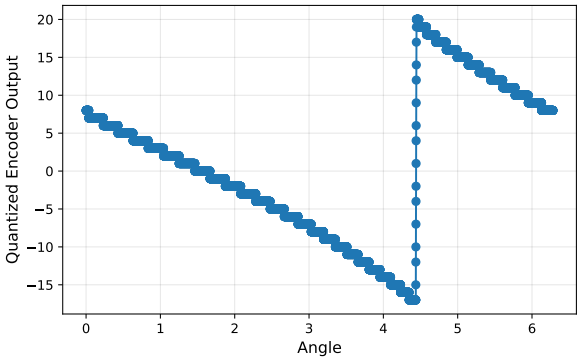


Figure 2.4: Quantized encoder output vs. angle θ for $\lambda = 512$ and 1-D latent (away from optimal tradeoff). The analysis transform is not sufficiently steep.

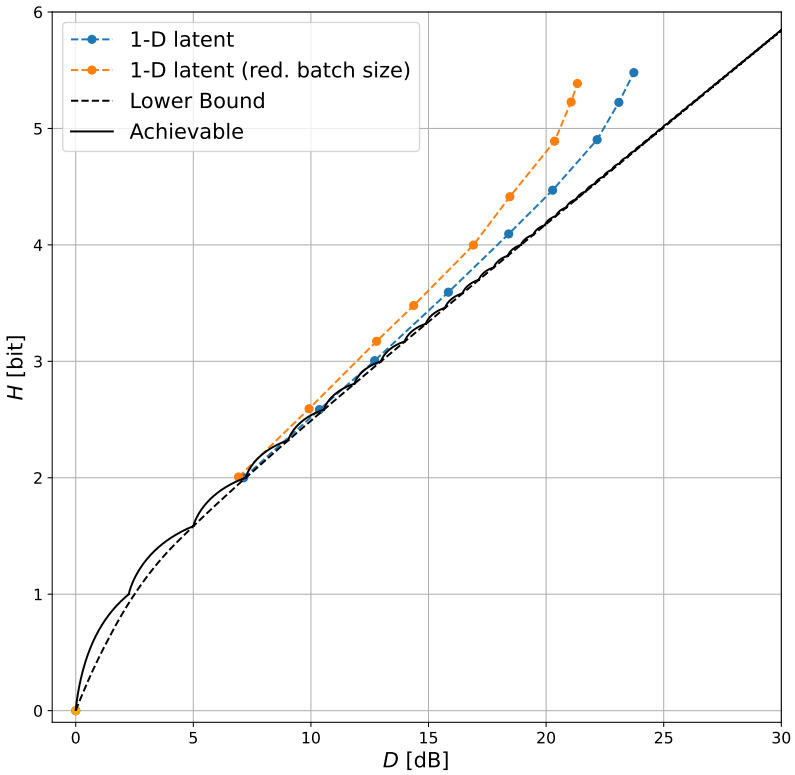


Figure 2.5: Circle entropy-distortion tradeoff with batch size reduced to 8 from 1024.

This is confirmed in Fig. 2.4, which shows the quantized output of the trained analysis transform g_a as a function of the angle θ 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 θ, 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. 2.5). Note that the extent of suboptimality is also dictated by the distortion arising from the quantization process since, for optimality, the distortion due to invertibility only needs to be negligible in comparison. This in turn implies that the analysis transform need not be arbitrarily steep and therefore the suboptimality is not completely dictated by the invertibility issue.

In the case of autoencoders, a similar phenomenon was uncovered in [11]. They investigate autoencoder behaviour to reconstruct training points on the unit circle, inspired by using autoencoders for anomaly detection of particles produced in relativistic collisions. They observe a "finite-sized break region" that is invariant to change in hyperparameters and neural network architecture. Their explanation, based on analysis restricted to the training dynamics of a fully-connected two-layer

neural network-based autoencoder, suggests that the invertibility issue is a complex interplay of stochastic training and architecture of neural networks.

2.3.2 Fourier Features: A Fix

The inability of the analysis transform to approximate a sufficiently steep function hints at an inability to learn a high frequency function. This insight is not new in multilayer perceptron-based (MLP-based) neural networks. Indeed, the inductive bias of neural networks towards smoother functions, a phenomenon termed "spectral bias", has been studied in the context of explaining their generalization capabilities in supervised learning by modeling the training dynamics as kernel regression with a neural tangent kernel (NTK). In particular, these works show that the training dynamics of a two-layer ReLU network can be approximated by a linear dynamical system whose convergence rate is dictated by the eigenvalues of the NTK matrix, which in turn are related to the frequencies of the learned functions.

Recently, [62] analyzed Fourier features [55] on some low-dimensional regression tasks relevant to computer vision. In particular, for 1D function regression, they found that Fourier features help learn high-frequency functions faster since they widen the NTK eigenvalue spectra thus enabling lower eigenvalues (which correspond to high frequencies) being learnt at a faster rate than before.

Consider the mapping of a vector $v \in \mathbb{R}^d$ to a $2m$-dimensional vector

$$\gamma(v) = \left[\sin\left(2\pi v^\top b_1\right), \cos\left(2\pi v^\top b_1\right), \cdots, \sin\left(2\pi v^\top b_m\right), \cos\left(2\pi v^\top b_m\right)\right].$$

In our simulations, we choose the frequency matrix $B = [b_1, b_2, \cdots, b_m]$ such that each entry of the matrix is drawn from a Gaussian distribution with mean 0 and standard deviation σ. Therefore

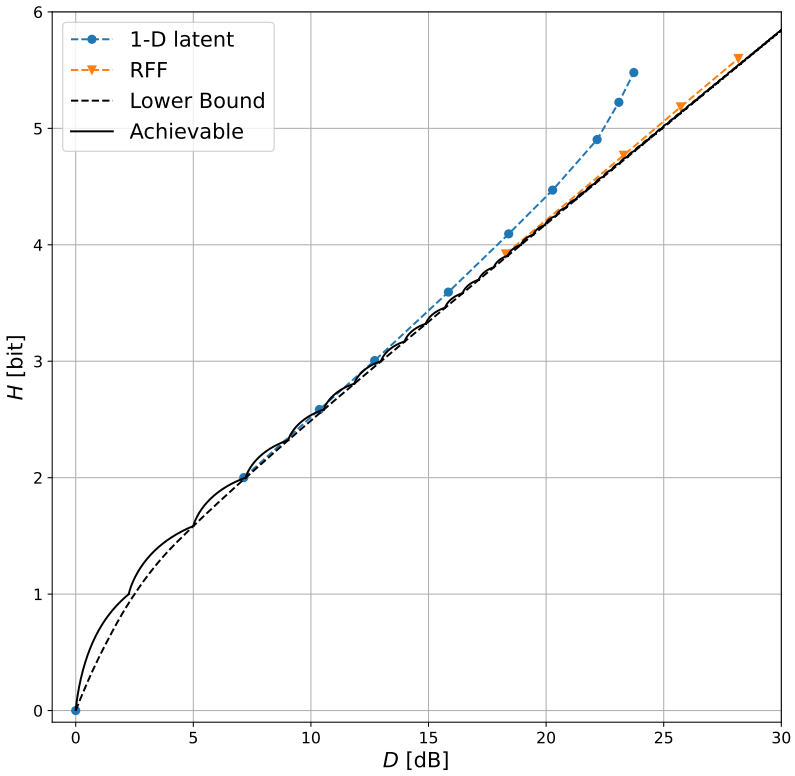


Figure 2.6: Circle entropy-distortion tradeoff with Random Fourier Features.

implementing Fourier features involves deciding where to include them, how many frequencies to include and what value of σ to choose. For the circle source, we see from Figure 2.6 that embedding $m = 500$ Fourier features just before the analysis transform with $\sigma = 10$ attains the optimal entropy-distortion tradeoff.

The suboptimality of vanilla ANN-based compressors 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.

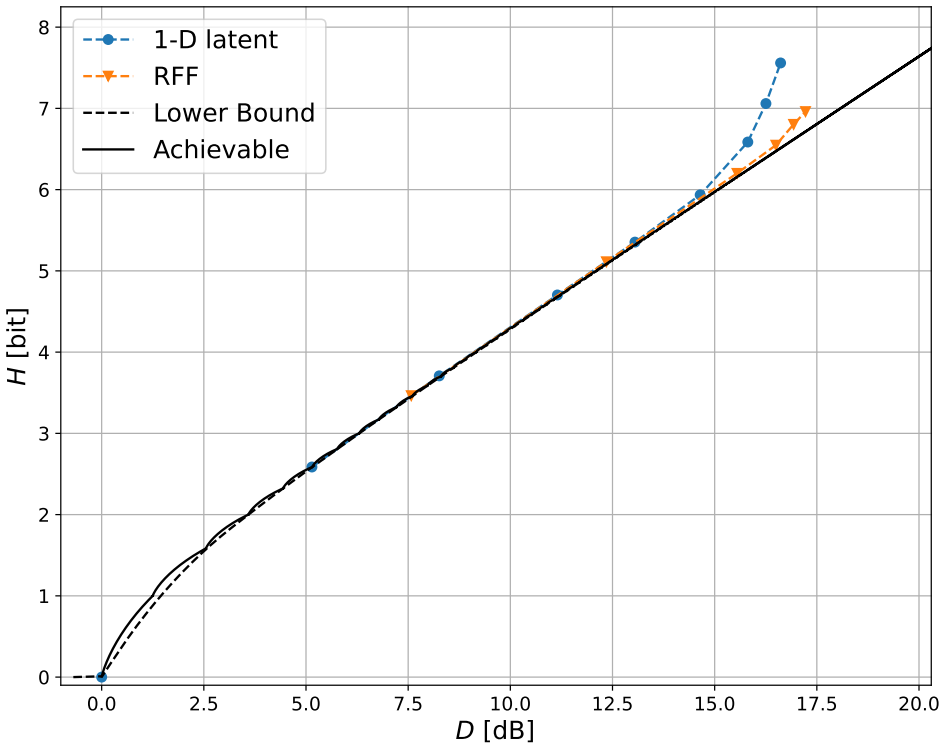
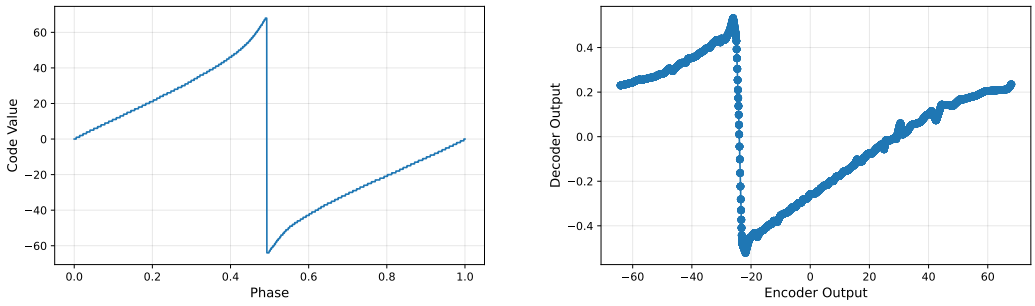


Figure 2.7: Entropy-distortion tradeoff for ramp.



(a) (b)

Figure 2.8: (a) Quantized encoder output vs. phase for $\lambda = 4096$ (away from optimal tradeoff). (b) Decoder output at randomly chosen index vs. encoder output for $\lambda = 4096$. Here the synthesis transform is insufficiently steep.

2.3.3 Ramp

Figure 2.7 shows the entropy-distortion tradeoff for the ANN-based compressor. We see that the performance is suboptimal at high rates. The problem is similar to that which arose with the circle. Per Theorem 15, we expect the analysis transform to output the phase V, or some scaled and shifted version of it such as $Y = \alpha(v + V \mod 1) + \beta$. The training process does indeed find such a solution for $g_a(\cdot)$, as shown in Fig. 2.8a. Now consider the synthesis transform, $g_s(\cdot)$, evaluated at a particular output time, t. The mapping we require in order to recover J_t from Y is

$$Y \mapsto ((Y - \beta)/\alpha + t - v) \mod 1. \tag{2.89}$$

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 (2.89) 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,

$$X_t = [(t + V) \mod 1] - \frac{1}{2}. \tag{2.90}$$

Thus the analysis transform can simply transmit any of the input samples directly, which is clearly a continuous map. On the other hand, Fig. 2.8b shows the reconstruction, evaluated at a fixed time index, as a function of the encoder output. We see that the discontinuity is insufficiently steep.

As with the circle, at low rates, the approximation error noted above is negligible compared with the quantization error, and thus this phenomenon does not lead to entropy-distortion suboptimality.

This reasoning assumes that the synthesis transform must be of the form in (2.89), which we have not established. We have merely shown, via the proof of Theorem 15, that it is an optimal approach.

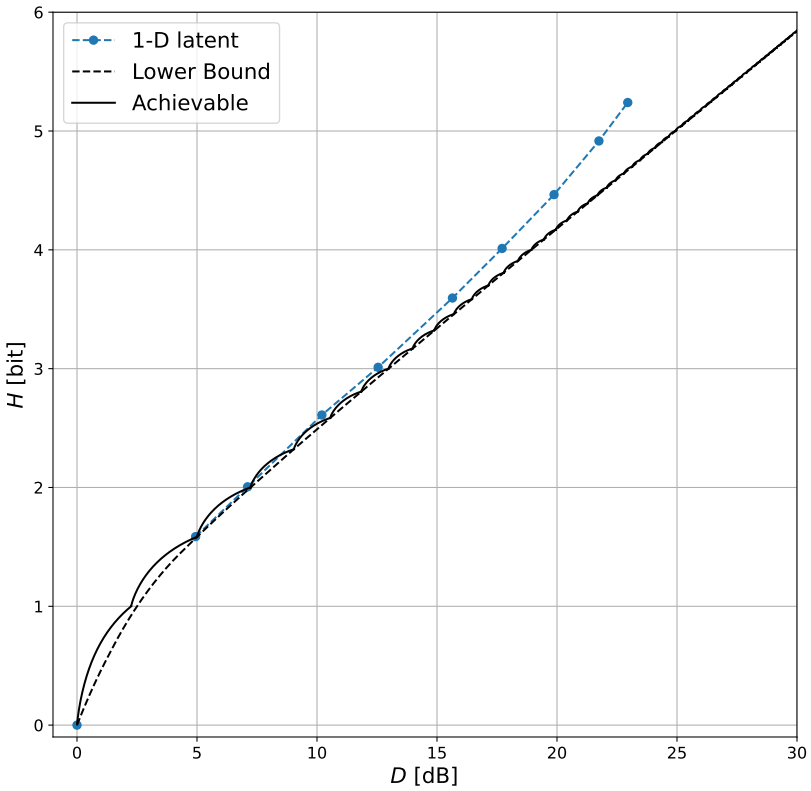


Figure 2.9: Entropy-distortion tradeoff for sinusoid.

As seen in Figure 2.7, Fourier features with $m = 500$ and $\sigma = 1$ significantly reduce the gap to optimality. Unlike the circle, the Fourier features are embedded after the output of the encoder, i.e., just before the synthesis transform.

2.3.4 Sinusoid

The sinusoid is completely described by the phase, therefore we have $d_c = 1$. From Figure 2.9, we see that the performance of the ANN-based compressors described in Section 2.1.2 is suboptimal

for high rates. The difficulty of compressing the sinusoid is similar to that of compressing the circle since the analysis transform tries to invert the $\Phi \mapsto P_t$ mapping. In future work, we investigate the performance of RFFs on the sinusoid process.

CHAPTER 3

ONE-BIT QUANTIZATION

In this chapter, we focus on characterizing the optimal one-bit quantizer for a given source under mean squared error (MSE). Despite the simplicity with which this problem can be stated, relatively little is known about it. For log-concave densities, there exists a unique locally optimal quantizer, which can be found using the Lloyd-Max algorithm [27, 66, 38, 47, 42]. For sources with a density of the form $f(x) = g(x^T Kx)$, where $g(\cdot)$ is decreasing and K is positive semidefinite, Magnani *et al.* [44] show that the optimal reconstructions lie on the major axis of the ellipsoid associated with K. On the other hand, it is known that the optimal quantizer is not necessarily symmetric about 0 even if the distribution itself is. Consider the distribution that is uniformly distributed across the three points $\{-1, 0, 1\}$. It is elementary to check that the best symmetric quantizer is outperformed by one that maps to the closest reconstruction among the set $\{-1, 1/2\}$. See Abaya and Wise [2] for an earlier example that is continuous and monotonically decreasing (cf. [44]).

We develop results toward a general theory of optimal one-bit quantization. Any optimal one-bit quantizer can evidently be implemented via a projection operation followed by a thresholding. We follow Magnani *et al.* in the sense that we focus on identifying the best direction in which to project; once this is identified, the optimal threshold can be found by a one-dimensional sweep. The optimal direction is controlled by a tension between the variance of the projected source and its "amenability" to one-bit quantization. On the one hand, quantizing high-variance directions results in a larger variance drop, i.e., a lower MSE. On the other hand, for a given variance, some distributions result in a lower variance drop under one-bit quantization than others (consider, for example, a standard Normal versus the uniform distribution on $\{-1, 1\}$; see [70] for a naturally-occurring example). We provide methods for resolving this tension, which we demonstrate on an example random process called the *stationary sawbridge* which we have encountered in Chapter 2.

For this infinite-dimensional process we characterize the optimal one-bit quantizer. Moreover, we show that it is found by an off-the-shelf ANN compressor trained via stochastic gradient descent (SGD).

We then consider variable-rate quantization given its ability to attain extremely low bit-rates. Like most previous literature on variable-rate quantization, we adopt the entropy of the encoded outputs as the proxy for rate. While entropy-constrained scalar and vector quantization has received much attention [21],[30], their analysis in the low-rate regime has been relatively less explored. [45] consider Gaussian sources with squared distortion error and establish the asymptotic optimal low-rate performance in terms of the slope of the optimal entropy-distortion function as the distortion approaches the variance of the source. They further demonstrate that both, uniform quantization with a threshold offset and variable-rate binary quantizers, attain this optimal performance in the asymptotic low-rate regime. Given that modern learned compressors based on the nonlinear transform coding paradigm [7] employ dithered uniform quantization, we analyze its asymptotic low-rate performance on vector Gaussian sources with the simplification of linear encoding and decoding transforms. We find that dither performs infinitely worse compared to the optimal low-rate slope.

Most of the prior work on optimal one-bit quantizers focuses on communication instead of compression [71, 50]. There the objective is to maximize mutual information or bit error rate instead of MSE. Nonetheless, the methods in this chapter may have some utility in that application.

3.1 Preliminaries

Let $\mathcal{H}$ be a real Hilbert space with a countable basis, which we make measurable by endowing it with its Borel σ-algebra. Let X be a random variable in $\mathcal{H}$. Without loss of generality, we assume

throughout that $\mathbb{E}[X] = 0$, by which we mean $\mathbb{E}[\langle q, X\rangle] = 0$ for all q such that $\|q\|^2 = 1$. We also assume that X has finite variance, by which we mean that $\mathbb{E}\left[\|X\|^2\right] < \infty$.

Definition 19. *A* one-bit quantizer *is an encoder* $f : \mathcal{H} \mapsto \{0, 1\}$ *and a decoder* $g : \{0, 1\} \mapsto \mathcal{H}$. *We denote the* quantization cells *by*

$$A_j = f^{-1}(j) \quad j \in \{0, 1\},$$

and the reconstructions *by*

$$\hat{x}_j = g(j) \quad j \in \{0, 1\}.$$

We will use Q *to refer to both* (f, g) *and* $g \circ f$. *We say that* Q *is a* symmetric *one-bit quantizer if* $\hat{x}_0 = -\hat{x}_1$.

We focus on mean-squared error (MSE) as a performance metric. For a given one-bit quantizer, the amount of variance reduced by quantization is given by the difference between the variance of the source and the mean-squared error of one-bit quantization.

Definition 20.

$$\mathrm{Vardrop}_{\mathrm{Q,X}} \stackrel{def}{=} \mathbb{E}\left[\|\mathrm{X}\|^2\right] - \mathbb{E}\left[\|\mathrm{X} - \mathrm{Q(X)}\|^2\right].$$

We define the variance drop of a source as the supremum of the variance drop over all one-bit quantizers of the source.

Definition 21.

$$\mathrm{Vardrop}_{\mathrm{X}} \stackrel{def}{=} \sup_{\mathrm{Q}} \mathrm{Vardrop}_{\mathrm{Q,X}}.$$

We will require the notions of symmetric real-valued random variables and log-concave probability density functions (pdf) in the rest of the chapter.

Definition 22. *A random variable* X *on* $\mathcal{H}$ *is symmetric if* X *and* $-X$ *have the same distribution.*

Definition 23. *A probability density function* $f : \mathbb{R}^d \mapsto \mathbb{R}_+$ *is* log-concave *if there exists a concave function* $\phi : \mathbb{R}^d \mapsto [-\infty, \infty)$ *such that for all* $x \in \mathbb{R}^d$, $f(x) = e^{\phi(x)}$.

We state a few known results concerning log-concave random vectors below that we will use in the rest of the chapter.

Proposition 24. *([25]) Let X be a d-dimensional log-concave random vector. Then for any* $v \in \mathbb{R}^d$, $X^\top v$ *is also log-concave.*

Proposition 25. *([54]) Let X and Y be two independent d-dimensional log-concave random vectors. Then X + Y is also a log-concave random vector.*

3.2 General Methods

The decision boundary of an optimal one-bit quantizer of a random vector is a hyperplane that is normal to the line joining the two reconstructions. Thus one-bit quantization of a random vector can be reduced to projecting the random vector along a direction and thresholding the projection. It would seem natural to project along the direction with the highest variance. Yet, as noted in the introduction, lower variance directions might be preferred if they are more amenable to one-bit quantization. We begin by making this tension precise.

Definition 26. *The* amenability (to one-bit quantization) *of a zero-mean random variable X in* $\mathcal{H}$ *is*

$$\zeta_X \stackrel{def}{=} 1 - \frac{\inf_Q \mathbb{E}\left[||X - Q(X)||^2\right]}{\mathbb{E}\left[||X||^2\right]}. \tag{3.1}$$

Evidently $0 \leq \zeta_X \leq 1$ and amenability is scale-free in the sense that $\zeta_X = \zeta_{aX}$ for nonzero $a \in \mathbb{R}$. Amenability captures what fraction of the variance of the source can be eliminated through one-bit

quantization. Note that we trivially have

$$\text{Vardrop}_X = \zeta_X \cdot \mathbb{E}\left[\|X\|^2\right]. \tag{3.2}$$

We shall see that in some important cases in which X is real-valued, amenability equals the following moment-based quantity.

Definition 27. *The* inverse squared coefficient of variation *(*iCoV^2*) of a real-valued, zero mean random variable is*

$$\xi_X \overset{def}{=} \frac{\mathbb{E}\left[|X|\right]^2}{\mathbb{E}\left[X^2\right]}. \tag{3.3}$$

We note that the iCoV^2 satisfies similar properties as amenability:

1. **Scale-free**: $\xi_X = \xi_{aX}$ for nonzero $a \in \mathbb{R}$.
2. **Bounded**: $0 \leq \xi_X \leq 1$ where the right hand side inequality follows from Cauchy Schwarz. Both extremes are approachable by distributions with uniformly bounded support. For $X \sim \text{Unif}\{-1, 1\}$, $\xi_X = 1$. For the lower limit, consider $X_{\varepsilon,\delta}$ with probability mass function

$$p_{X_{\varepsilon,\delta}}(\pm 1) = \delta, p_{X_{\varepsilon,\delta}}(\pm\varepsilon) = 0.5 - \delta.$$

It can be verified that

$$\xi_{X_{\varepsilon,\delta}} = \frac{\mathbb{E}\left[\left|X_{\varepsilon,\delta}\right|\right]^2}{\text{Var}_{X_{\varepsilon,\delta}}} = \frac{((1-2\delta)\varepsilon + 2\delta)^2}{(1-2\delta)\varepsilon^2 + 2\delta}.$$

Finally, $\lim_{\delta \mapsto 0}\left(\lim_{\varepsilon \mapsto 0} \xi_{X_{\varepsilon,\delta}}\right) = 0$.

This is not coincidental, as we shall see that the two concepts are equivalent for an important family of distributions.

The iCoV^2 of a few standard distributions whose mean is 0 is given in Table 3.1.

A key stepping stone for finding the direction to project for optimal one-bit quantization is a relation between the variance drop of any real-valued, zero-mean random variable whose optimal one-bit quantizer is symmetric, to its amenability. Note that this relation holds in particular for a symmetric random variable whose pdf is log-concave, since its optimal one-bit quantizer is known to be symmetric [38].

Lemma 28. *Let W be a real-valued, zero-mean random variable with a density. Then*

1. $$\mathrm{Vardrop}_{\mathrm{W}} = \sup_{\mathrm{w}} \left(\mathbb{E}\left[\mathrm{W} \mid \mathrm{W} \geq \mathrm{w}\right]\right)^2 \frac{\Pr\left(\mathrm{W} \geq \mathrm{w}\right)}{\Pr\left(\mathrm{W} < \mathrm{w}\right)}, \tag{3.4}$$

   where the supremum is over all w such that $\Pr(W \geq w) > 0$ *and* $\Pr(W < w) > 0$.

2. *If W has an optimal quantizer that is symmetric, then*

   $$\mathrm{Vardrop}_{\mathrm{W}} = \mathbb{E}\left[|\mathrm{W}|\right]^2 = \xi_{\mathrm{W}} \mathrm{Var}_{\mathrm{W}},$$

   that is, $\zeta_W = \xi_W$. *In particular, the conclusion holds if W is symmetric with a density that is log-concave.*

Proof. Given a quantizer for which both reconstructions have nonzero probability, the MSE is only reduced by taking the quantization cells to be of the form $(-\infty, w)$ and $[w, \infty)$ for some specially chosen $w \in \mathbb{R}$. Then the MSE is only further reduced by taking the reconstructions to be $\mathbb{E}\left[W \mid W < w\right]$ and $\mathbb{E}\left[W \mid W \geq w\right]$. Therefore the mean-squared error is

$$\begin{aligned}
&\Pr\left(W \geq w\right) \mathbb{E}\left[\left(W - \mathbb{E}\left[W \mid W \geq w\right]\right)^2 \mid W \geq w\right] \\
&+ \Pr\left(W < w\right) \mathbb{E}\left[\left(W - \mathbb{E}\left[W \mid W < w\right]\right)^2 \mid W < w\right] \\
&= \mathbb{E}\left[W^2\right] - \Pr\left(W \geq w\right) \left(\mathbb{E}\left[W \mid W \geq w\right]\right)^2 \\
&\qquad - \Pr\left(W < w\right) \left(\mathbb{E}\left[W \mid W < w\right]\right)^2
\end{aligned} \tag{3.5}$$

Since W is zero-mean,

$$\Pr(W \geq w)\,\mathbb{E}[W \mid W \geq w] + \Pr(W < w)\,\mathbb{E}[W \mid W < w] = 0.$$

Thus we have

$$\mathbb{E}[W \mid W < w] = -\frac{\Pr(W \geq w)\,\mathbb{E}[W \mid W \geq w]}{\Pr(W < w)}.$$

Substituting this into (3.5), we obtain that the variance drop associated with this quantizer is upper bounded by

$$\sup_{w} \left(\mathbb{E}[W \mid W \geq w]\right)^2 \frac{\Pr(W \geq w)}{\Pr(W < w)}. \tag{3.6}$$

Since any quantizer for which the output is a.s. constant is dominated by one for which both reconstructions have positive probability, (3.7) implies

$$\mathrm{Vardrop_W} \leq \sup_{\mathrm{w}} \left(\mathbb{E}[\mathrm{W} \mid \mathrm{W} \geq \mathrm{w}]\right)^2 \frac{\Pr(\mathrm{W} \geq \mathrm{w})}{\Pr(\mathrm{W} < \mathrm{w})}. \tag{3.7}$$

Now fix w such that $\Pr(W \geq w) > 0$ and $\Pr(W < w) > 0$. Consider the quantizer

$$f(x) = \begin{cases} 0 & \text{if } x < w \\ 1 & \text{if } x \geq w, \end{cases} \tag{3.8}$$

and

$$g(j) = \begin{cases} \mathbb{E}[W|W \geq w] & \text{if } j = 1 \\ \mathbb{E}[W|W < w] & \text{if } j = 0. \end{cases} \tag{3.9}$$

By the previous calculation, the variance drop associated with this quantizer is

$$\mathbb{E}[W|W \geq w]^2 \frac{\Pr(W \geq w)}{\Pr(W < w)}. \tag{3.10}$$

Thus we have equality in (3.7), which establishes (a).

By considering W in place of $-W$, as necessary, we may assume that an optimal quantizer is

$$f(w) = \begin{cases} 0 & \text{if } w < 0 \\ 1 & \text{if } w \geq 0 \end{cases} \tag{3.11}$$

$$g(b) = \begin{cases} E[W|W < 0] & \text{if } b = 0 \\ E[W|W \geq 0] & \text{if } b = 1 \end{cases} \tag{3.12}$$

and $g(0) = -g(1)$, i.e., $E[W|W < 0] = -E[W|W \geq 0]$. Since W is zero-mean,

$$\Pr(W \geq 0)E[W|W \geq 0] = \Pr(W < 0)E[W|W < 0]. \tag{3.13}$$

which implies that $\Pr(W \geq 0) = \Pr(W < 0) = 1/2$. Then from (a) we have

$$\mathrm{Vardrop_W} = \mathrm{E[W|W \geq 0]^2}. \tag{3.14}$$

To relate this to $\mathrm{iCoV^2}$, note that

$$E[|W|] = \Pr(W \geq 0)E[|W||W \geq 0] + \Pr(W < 0)E[|W|| < 0] \tag{3.15}$$

$$= \Pr(W \geq 0)E[|W||W \geq 0] - \Pr(W < 0)E[W| < 0] \tag{3.16}$$

$$= E[W|W \geq 0]. \tag{3.17}$$

Thus

$$\mathrm{Vardrop_W} = \mathrm{E[|W|]^2} = \xi_{\mathrm{W}}\mathrm{Var_W}. \tag{3.18}$$

From (3.2), $\xi_W = \zeta_W$. If W is symmetric and log-concave then there exists a unique locally-optimal one-bit quantizer [66, 38]. If W is symmetric then the one-bit quantizer

$$f(w) = \begin{cases} 0 & \text{if } w < 0 \\ 1 & \text{if } w \geq 0 \end{cases} \tag{3.19}$$

$$g(j) = \begin{cases} E[W|W < 0] & \text{if } j = 0 \\ -E[W|W \geq 0] & \text{if } j = 1 \end{cases} \tag{3.20}$$

is locally optimal since

$$E[W|W \geq 0] = -E[W|W < 0]. \tag{3.21}$$

Since it is symmetric, the result follows. □

We now consider the general problem of one-bit quantization of random variables in the Hilbert space $\mathcal{H}$. We first show that the variance drop of a random variable in Hilbert space is the supremum of the variance drop of its projection over all directions. If the projection is symmetric and log-concave for every direction then using Lemma 28, the variance drop of the projection can be related to its amenability.

Theorem 29. *Let $\mathcal{H}$ be a Hilbert space with a countable basis and let X be a zero-mean, finite variance random variable in $\mathcal{H}$. The following are true.*

(a)

$$\text{Vardrop}_{\text{X}} = \sup_{q \in \mathcal{H}, \|q\|=1} \text{Vardrop}_{\langle \text{X,q} \rangle} \tag{3.22}$$

$$= \sup_{q \in \mathcal{H}, \|q\|=1} \zeta_{\langle X,q \rangle} \text{Var}_{\langle X,q \rangle}. \tag{3.23}$$

(b) If $\langle X, q \rangle$ is symmetric and log-concave for all q, then

$$\text{Vardrop}_{\text{X}} = \sup_{\text{q} \in \mathcal{H}, \|\text{q}\|=1} \xi_{\langle \text{X,q} \rangle} \text{Var}_{\langle \text{X,q} \rangle}.$$

Proof. (a) Let Q be any one-bit quantizer. Define $q \stackrel{\text{def}}{=} \frac{\hat{x}_1 - \hat{x}_0}{\|\hat{x}_1 - \hat{x}_0\|}$. Let $\{q, b_1, b_2, \ldots\}$ be an orthonormal

basis for $\mathcal{H}$. Then

$$\begin{aligned}
&\mathbb{E}\left[\|X\|^2\right] - \mathbb{E}\left[\|X - Q(X)\|^2\right] \\
&= \mathbb{E}\left[\langle X, q\rangle^2\right] + \sum_{i=1}^{\infty} \mathbb{E}\left[\langle X, b_i\rangle^2\right] \\
&\quad - \mathbb{E}\left[\left\|\langle X, q\rangle\, q + \sum_{i=1}^{\infty} \langle X, b_i\rangle\, b_i \right.\right. \\
&\qquad \left.\left. - \langle Q(X), q\rangle\, q - \sum_{i=1}^{\infty} \langle Q(X), b_i\rangle\, b_i \right\|^2\right]. \qquad (3.24)
\end{aligned}$$

Let $\overline{q} = \Pr(f(X) = 0)\,\hat{x}_0 + \Pr(f(X) = 1)\,\hat{x}_1$. Then

$$\begin{aligned}
\langle Q(X), b_i\rangle &= \langle Q(X) + \overline{q} - \overline{q}, b_i\rangle \\
&= \langle Q(X) - \overline{q}, b_i\rangle + \langle \overline{q}, b_i\rangle = \langle \overline{q}, b_i\rangle,
\end{aligned}$$

where the last equality follows since $Q(X) - \overline{q} = cq$ for $c \in \mathbb{R}$ a.s. and is therefore orthogonal to b_i. Substituting in (3.24),

$$\begin{aligned}
&\mathbb{E}\left[\|X\|^2\right] - \mathbb{E}\left[\|X - Q(X)\|^2\right] \\
&= \mathbb{E}\left[\langle X, q\rangle^2\right] - \mathbb{E}\left[(\langle X, q\rangle - \langle Q(X), q\rangle)^2\right] \\
&+ \sum_{i=1}^{\infty} \mathbb{E}\left[\langle X, b_i\rangle^2\right] - \sum_{i=1}^{\infty} \mathbb{E}\left[(\langle X, b_i\rangle - \langle \overline{q}, b_i\rangle)^2\right]. \\
&\leq \mathbb{E}\left[\langle X, q\rangle^2\right] - \mathbb{E}\left[(\langle X, q\rangle - \langle Q(X), q\rangle)^2\right] \\
&\leq \mathrm{Vardrop}_{\langle \mathrm{X,q}\rangle} \qquad (3.25) \\
&\leq \sup_{q \in \mathcal{H}, \|q\| = 1} \mathrm{Vardrop}_{\langle \mathrm{X,q}\rangle}. \qquad (3.26)
\end{aligned}$$

Conversely, take any $q \in \mathcal{H}$ such that $\|q\| = 1$. Let $Q(\cdot)$ be a one-bit quantizer on $\mathbb{R}$ satisfying

$$\begin{aligned}
&\mathbb{E}\left[\langle X, q\rangle^2\right] - \mathbb{E}\left[(\langle X, q\rangle - Q(\langle X, q\rangle))^2\right] \\
&\qquad \geq \mathrm{Vardrop}_{\langle \mathrm{X,q}\rangle} - \varepsilon. \qquad (3.27)
\end{aligned}$$

Construct a one-bit quantizer $Q^*(\cdot)$ on $\mathcal{H}$ where

$$g^*(0) = g(0)q$$

$$g^*(1) = g(1)q,$$

and $f^*(x) = f(\langle x, q\rangle)$. Then

$$\text{Vardrop}_{\text{X}} \geq \mathbb{E}\left[||\text{X}||^2\right] - \mathbb{E}\left[||\text{X} - \text{Q}^*(\text{X})||^2\right].$$

Let $\{q, b_1, b_2, \ldots\}$ be an orthonormal basis for $\mathcal{H}$. Note that $\langle Q^*(x), b_i\rangle = 0$ for all i and x. Using the decomposition in (3.24), we have

$$\begin{aligned}
\text{Vardrop}_{\text{X}} &\geq \mathbb{E}\left[\langle X, q\rangle^2\right] + \sum_{i=1}^{\infty} \mathbb{E}\left[\langle X, b_i\rangle^2\right] \\
&\quad - \mathbb{E}\left[\left\| \langle X, q\rangle\, q + \sum_{i=1}^{\infty} \langle X, b_i\rangle\, b_i \right. \right. \\
&\qquad \left. \left. - \langle Q^*(X), q\rangle\, q \right\|^2 \right] \\
&= \mathbb{E}\left[\langle X, q\rangle^2\right] - \mathbb{E}\left[(\langle X, q\rangle - \langle Q^*(X), q\rangle)^2\right] \\
&= \mathbb{E}\left[\langle X, q\rangle^2\right] - \mathbb{E}\left[(\langle X, q\rangle - Q(\langle X, q\rangle))^2\right] \\
&\geq \text{Vardrop}_{\langle \text{X,q}\rangle} - \varepsilon.
\end{aligned}$$

But ε and q were arbitrary. Therefore,

$$\text{Vardrop}_{\text{X}} \geq \sup_{\text{q}\in\mathcal{H}, ||\text{q}||=1} \text{Vardrop}_{\langle \text{X,q}\rangle}.$$

Part (b) follows from (a) and Lemma 28.

□

Since the optimal direction to project along requires that the product of amenability and variance of the projection be maximum, projecting along the direction of highest variance need not always be optimal. We now look at an example that illustrates this point.

Example: Let $\overline{S} = \left[S_1, S_2\right]$ where S_1 and S_2 are independent Laplace random variables with mean zero and variance 2. We will show that projecting along $\left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]$ results in a higher variance drop compared to projecting along either of the coordinate vectors. First note that Theorem 29 holds by Proposition 24 since $\overline{S}$ is a symmetric, log-concave random vector. Therefore, it is sufficient to prove that $\mathbb{E}\left[\left|\frac{S_1+S_2}{\sqrt{2}}\right|\right] > \mathbb{E}\left[|S_1|\right] = \mathbb{E}\left[|S_2|\right]$. The pdf of $S_1 + S_2$ is $\frac{1}{4}e^{-|z|}(|z| + 1)$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\left|\frac{S_1 + S_2}{\sqrt{2}}\right|\right] &= \frac{1}{\sqrt{2}}\left[2\int_0^\infty z \cdot \frac{(z+1)e^{-z}}{4} dz\right] \\ &= \frac{3}{2\sqrt{2}} > \mathbb{E}\left[|S_1|\right] = \mathbb{E}\left[|S_2|\right] = 1.\end{aligned}$$

3.2.1 Elliptical Distributions

The decomposition in (3.23) exposes a tension in the choice of the direction along which to project for the purpose of one-bit quantization. On the one hand, we wish to quantize high-variance directions since they afford proportionally larger variance drops, other things equal. On the other hand, some distributions are particularly amenable to one-bit quantization, and as a result they afford a larger variance drop for a given *a priori* variance. The optimal choice of direction must resolve the tension between these competing objectives.

If the amenability is invariant to the direction, then evidently there is no tension to resolve and it is optimal to project along the direction of maximum variance. This occurs in particular when the distribution of $\langle X, q\rangle$ is constant over q up to scalings, as happens when X is multivariate Gaussian, or more generally, when X has an elliptical distribution.

Definition 30. *A zero-mean random variable X in the Hilbert space $\mathcal{H}$ has an* elliptical *distribution if (i) for all $q_1, q_2 \in \mathcal{H}$, we have $\langle X, q_1\rangle \overset{d}{=} \alpha \langle X, q_2\rangle$ for some α, (ii) There exists some q such that*

$\langle X, q\rangle \overset{d}{=} -\langle X, q\rangle$.

Note that this definition is equivalent to Definition 2.2 in [17] that requires that each one-dimensional projection of X be a spherical distribution in 1-D, i.e., be symmetric. As we shall see next, in Euclidean spaces, an elliptically-distributed random variable can be characterized in multiple ways. In a Hilbert space $\mathcal{H}$, the following result from [17] provides a useful representation.

Proposition 31. *If X is a zero-mean, elliptically-distributed random variable on $\mathcal{H}$, then there exists a zero-mean Gaussian element V and a random variable $S \geq 0$ independent of V such that $X \overset{d}{=} SV$.*

The term "elliptically-distributed" is justified by the following fact.

Proposition 32 ([17])**.** *If $\mathcal{H} = \mathbb{R}^d$ and X is a zero-mean, finite variance random variable on $\mathcal{H}$ then X is elliptically distributed if and only if its characteristic function can be written as*

$$\psi_X(t) = \varphi(t^T K t), \tag{3.28}$$

where K is a positive semidefinite matrix and $\varphi(\cdot)$ is the characteristic function of a scalar random variable. If X has a density, $\phi(\cdot)$, then it is elliptically distributed if f can be written as

$$\phi(x) = \frac{|K|^{1/2}}{h}(x^T K^{-1} x), \tag{3.29}$$

where K is positive definite and $h(\cdot)$ is the density of a scalar random variable.

A consequence of $\langle X, q\rangle$ having the same distribution up to scaling is that its amenability is invariant with respect to q.

Theorem 33. *If X is a zero-mean and elliptically-distributed random variable in the finite-dimensional Hilbert space $\mathcal{H}$ then*

$$\underset{q\in\mathcal{H}:\|q\|=1}{\arg\min}\ \mathrm{Vardrop}_{\langle \mathrm{X,q}\rangle} = \underset{\mathrm{q}\in\mathcal{H}:\|\mathrm{q}\|=1}{\arg\min}\ \mathrm{Var}_{\langle \mathrm{X,q}\rangle}. \tag{3.30}$$

Proof. For a finite-dimensional Hilbert space, the unit hypersphere is compact since it is closed and bounded. Since $\mathrm{Var}_{\langle X,q\rangle}$ is a continuous function on a compact set, the minimum on the right hand side exists. Since amenability is scale-free, for elliptical distributions, $\zeta_{\langle X,q\rangle}$ does not depend on q. The proof follows from (3.7). □

Corollary 34. *If X is a zero-mean, finite variance random variable with density of the form*

$$\phi(x) = h(x^T K^{-1} x) \tag{3.31}$$

for some positive definite matrix K and some function $h(\cdot)$, then an optimal one-bit quantizer has reconstructions that are aligned with an eigenvector of K associated with a maximal eigenvalue.

Proof. By Proposition 32, X is elliptically distributed, so $\hat{x}_0$ and $\hat{x}_1$ can be taken to be aligned with the vector $q \in \mathbb{R}^d$ that maximizes $\mathrm{Var}_{\langle X,q\rangle}$, which is evidently an eigenvector of K with maximal eigenvalue. □

Corollary 34 extends and corrects Theorem 2 of Magnani, Ghosh, and Gray [44]. Translated to our setup, they claim that it is optimal to quantize along the eigenvector associated with the minimum eigenvalue of K, which results from a sign error in their proof[1]. To reach this conclusion, they also place more assumptions on K and $h(\cdot)$ than is done here.

3.3 Examples

We now consider applications of the general methods in the previous section to find the optimal one-bit quantizer of various sources. We have already considered a few; namely, elliptical distributions, 2-D Laplace random variable. We will first consider a few standard distributions and

[1] Specifically, equation (4) in [44] incorrectly assumes that the second term is positive, when in fact it is negative.

then focus on the main application of this section, the optimal one-bit quantizer of the stationary sawbridge process.

3.3.1 Standard Distributions

The amenability of a few standard distributions whose mean is 0 is given in Table 3.1.

Table 3.1: Amenability of some standard distributions.

Distribution	**Amenability**
Unif	$3/4$
Unif*Unif	$2/3$
Gaussian	$2/\pi$
Laplacian	$1/2$

Proof of Table 3.1. All distributions in the table are log-concave and symmetric. Therefore, by Lemma 28, $\zeta_X = \xi_X$.

- For $X \sim \text{Unif}\left[-\frac{1}{2}, \frac{1}{2}\right], \xi_X = \frac{\mathbb{E}\left[|X|^2\right]}{\mathbb{E}\left[X^2\right]} = \frac{(1/4)^2}{1/12} = \frac{3}{4}$.
- For $X \sim \text{Unif}\left[-\frac{1}{2}, \frac{1}{2}\right] * \text{Unif}\left[-\frac{1}{2}, \frac{1}{2}\right]$,

$$\xi_X = \frac{\mathbb{E}\left[|X|^2\right]}{\mathbb{E}[X^2]} = \frac{\left(\int_0^1 2(1-x)x dx\right)^2}{\int_{-1}^0 x^2(1+x)dx + \int_0^1 (1-x)x^2 dx} = \frac{1/9}{1/6} = \frac{2}{3}.$$

- For $X \sim \mathcal{N}(0,1), \xi_X = \frac{\mathbb{E}\left[|X|^2\right]}{\mathbb{E}\left[X^2\right]} = \frac{2}{\pi}$.
- For $X \sim \text{Lap}(0,1)$, since $|X| \sim \text{Exp}(1), \mathbb{E}[|X|] = 1$. Therefore, $\xi_X = \frac{1}{2}$.

□

3.3.2 Sawbridge Family

Wagner and Ballé [70] studied the sawbridge process, which is defined as

$$X_t \stackrel{\text{def}}{=} t - \mathbf{1}\left(t \geq U\right) \qquad t \in [0,1]\,,$$

where $U \sim \text{Unif}\,[0,1]$. We denote the entire process $\{X_t\}_{t=0}^1$ by X and call it the *nonstationary sawbridge* to distinguish it from the *stationary sawbridge*

$$Y_t \stackrel{\text{def}}{=} X_{(t+V) \bmod 1} \qquad t \in [0,1]\,, \tag{3.32}$$

where $U, V \sim \text{Unif}\,[0,1]$ and $U \perp\!\!\!\perp V$. We denote the entire process $\{Y_t\}_{t=0}^1$ by Y.

Since the stationary sawbridge is a rotation of the nonstationary sawbridge in time, both the processes have the same average value or DC, $\int_0^1 X_t dt = \int_0^1 Y_t dt = U - 0.5$. For the nonstationary sawbridge, it is known from Corollary 2 in [70] that an optimal one-bit quantizer is the sign of the DC. From Theorem 29 we know that finding an optimal one-bit quantizer is equivalent to finding an optimal direction to project upon and then quantizing the projection. It should be noted that the constant function equal to 1 is not the highest variance eigenfunction of X providing another instance where projecting along a direction different from the highest variance direction is optimal. As we shall see below, this is not the case for stationary sawbridge. Our main result in this section is that the optimal direction to project upon is the constant function equal to 1 and therefore, the sign of the DC is an optimal one-bit quantizer for the stationary sawbridge. We now specify the eigenfunctions and eigenvalues of the stationary sawbridge.

Lemma 35. *The functions* $\psi_{1,t} = 1, \psi_{2k,t} = \sqrt{2}\sin\left(2\pi kt\right), \psi_{2k+1,t} = \sqrt{2}\cos\left(2\pi kt\right)$ *for* $k \geq 1$ *form an orthonormal basis of* $L^2\,[0,1]$ *and are the eigenfunctions of the stationary sawbridge with eigenvalues* $\lambda_1 = \frac{1}{12}, \lambda_{2k} = \lambda_{2k+1} = \frac{1}{4\pi^2 k^2}$.

The proof of Lemma 35 is in section 3.3.2.

Theorem 36. *Let* $f^* : L^2[0,1] \mapsto \{0,1\}$ *be defined as* $f^*(Y) = 1$ *if* $\int_0^1 Y_t dt > 0$ *and* $f^*(Y) = 0$ *otherwise. Define* $g^* : \{0,1\} \mapsto L^2[0,1]$ *as* $g^*(0) = -0.25, g^*(1) = 0.25$. *Then* $g^* \circ f^*$ *is an optimal one-bit quantizer of* Y.

Proof. From Theorem 29, we know that

$$\text{Vardrop}_{\text{Y}} = \sup_{q_t \in L^2[0,1], ||q_t||=1} \text{Vardrop}_{\int_0^1 q_t Y_t dt}$$

Therefore, finding the unit norm function q_t that maximizes the variance drop of the projection is sufficient to obtain an optimal one-bit quantizer of Y. Define the projection of Y_t on q_t as $Z \overset{\text{def}}{=} \int_0^1 q_t Y_t dt$. Then for $T \in \mathbb{R}$, an optimal decision rule for quantizing Z can be written as

$$Z \underset{1}{\overset{0}{\lessgtr}} T.$$

We prove that $q_t^* = 1$ is optimal and that the quantizer for this choice is symmetric, $T^* = 0$. The proofs of Lemmas 37, 38, 39 are in section 3.3.2.

Lemma 37. *For a unit norm* q_t, *define* $Z \overset{def}{=} \int_0^1 q_t Y_t dt$. *Let* $\theta \overset{def}{=} \left(\int_0^1 q_t dt\right)^2$. *Then,*

1. $Z = \sqrt{\theta} Z_{DC} + \sqrt{1-\theta} Z_{AC}$, *where* $Z_{DC} \overset{def}{=} sgn\left(\int_0^1 q_t dt\right) \int_0^1 Y_t dt$ *and* $Z_{AC} \overset{def}{=} \int_0^1 g_t Y_t dt$ *where* g_t *is unit norm and* $\int_0^1 g_t dt = 0$.
2. Z_{AC} *and* Z_{DC} *are independent.*

Since q_t is arbitrary, it suffices to show that $\text{Vardrop}_{\text{Z}_{\text{DC}}} = \max_{\theta \in [0,1]} \text{Vardrop}_{\text{Z}}$. Consider two cases a) $\theta \leq \frac{5}{8}$ and b) $\frac{5}{8} < \theta < 1$. The following lemma proves that the optimal θ cannot be smaller than $\frac{5}{8}$.

Lemma 38. *If* $\theta \leq \frac{5}{8}$, $\text{Vardrop}_{\text{Z}} \leq \text{Var}_{\text{Z}} < \text{Vardrop}_{\text{Z}_{\text{DC}}}$.

For large θ, a variance argument like before does not work because the variance of the DC is high. We use the structure of the probability density function of Z, f_Z, to show that the optimal quantizer of Z is symmetric.

Let the support of $\sqrt{\theta}Z_{DC}$ be $[-a, a]$, and that of $\sqrt{1-\theta}Z_{AC}$ be $[-b, c]$ where $c \leq b$ without loss of generality. Note that for $\theta > \frac{5}{8}$, $a > \frac{\sqrt{5}}{4\sqrt{2}}$ and $b < \frac{1}{4\sqrt{2}}$. Also, the support of Z is $[-(a+b), a+c]$ with $f_Z(z) = \frac{1}{2a}$ for $z \in [-(a-c), a-b]$. Note that $\frac{a}{2} < a - b$.

We now construct a random variable $\widetilde{Z} = \sqrt{\theta}Z_{DC} + \sqrt{1-\theta}\widetilde{Z_{AC}}$, where $\sqrt{1-\theta}\widetilde{Z_{AC}} = -b$ with probability $\frac{c}{c+b}$ and c with probability $\frac{b}{c+b}$. We show that $\text{Vardrop}_{Z_{DC}} \geq \text{Vardrop}_{\widetilde{Z}} \geq \text{Vardrop}_Z$ with equality holding for $\theta = 1$.

Lemma 39. *For $\theta > \frac{5}{8}$, $\text{Vardrop}_{Z_{DC}} \geq \text{Vardrop}_{\widetilde{Z}} \geq \text{Vardrop}_Z$, where equality holds for $\theta = 1$.*

Therefore, the optimal direction to quantize is $q_t^* = 1$ and the optimal quantizer of the projection is symmetric because the uniform distribution is log-concave. This corresponds to the encoder $f^*(Y) = 1$ if $Z_{DC} > 0$ and $f^*(Y) = 0$ otherwise. By the Lloyd-Max conditions, the reconstructions are given by $g^*(1) = \mathbb{E}\left[Y_t \mid f^*(Y) = 1\right] = 0.25$ and $g^*(0) = \mathbb{E}\left[Y_t \mid f^*(Y) = 0\right] = -0.25$.

□

Proofs of Lemmas

We list the proofs of unproven lemmas here.

Proof of Lemma 35. Define $R_t \stackrel{\text{def}}{=} (t+V) \bmod 1$. The autocorrelation of Y_t is

$$\begin{aligned} K(s,t) &= \mathbb{E}\left[Y_s Y_t\right] \\ &= \mathbb{E}\left[\left(R_s - \mathbf{1}\left(R_s \geq U\right)\right)\left(R_t - \mathbf{1}\left(R_t \geq U\right)\right)\right] \\ &= \mathbb{E}\left[R_s R_t\right] + \mathbb{E}\left[\mathbf{1}\left(\min\left(R_s, R_t\right) \geq U\right)\right] \\ &\quad - \mathbb{E}\left[R_s \mathbf{1}\left(R_t \geq U\right)\right] - \mathbb{E}\left[R_t \mathbf{1}\left(R_s \geq U\right)\right] \\ &= \frac{(s-t)^2}{2} - \frac{|s-t|}{2} + \frac{1}{6}. \end{aligned}$$

If $\{\psi_{k,t}\}_{k=1}^{\infty}$ and $\{\lambda_k\}_{k=1}^{\infty}$ are the eigenfunctions and eigenvalues of K, then for all k and $s \in [0,1]$,

$$\int_0^1 K(s,t)\psi_{k,t}dt = \lambda_k \psi_{k,s}.$$

By differentiating both sides w.r.t s and solving the resultant differential equation, it can be shown that the eigenfunctions are $\psi_{1,t} = 1, \psi_{2k,t} = \sqrt{2}\sin\left(2\pi k t\right), \psi_{2k+1,t} = \sqrt{2}\cos\left(2\pi k t\right)$ for $k \geq 1$. The corresponding eigenvalues are $\lambda_1 = \frac{1}{12}, \lambda_{2k} = \lambda_{2k+1} = \frac{1}{4\pi^2 k^2}$. □

Proof of Lemma 37. Since q is unit norm, $\theta \in [0,1]$. We can decompose q_t into its DC and AC,

$$q_t = \operatorname{sgn}\left(\int_0^1 q_t dt\right)\sqrt{\theta} + \sqrt{1-\theta}g_t, \tag{3.33}$$

where g_t is unit norm and because of orthogonality, $\int_0^1 g_t dt = 0$. Therefore

$$Z = \sqrt{\theta}Z_{DC} + \sqrt{1-\theta}Z_{AC}.$$

The nonstationary sawbridge can be written as

$$X_t = \left(t - U_{DC} - \frac{1}{2}\right) \bmod 1 - \frac{1}{2} + U_{DC}$$

where $U_{DC} \sim \text{Unif}\,[-0.5, 0.5]$. Thus

$$\begin{aligned} Y_t &= \left((t+V) \bmod 1 - U_{DC} - \frac{1}{2}\right) \bmod 1 - \frac{1}{2} + U_{DC} \\ &= \left(t + V - U_{DC} - \frac{1}{2}\right) \bmod 1 - \frac{1}{2} + U_{DC} \\ &= \left(t + (V - U_{DC}) \bmod 1 - \frac{1}{2}\right) \bmod 1 - \frac{1}{2} + U_{DC}. \end{aligned}$$

Since $(V - U_{DC}) \bmod 1$ is independent of U_{DC} and since Z_{DC} depends only on U_{DC} and Z_{AC} depends only on $(V - U_{DC}) \bmod 1$, Z_{DC} and Z_{AC} are independent.

□

Proof of Lemma 38. $\text{Var}\,(Z_{DC}) = \frac{1}{12}$. By the Karhunen-Loève theorem, we can express Y as

$$Y_t = G_1\psi_{1,t} + \sum_{k=2}^{\infty} G_k\psi_{k,t}, \tag{3.34}$$

where $\{\psi_{k,t}\}_{k=1}^{\infty}$ are eigenfunctions of K, and $G_k \overset{\text{def}}{=} \int_0^1 Y_t\psi_{k,t}dt$ for $k \geq 1$. By Lemma 35, since $\{\psi_{k,t}\}_{k=1}^{\infty}$ is an orthonormal basis for $L^2\,[0,1]$, for $\{c_k\}_{k=1}^{\infty} \in \mathbb{R}$, we can represent g as

$$g_t = c_1\psi_{1,t} + \sum_{k=2}^{\infty} c_k\psi_{k,t}. \tag{3.35}$$

Since $\int_0^1 g_t dt = 0$ and $\psi_{k,t}$ is orthogonal to $\psi_{1,t} = 1$ for $k \geq 2$, $c_1 = 0$. This implies

$$Z_{AC} = \int_0^1 g_t Y_t dt = \sum_{k=2}^{\infty} G_k c_k. \tag{3.36}$$

Since $Z_{AC} = \sum_{k=2}^{\infty} G_k c_k$,

$$\text{Var}\,(Z_{AC}) = \sum_{k=2}^{\infty} \text{Var}\,(G_k)\, c_k^2 = \sum_{k=2}^{\infty} \lambda_k c_k^2.$$

Further, since g is unit norm, $\sum_{k=2}^{\infty} c_k^2 = 1$. Therefore,

$$\text{Var}\,(Z_{AC}) \leq \max_{k\geq 2} \lambda_k = \frac{1}{4\pi^2}.$$

For $U_{DC} = U - 0.5$,

$$\begin{aligned} Z_{AC} &= \int_0^1 g_t Y_t dt \\ &= \int_0^1 g_t \left(Y_t - U_{DC} + U_{DC}\right) dt \\ &= \int_0^1 g_t \left(Y_t - U_{DC}\right) dt \\ &\leq \sqrt{\int_0^1 \left(Y_t - U_{DC}\right)^2 dt} \leq \frac{1}{\sqrt{12}}. \end{aligned}$$

Therefore, Z_{AC} lies within $\left[-\sqrt{\frac{1}{12}}, \sqrt{\frac{1}{12}}\right]$ almost surely.

For $\theta \leq \frac{5}{8}$,

$$\begin{aligned} \text{Var}\left(Z\right) &= \theta \text{Var}\left(Z_{DC}\right) + \left(1 - \theta\right) \text{Var}\left(Z_{AC}\right) \\ &\leq \frac{1}{4\pi^2} + \frac{5}{8}\left(\frac{1}{12} - \frac{1}{4\pi^2}\right) \leq \frac{1}{16}. \end{aligned}$$

□

Proof of Lemma 39. We first prove that for both Z and $\widetilde{Z}$ the median is 0.

$$\begin{aligned} \Pr\left(Z \geq 0\right) &= \int_0^{a+c} f_Z(z) dz \\ &= \int_0^{a-b} \frac{1}{2a} dz + \int_{a-b}^{a+c} f_Z(z) dz \\ &= \frac{1}{2} - \frac{b}{2a} + \int_{a-b}^{a+c} f_Z(z) dz, \end{aligned} \tag{3.37}$$

where f_Z is the pdf of Z. Since Z is the sum of independent random variables, f_Z can be written as a convolution of $f_{\sqrt{\theta} Z_{DC}}$ and $f_{\sqrt{1-\theta} Z_{AC}}$. For simplicity of notation we denote the pdf of $\sqrt{1-\theta} Z_{AC}$ as

f_{AC} and denote its cumulative distribution function (cdf) as F_{AC}.

$$\begin{aligned}\int_{a-b}^{a+c} f_Z(z)dz &= \int_{a-b}^{a+c}\left(\int_a^{z-c}\frac{1}{2a}f_{AC}\left(z-\tau\right)d\tau\right)dz\\ &= \frac{1}{2a}\int_{a-b}^{a+c}\left(\int_a^{z-c} f_{AC}\left(\gamma\right)d\gamma\right)dz\\ &= \frac{1}{2a}\int_{a-b}^{a+c} 1 - F_{AC}\left(z-a\right)dz\\ &= \frac{b}{2a},\end{aligned} \tag{3.38}$$

where in the last equality we use the identity $\int_\ell^u F(x)dx = u - \mathbb{E}\left[X\right]$ for a random variable X with cdf F whose support is $[\ell, u]$ where $\ell, u \in \mathbb{R}$. Substituting (3.37) in (3.38), we get $\Pr\left(Z \geq 0\right) = \Pr\left(Z < 0\right) = \frac{1}{2}$. Note that for the proof above we only require that $\sqrt{1-\theta}Z_{AC}$ is supported on $[-b, c]$ and its mean is 0. Therefore, $\Pr\left(\widetilde{Z} \geq 0\right) = \Pr\left(\widetilde{Z} < 0\right) = \frac{1}{2}$.

We now compute $\text{Vardrop}_{\widetilde{Z}}$ using Lemma 28. The optimal w in (3.4) lies in $\left[-\frac{a}{2}, \frac{a}{2}\right]$. For $w \in \left[-\frac{a}{2}, \frac{a}{2}\right]$, $\Pr\left(\widetilde{Z} \geq w\right) = \frac{1}{2} - \frac{w}{2a}$ and

$$\begin{aligned}\mathbb{E}\left[\widetilde{Z} \mid \widetilde{Z} \geq w\right] &= \frac{1}{\frac{1}{2}-\frac{w}{2a}}\left[\int_w^{a-b}\frac{z}{2a}dz\right.\\ &\qquad \left.+\int_{a-b}^{a+c} z\frac{1}{2a}\frac{b}{(c+b)}\right]\\ &= \frac{1}{4a\left(\frac{1}{2}-\frac{w}{2a}\right)}\left(a^2+bc-w^2\right).\end{aligned}$$

It can be shown that

$$\begin{aligned}&\arg\max_{w\in\left[-\frac{a}{2},\frac{a}{2}\right]} \mathbb{E}\left[\widetilde{Z} \mid \widetilde{Z} \geq w\right]^2 \frac{\Pr\left(\widetilde{Z} \geq w\right)}{\Pr\left(\widetilde{Z} < w\right)}\\ &= \arg\max_{w\in\left[-\frac{a}{2},\frac{a}{2}\right]} \frac{\left(\frac{a^2+bc-w^2}{2a}\right)^2}{1-\frac{w^2}{a^2}} = 0.\end{aligned}$$

Therefore,

$$\text{Vardrop}_{\widetilde{Z}} = \left(\frac{\mathrm{a}^2+\mathrm{bc}}{2\mathrm{a}}\right)^2 \leq \text{Vardrop}_{Z_{\text{DC}}} = \frac{1}{16} \tag{3.39}$$

for $a = \frac{\sqrt{\theta}}{2}$ and $b = \frac{\sqrt{1-\theta}}{\sqrt{12}}$. Equality holds for $\theta = 1$.

We now show that $\text{Vardrop}_Z \leq \text{Vardrop}_{\overline{Z}}$. We again note that for the optimal quantizer of Z, $w \in \left[-\frac{a}{2}, \frac{a}{2}\right]$. Therefore, since $\Pr(Z \geq 0) = \Pr(Z < 0) = \frac{1}{2}$,

$$\mathbb{E}[Z \mid Z \geq w]^2 \frac{\Pr(Z \geq w)}{\Pr(Z < w)} = \frac{\left(\int_w^{a+c} z f_Z(z)\, dz\right)^2}{\frac{1}{4} - \frac{w^2}{4a^2}}. \tag{3.40}$$

$$\int_m^{a+c} z f_Z(z)\, dz = \int_w^{a-b} \frac{z}{2a} dz + \int_{a-b}^{a+c} z f_Z(z)\, dz. \tag{3.41}$$

Note that from (3.38),

$$\begin{aligned}\int_{a-b}^{a+c} z f_Z(z)\, dz &= \int_{a-b}^{a+c} \frac{z}{2a}\left(1 - F_{AC}(z-a)\right) dz \\ &= \frac{1}{2a}\int_{-b}^{c} (a+\tau)\left(1 - F_{AC}(\tau)\right) d\tau. \end{aligned} \tag{3.42}$$

Integrating by parts, we have

$$\begin{aligned}&\int_{-b}^{c} (a+\tau)\left(1 - F_{AC}(\tau)\right) d\tau \\ &= ab - \frac{b^2}{2} + \int_{-b}^{c} \left(a\tau + \frac{\tau^2}{2}\right) f_{AC}(\tau)\, d\tau \\ &= ab - \frac{b^2}{2} + \frac{\int_{-b}^{c} \tau^2 f_{AC}(\tau) d\tau}{2}. \end{aligned} \tag{3.43}$$

We now prove that the last term is bounded by $\frac{bc}{2}$. Since τ^2 is convex, by Jensen's inequality we have

$$\tau^2 \leq b^2\left(1 - \frac{\tau + b}{b + c}\right) + c^2\left(\frac{\tau + b}{b + c}\right)$$

Thus we have

$$\frac{\int_{-b}^{c} \tau^2 f_{AC}(\tau) d\tau}{2} \leq \frac{b^2 \frac{c}{b+c} + c^2 \frac{b}{b+c}}{2} = \frac{bc}{2} \tag{3.44}$$

where we use the fact that the mean of $\sqrt{1-\theta} Z_{AC}$ is 0. Substituting (3.44) in (3.43),

$$\int_{-b}^{c} (a+\tau)\left(1 - F_{AC}(\tau)\right) \leq ab - \frac{b^2}{2} + \frac{bc}{2}. \tag{3.45}$$

Substituting (3.45) in (3.42),

$$\begin{aligned}\int_{a-b}^{a+c} z f_Z(z)\, dz &\leq ab - \frac{b^2}{2} + \frac{bc}{2} \\ &= \int_{a-b}^{a+c} z \frac{b}{b+c} dz.\end{aligned}$$

Therefore,

$$\begin{aligned}\int_{w}^{a+c} z f_Z(z)\, dz &\leq \int_{w}^{a-b} \frac{z}{2a} dz + \int_{a-b}^{a+c} \frac{z}{2a} \cdot \frac{b}{b+c} dz \\ &= \int_{w}^{a+c} z f_{\widetilde{Z}}(z)\, dz. \end{aligned} \tag{3.46}$$

Substituting (3.46) in (3.40),

$$\begin{aligned}&\mathbb{E}\left[Z \mid Z \geq w\right]^2 \frac{\Pr(Z \geq w)}{\Pr(Z < w)} \\ &\leq \mathbb{E}\left[\widetilde{Z} \mid \widetilde{Z} \geq w\right]^2 \frac{\Pr\left(\widetilde{Z} \geq w\right)}{\Pr\left(\widetilde{Z} < w\right)}.\end{aligned}$$

Therefore,

$$\mathrm{Vardrop}_{Z} \leq \mathrm{Vardrop}_{\widetilde{Z}} \leq \mathrm{Vardrop}_{Z_{\mathrm{DC}}}.$$

□

3.4 Numerical Results

We experimentally verify that the optimal one-bit quantizer of the stationary sawbridge is found by neural-network-based variable-rate compressors trained using stochastic gradient descent (SGD).

A neural-network-based compressor consists of an encoder-decoder pair and a factorized entropy model for entropy coding of the latent components. All three components are implemented using fully connected neural networks as in [70] and are trained using the nonlinear transform coding approach in [7]. A single realization of the stationary sawbridge is a vector of 1024 equally spaced points between 0 and 1. At train time, this vector is passed through the encoder and the output of the encoder is quantized using a differentiable approximation of rounding by soft-rounding and adding uniform noise [4]. The soft-quantized latents are then fed to the decoder to obtain the reconstruction. At test time, the latents are quantized by rounding to the nearest integer. The objective function is the rate-distortion Lagrangian where the rate is computed by the entropy model, and the distortion is the mean-squared error between the inputs and the reconstructions. The encoder, decoder and the entropy model are trained using SGD until convergence.

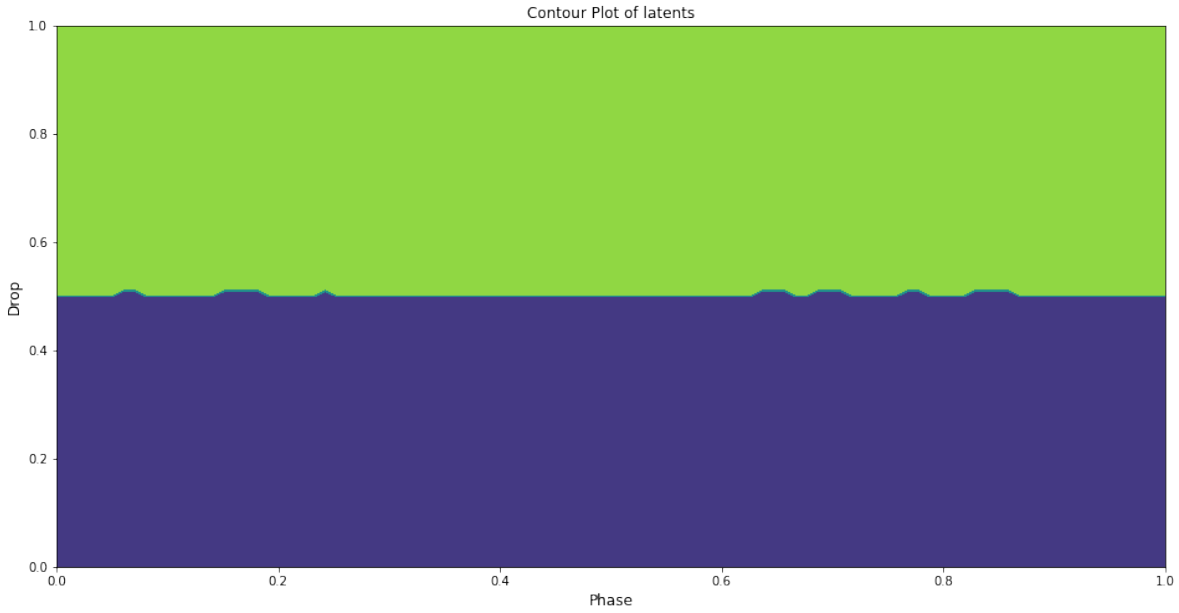


Figure 3.1: Contour plot for stationary sawbridge.

Fig 3.1 is a contour plot of the quantized latent as we vary the drop and phase parameter corresponding to variables U and V in (3.32). Note that the quantized latents are the quantized encoder outputs that are then fed to the decoder. Each of the two shaded regions of Fig 3.1 correspond to a

single quantized latent vector that differ only in a single latent component. Since the regions correspond to whether the drop is greater than 0.5 or not, neural-network-based compressors trained using SGD converge to an optimal one-bit quantizer.

3.5 Variable-Rate Quantization in the Low-rate Regime

So far we have focused on the case where the rate is $\log_2(m) = 1$ bit, irrespective of the input, where m the number of reconstructions is fixed to two. In this section, we relax this requirement by allowing the rate to be dependent on the input, i.e., we consider variable-rate quantization. It is common to use the entropy of the quantized outputs as a proxy for the rate since it is a lower bound to $\log_2(m)$ thereby making it possible to achieve rates much smaller than $\log_2(m)$. This problem of quantization that minimizes the entropy of the encoded outputs subject to a distortion constraint is also called entropy-constrained vector quantization. In addition, modern neural-network based compressors that are based on the nonlinear transform coding paradigm, also perform variable-rate quantization and minimize the entropy-distortion Lagrangian using stochastic minimization algorithms.

We are interested in the low-rate slope of the vector Gaussian source and quantization schemes to achieve the same. In particular, we analyze the low-rate slope of dithered uniform quantization and show that it is suboptimal.

3.5.1 Low-rate Slope for Vector Gaussian Source

For a source X in a space $\mathcal{M}$ equipped with a well-defined norm $\|.\|$, a variable-rate quantizer Q consists of an encoder $f : \mathcal{M} \mapsto \mathbb{N}$ and a decoder $g : \mathbb{N} \mapsto \mathcal{M}$. Its entropy and distortion is given

by

$$H(Q) = -\sum_{i\in\mathbb{N}} \Pr\left(f(X)=i\right)\log\left(\Pr\left(f(X)=i\right)\right),$$
$$D(Q) = \mathbb{E}\left[\|X - g(f(X))\|^2\right].$$

The entropy-distortion function of the source X is given by

$$E(D) = \inf H(Q)$$
$$\text{s.t. } D(Q) \le D.$$

Definition 40. *The low-rate slope of a source X with finite variance is defined as*

$$\lim_{D\to\mathrm{Var}(X)} \frac{E(D)}{\mathrm{Var}(X) - D}.$$

We shall also refer to the low-rate slope of a particular quantization scheme where, in the above definition, instead of the entropy-distortion function of the source, we use the entropy-distortion trade-off attained by the quantization scheme. The following lemma specifies the low-rate slope of a vector Gaussian source.

Lemma 41. *Let $X \sim \mathcal{N}(0, K)$ where K is a positive definite matrix such that $\sigma_1^2 \ge \sigma_2^2 \ge \cdots \ge \sigma_d^2 > 0$ are the eigenvalues of K. The low-rate slope of X is $\frac{\log e}{2\sigma_1^2}$.*

Proof. By the converse of the Shannon coding theorem, the rate-distortion function, denoted by $R(D)$, is a lower bound to the entropy-distortion function. Therefore,

$$\lim_{D\to\mathrm{Var}(X)} \frac{R(D)}{\mathrm{Var}(X) - D} \le \lim_{D\to\mathrm{Var}(X)} \frac{E(D)}{\mathrm{Var}(X) - D}.$$

Therefore, $\frac{\log e}{2\sigma_1^2}$ is a lower bound on the low-rate slope of vector Gaussians. This lower bound can be achieved by considering a scheme that does entropy-constrained scalar quantization of the

maximum variance component. As shown in [45], the low-rate slope of entropy-constrained scalar quantization of a Gaussian source of the form $\mathcal{N}\left(0, \sigma^2\right)$ is $\frac{\log e}{2\sigma^2}$. Since the lower bound to the low-rate slope of the vector Gaussian source is attained, the low-rate slope is $\frac{\log e}{2\sigma_1^2}$. □

While the scheme in the vector-Gaussian case involves quantizing the maximum variance direction, we note that, inline with our observations from Section 3.2, this need not always be the case. Consider the following example.

Example 42. *Let* $X = [X_1, X_2]$ *such that* $X_1 \sim \mathcal{N}\left(0, \sigma^2\right)$ *and* $X_2 \sim Lap\left(0, \frac{\sigma}{2}\right)$ *and* X_1 *and* X_2 *are independent random variables. Note that* $\mathrm{Var}(X_1) = \sigma^2 > \mathrm{Var}(X_2) = \frac{\sigma^2}{2}$. *However, the low-rate slope for quantizing along the maximum variance direction is* $\frac{\log e}{2\sigma^2}$ *while the low-rate slope for quantizing along the minimum variance direction is* 0 *[60].*

3.5.2 Low-rate Slope of Dithered Quantization

[45] previously considered uniform quantization with an offset and showed that it attains the optimal low-rate slope for scalar Gaussians. In this section, we show that dithered uniform quantization is suboptimal in the low-rate regime for entropy-constrained scalar quantization of Gaussian sources. While traditionally dithered quantization has been used primarily because of its property of making quantization noise independent of the source, its utility in modern neural-network-based compression is due to the fact that it enables the entropy-distortion objective to be differentiable.

We consider the following formulation of dithered quantization. Consider a function $Q(.)$ that maps a real number to the nearest integer and let ε be a random variable uniformly distributed over $\left[-\frac{1}{2}, \frac{1}{2}\right]$ whose variance is $\alpha^2 = \frac{1}{12}$. Dithered quantization then refers to adding ε to the source before quantization at the encoder and subtracting the added ε again at the decoder after quantization. [73] showed that $Q(X + \varepsilon) - \varepsilon$ and $X + \varepsilon$ have the same distribution.

For scalar quantization, we consider a quantizer where a source X is multiplied by a scalar s and then undergoes dithered quantization as mentioned above. Given the above equivalence of $Q(X+\varepsilon)-\varepsilon$ and $X+\varepsilon$ in distribution, the quantity $Q(sX+\varepsilon)-\varepsilon$ can be optimally rescaled by the linear least squares estimate (LLSE) of X given $sX+\varepsilon$ to obtain the reconstruction. We now show that the low-rate slope for dithered quantization is strictly larger than $\frac{\log e}{2\sigma^2}$ for scalar Gaussian sources of the form $\mathcal{N}\left(0,\sigma^2\right)$.

Theorem 43. *The low-rate slope of dithered uniform quantization for $X \sim \mathcal{N}\left(0,\sigma^2\right)$ is strictly greater than $\frac{\log e}{2\sigma^2}$.*

Proof. For $X \sim \mathcal{N}\left(0,\sigma^2\right)$, the overall distortion is the conditional variance of X given $sX+\varepsilon$ and is given by

$$D(s) = \sigma^2 - s\sigma^2\left(s^2\sigma^2+\alpha^2\right)^{-1}s\sigma^2 = \frac{\alpha^2\sigma^2}{s^2\sigma^2+\alpha^2}. \tag{3.47}$$

The entropy of the discrete random vector $Q(sX+\varepsilon)$ can be written as

$$H(s) = H(Q(sX+\varepsilon)-\varepsilon \mid \varepsilon).$$

[73] also showed that $H(Q(X+\varepsilon)-\varepsilon \mid \varepsilon) = I(X+\varepsilon;X) = h(X+\varepsilon)$. Therefore, the low-rate slope is given by

$$\lim_{D\to\sigma^2}\frac{H(s)}{\sigma^2-D(s)} = \lim_{s\to 0}\frac{H(s)}{\sigma^2-D(s)} = \lim_{s\to 0}\frac{h(sX+\varepsilon)}{\sigma^2-\frac{\alpha^2\sigma^2}{s^2\sigma^2+\alpha^2}}.$$

By the entropy-power inequality, we have that

$$2^{2h(sX+\varepsilon)} \geq 2^{2h(sX)} + 2^{2h(\varepsilon)}.$$

Taking log on both sides

$$\begin{aligned} h(sX+\varepsilon) &\geq \frac{1}{2}\log\left(2^{2h(sX)}+1\right) \\ &= \frac{1}{2}\log\left(2\pi e s^2\sigma^2+1\right). \end{aligned}$$

The low-rate slope can be lower bounded as a result.

$$\begin{aligned} \lim_{s\to 0}\frac{h(sX+\varepsilon)}{\sigma^2-\frac{\alpha^2\sigma^2}{s^2\sigma^2+\alpha^2}} &\geq \frac{1}{2}\lim_{s\to 0}\frac{\log\left(2\pi e s^2\sigma^2+1\right)}{\sigma^2-\frac{\alpha^2\sigma^2}{s^2\sigma^2+\alpha^2}} \\ &= \frac{\log e}{2\sigma^2}\times 2\pi e\alpha^2 > \frac{\log e}{2\sigma^2}. \end{aligned}$$

□

A more careful calculation below shows that in fact, the low-rate slope of dithered uniform quantization is infinite.

Theorem 44. *The low-rate slope of dithered uniform quantization for $X\sim\mathcal{N}\left(0,\sigma^2\right)$ is infinite.*

Proof. From the proof of Theorem 43, the low-rate slope is given by

$$\lim_{s\to 0}\frac{h(sX+\varepsilon)}{\sigma^2-\frac{\alpha^2\sigma^2}{s^2\sigma^2+\alpha^2}} = .\lim_{s\to 0}\frac{\frac{\partial}{\partial s}h(sX+\varepsilon)}{\frac{\partial}{\partial s}\frac{s^2\sigma^4}{s^2\sigma^2+\alpha^2}} = \lim_{s\to 0}\frac{\frac{\partial}{\partial s}h(sX+\varepsilon)}{\frac{2s\sigma^4\alpha^2}{\left(s^2\sigma^2+\alpha^2\right)^2}}. \tag{3.48}$$

By de Bruijn's identity [23][Theorem 17.7.2],

$$\frac{\partial}{\partial s}h(sX+\varepsilon) = sJ(sX+\varepsilon), \tag{3.49}$$

where $J(.)$ is the Fisher information. Substituting the above in (3.48),

$$\lim_{s\to 0}\frac{h(sX+\varepsilon)}{\sigma^2-\frac{\alpha^2\sigma^2}{s^2\sigma^2+\alpha^2}} = \frac{\alpha^2}{\sigma^4}\lim_{s\to 0}J(sX+\varepsilon). \tag{3.50}$$

Denote the tail distribution function of the standard normal distribution. by $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$. The probability density function of $Y = sX + \varepsilon$, $g_s(y)$ can be expressed as

$$g_s(y) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{\sqrt{2\pi\sigma^2 s^2}} e^{-\frac{(y-x)^2}{2\sigma^2 s^2}} dx = Q\left(\frac{-\frac{1}{2}-y}{\sigma s}\right) - Q\left(\frac{\frac{1}{2}-y}{\sigma s}\right).$$

The Fisher information of Y in terms of $g_s(y)$ is given by

$$J(sX+\varepsilon) = \int_{-\infty}^{\infty} g_s(y)\left(\frac{g_s'(y)}{g_s(y)}\right)^2 dy. \tag{3.51}$$

Let $G(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Since $Q'(x) = -G(x)$,

$$J(sX+\varepsilon) = \int_{-\infty}^{\infty} \frac{1}{\sigma^2 s^2} \frac{\left(G\left(\frac{\frac{1}{2}-y}{\sigma s}\right) - G\left(\frac{-\frac{1}{2}-y}{\sigma s}\right)\right)^2}{Q\left(\frac{-\frac{1}{2}-y}{\sigma s}\right) - Q\left(\frac{\frac{1}{2}-y}{\sigma s}\right)} dy. \tag{3.52}$$

Since the denominator within the integral is at most 1, we can lower bound $J(sX + \varepsilon)$ as

$$\begin{aligned}
J(sX+\varepsilon) &\geq \frac{1}{\sigma^2 s^2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(G\left(\frac{\frac{1}{2}-y}{\sigma s}\right) - G\left(\frac{-\frac{1}{2}-y}{\sigma s}\right)\right)^2 dy && (3.53)\\
&= \frac{1}{\sigma^2 s^2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(G^2\left(\frac{\frac{1}{2}-y}{\sigma s}\right) + G^2\left(\frac{-\frac{1}{2}-y}{\sigma s}\right) - 2G\left(\frac{\frac{1}{2}-y}{\sigma s}\right)G\left(\frac{-\frac{1}{2}-y}{\sigma s}\right)\right) dy. && (3.54)\\
&= \frac{1}{\sigma^2 s^2}\left(\frac{1}{\sqrt{4\pi\sigma^2 s^2}}\left(Q\left(\frac{-\sqrt{2}}{\sigma s}\right) - Q(0)\right) + \frac{1}{\sqrt{4\pi\sigma^2 s^2}}\left(Q(0) - Q\left(\frac{\sqrt{2}}{\sigma s}\right)\right)\right. && (3.55)\\
&\quad \left. -2\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{1}{\sqrt{2\pi\sigma^2 s^2}}\right)^2 e^{-\frac{\frac{1}{2}+2y^2}{2\sigma^2 s^2}}\right) && (3.56)\\
&= \frac{1}{\sigma^2 s^2}\left(\frac{1}{\sqrt{4\pi\sigma^2 s^2}}\left(Q\left(\frac{-\sqrt{2}}{\sigma s}\right) - Q\left(\frac{\sqrt{2}}{\sigma s}\right)\right) - 2\frac{e^{-\frac{1}{4\sigma^2 s^2}}}{\sqrt{4\pi\sigma^2 s^2}}\left(Q\left(\frac{-\sqrt{2}}{\sigma s}\right) - Q\left(\frac{\sqrt{2}}{\sigma s}\right)\right)\right) && (3.57)\\
&= \frac{1}{(\sigma^2 s^2)^{3/2}}\left(\frac{1}{\sqrt{4\pi}}\right)\left(1 - 2e^{-\frac{1}{4\sigma^2 s^2}}\right)\left(1 - 2Q\left(\frac{\sqrt{2}}{\sigma s}\right)\right). && (3.58)
\end{aligned}$$

Therefore, $\lim_{s\to 0} J(sX+\varepsilon) = \infty$. From (3.50), we have that the low-rate slope is infinite.

□

For vector Gaussian sources, we analyze a formulation of dithered uniform quantization that is identical to the nonlinear transform coding approach (NTC) introduced in [7] with linear transforms and a factorized entropy model.

Let $W, T \in \mathbb{R}^{d\times d}$ be the encoder-decoder pair such that the encoder output $W^\top X$ is quantized with dither $\varepsilon \in \mathbb{R}^d$ where each component is an independent Unif$\left[-\frac{1}{2}, \frac{1}{2}\right]$ random variable. Applying the function $Q(.)$ to quantize each component of the dithered encoder output, the reconstruction is obtained by subtracting the dither from the quantized encoder output and passing that through the decoding transform T.

Theorem 45. *Let $X \sim \mathcal{N}(0, K)$, where $K \in \mathbb{R}^{d\times d}$ is a positive definite covariance matrix. The low-rate slope of dithered uniform quantization for X is infinite.*

Proof. For the above dithered uniform quantization formulation, [14] showed that the optimal linear encoder's nonzero rows, to quantize a vector Gaussian source, are eigenvectors of its covariance matrix, i.e., the optimal W is given by $W = US$ where S is a diagonal matrix with entries $s_i \geq 0$ and U is the normalized eigenvector matrix. The optimal T is the LLSE estimate of X given $W^\top X + \varepsilon$. Since the dither for each component is independent, the rate can be written as $\sum_{i=1}^{d} h(s_i u_i^\top X + \varepsilon_i)$ where $u_i^\top X \sim \mathcal{N}\left(0, \sigma_i^2\right)$. Similarly, the distortion is $\sum_{i=1}^{d} D(s_i)$ where $D(.)$ is defined as in (3.47). The low-rate slope in the limit $D \to \text{tr}(K)$ is obtained when each of the s_i's approach 0. Since,

$$\frac{\sum_{i=1}^{d} h\left(s_i u_i^\top X + \varepsilon_i\right)}{\text{tr}(K) - \sum_{i=1}^{d} D(s_i)} \geq d \min_i \frac{h\left(s_i u_i^\top X + \varepsilon_i\right)}{\sigma_i^2 - D(s_i)},$$

the low-rate slope of vector Gaussian source under dithered uniform quantization as formulated above is infinite. □

CHAPTER 4

PRINCIPAL BIT ANALYSIS: AUTOENCODING FOR SCHUR-CONCAVE LOSS

Autoencoders are an effective method for representation learning and dimensionality reduction. Given a centered dataset $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ (i.e., $\sum_i \boldsymbol{x}_i = 0$), an autoencoder (with *latent dimension* $k \leq d$) consists of an *encoder* $f : \mathbb{R}^d \mapsto \mathbb{R}^k$ and a *decoder* $g : \mathbb{R}^k \mapsto \mathbb{R}^d$. The goal is to select f and g from prespecified classes C_f and C_g respectively such that if a random point $\boldsymbol{x}$ is picked from the data set then $g(f(\boldsymbol{x}))$ is close to $\boldsymbol{x}$ in some sense, for example in mean squared error. If C_f and C_g consist of linear mappings then the autoencoder is called a *linear autoencoder*.

Autoencoders have achieved striking successes when f and g are selected through training from the class of functions realized by multilayer perceptrons of a given architecture [33]. Yet, the canonical autoencoder formulation described above has a notable failing, namely that for linear autoencoders, optimal choices of f and g do not necessarily identify the principal components of the dataset; they merely identify the principal subspace [18, 6]. That is, the components of $f(\boldsymbol{x})$ are not necessarily proportional to projections of $\boldsymbol{x}$ against the eigenvectors of the covariance matrix

$$\boldsymbol{K} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \cdot \boldsymbol{x}_i^\top, \tag{4.1}$$

which we assume without loss of generality is full rank. Thus, linear autoencoders do not recover Principal Component Analysis (PCA). The reason for this is that both the objective (the distortion) and the constraint (the dimensionality of the latents) are invariant to an invertible transformation applied after the encoder with its inverse applied before the decoder. It is desirable for linear autoencoders to recover PCA for two reasons. First, from a representation learning standpoint, it guarantees that the autoencoder recovers uncorrelated features. Second, since a conventional linear autoencoder has a large number of globally optimal solutions corresponding to different bases of the principal subspace, it is preferable to eliminate this indeterminism.

Autoencoders are sometimes described as "compressing" the data [57, 18, 41, 15], even though f can be invertible even when $k < d$. We show that by embracing this compression-view, one can obtain autoencoders that are able to recover PCA. Specifically, we consider linear autoencoders with quantized (or, equivalently, noisy) latent variables with a constraint on the estimated number of bits required to transmit the quantized latents under fixed-rate coding. We call this problem *Principal Bit Analysis (PBA)*. The constraint turns out to be a strictly Schur-concave function of the set of variances of the latent variables (see the supplementary for a review of Schur-concavity). Although finding the optimal f and g for this loss function is a nonconvex optimization problem, we show that for any strictly Schur-concave loss function, an optimal f must send projections of the data along the principal components, assuming that the empirical covariance matrix of the data has only simple eigenvalues. That is, imposing a strictly Schur-concave loss in place of a simple dimensionality constraint suffices to ensure recovery of PCA. The idea is that the strict concavity of the loss function eliminates the rotational invariance described above. As we show, even a slight amount of "curvature" in the constraint forces the autoencoder to spread the variances of the latents out as much as possible, resulting in recovery of PCA. If the loss function is merely Schur-concave, then projecting along the principal components is optimal, but not necessarily uniquely so.

Using this theorem, we can efficiently solve PBA. We validate the solution experimentally by using it to construct a fixed-rate compression algorithm for arbitrary vector-valued data sources. We find that the PBA-derived compressor beats existing linear, fixed-rate compressors both in terms of mean squared error, for which it is optimized, and in terms of the structural similarity index measure (SSIM) and downstream classification accuracy, for which it is not.

A number of variable-rate multimedia compressors have recently been proposed that are either related to, or directly inspired by, autoencoders [67, 65, 9, 64, 63, 56, 31, 3, 10, 74, 5, 7]. As a second application of our result, we show that for Gaussian sources, a linear form of such a compressor is guaranteed to recover PCA. Thus we show that ideas from compression can be

fruitfully fed back into the original autoencoder problem.

The contributions of this chapter are

- We propose a novel linear autoencoder formulation in which the constraint is Schur-concave. We show that this generalizes conventional linear autoencoding.
- If the constraint is strictly Schur-concave and the covariance matrix of the data has only simple eigenvalues, then we show that the autoencoder provably recovers PCA, providing a new remedy for a known limitation of linear autoencoders.
- We use the new linear autoencoder formulation to efficiently solve a fixed-rate compression problem that we call *Principal Bit Analysis (PBA).*
- We demonstate experimentally that PBA outperforms existing fixed-rate compressors on a variety of data sets and metrics.
- We show that a linear, variable-rate compressor that is representative of many autoencoder-based compressors in the literature effectively has a strictly Schur-concave loss, and therefore it recovers PCA.

4.1 Linear Autoencoding with a Schur-Concave Constraint

Throughout this chapter we consider C_f and C_g to be the class of linear functions. The functions $f \in C_f$ and $g \in C_g$ can then be represented by d-by-d matrices, respectively, which we denote by $\boldsymbol{W}$ and $\boldsymbol{T}$, respectively. Thus we have

$$f(\boldsymbol{x}) = \boldsymbol{W}^\top \boldsymbol{x} \tag{4.2}$$

$$g(\boldsymbol{x}) = \boldsymbol{T}\boldsymbol{x}. \tag{4.3}$$

We wish to design $\boldsymbol{W}$ and $\boldsymbol{T}$ to minimize the mean squared error when the latent variables $\boldsymbol{W}^\top \boldsymbol{x}$ are quantized, subject to a constraint on the number of bits needed to represent the quantized latents. We accomplish this via two modifications of the canonical autoencoder. First, we perturb the d latent variables with zero-mean additive noise with covariance matrix $\sigma^2 \boldsymbol{I}$, which we denote by $\boldsymbol{\varepsilon}$. Thus the input to the decoder is

$$\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{\varepsilon} \tag{4.4}$$

and our objective is to minimize the mean squared error

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\left\| \boldsymbol{x}_i - \boldsymbol{T} \left(\boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{\varepsilon} \right) \right\|_2^2 \right]. \tag{4.5}$$

This is equivalent to quantizing the latents, in the following sense [73]. Let $Q(\cdot)$ be the function that maps any real number to its nearest integer and ε be a random variable uniformly distributed over $[-1/2, 1/2]$. Then for X independent of ε, the quantities $Q(X+\varepsilon) - \varepsilon$ and $X + \varepsilon$ have the same joint distribution with X. Thus (4.5) is exactly the mean squared error if the latents are quantized to the nearest integer and $\sigma^2 = \frac{1}{12}$, assuming that the quantization is dithered. The overall system is depicted in Fig. 4.1.

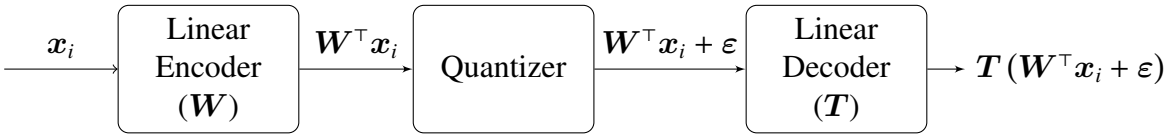


Figure 4.1: Compression Block Diagram

We wish to constrain the number of bits needed to describe the latent variables. We assume that the jth quantized latent is clipped to the interval

$$\left[-\frac{\sqrt{(2a)^2 \boldsymbol{w}_j^\top \boldsymbol{K} \boldsymbol{w}_j + 1}}{2}, \frac{\sqrt{(2a)^2 \boldsymbol{w}_j^\top \boldsymbol{K} \boldsymbol{w}_j + 1}}{2} \right],$$

where $a > 0$ is a hyperparameter and the covariance matrix $\boldsymbol{K}$ is as defined in (4.1). The idea is that for sufficiently large a, the interval

$$\left(-a\sqrt{\boldsymbol{w}_j^\top \boldsymbol{K}\boldsymbol{w}_j}, a\sqrt{\boldsymbol{w}_j^\top \boldsymbol{K}\boldsymbol{w}_j}\right]$$

contains the latent with high probability, and adding 1 accounts for the expansion due to the dither. The number of bits needed for the jth latent is then

$$\log\left(\sqrt{4a^2\boldsymbol{w}_j^\top \boldsymbol{K}\boldsymbol{w}_j + 1}\right) = \frac{1}{2}\log\left(4a^2\boldsymbol{w}_j^\top \boldsymbol{K}\boldsymbol{w}_j + 1\right). \tag{4.6}$$

We arrive at our optimization problem:

$$\begin{aligned} \inf_{\boldsymbol{W},\boldsymbol{T}} \quad & \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\left\|\boldsymbol{x}_i - \boldsymbol{T}\left(\boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{\varepsilon}\right)\right\|_2^2\right] \\ \text{subject to} \quad & R \geq \sum_{j=1}^{d} \frac{1}{2}\log\left(4a^2\boldsymbol{w}_i^\top \boldsymbol{K}\boldsymbol{w}_i + 1\right). \end{aligned} \tag{4.7}$$

Note that the function

$$\{\boldsymbol{w}_j^\top \boldsymbol{K}\boldsymbol{w}_j\}_{j=1}^d \mapsto \sum_{j=1}^{d} \frac{1}{2}\log\left(4a^2\boldsymbol{w}_i^\top \boldsymbol{K}\boldsymbol{w}_i + 1\right)$$

is strictly Schur-concave (see Appendix A for a brief review of Schur-concavity). Our first result only requires that the constraint is Schur-concave in the set of latent variances, so we will consider the more general problem

$$\begin{aligned} \inf_{\boldsymbol{W},\boldsymbol{T}} \quad & \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\left\|\boldsymbol{x}_i - \boldsymbol{T}\left(\boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{\varepsilon}\right)\right\|_2^2\right] \\ \text{subject to} \quad & R \geq \rho\left(\left\{\boldsymbol{w}_j^\top \boldsymbol{K}\boldsymbol{w}_j\right\}_{j=1}^d\right) \end{aligned} \tag{4.8}$$

where $\rho(\cdot)$ is any Schur-concave function.

Expressing the objective in (4.8) in terms of $\boldsymbol{K}$, the optimization problem reduces to

$$\begin{aligned} \inf_{\boldsymbol{W},\boldsymbol{T}} \quad & \operatorname{tr}(\boldsymbol{K}) - 2\operatorname{tr}(\boldsymbol{K}\boldsymbol{W}\boldsymbol{T}^\top) + \operatorname{tr}\left(\boldsymbol{T}\left(\boldsymbol{W}^\top \boldsymbol{K}\boldsymbol{W} + \sigma^2\boldsymbol{I}\right)\boldsymbol{T}^\top\right) \\ \text{subject to} \quad & R \geq \rho\left(\left\{\boldsymbol{w}_j^\top \boldsymbol{K}\boldsymbol{w}_j\right\}_{j=1}^d\right). \end{aligned} \tag{4.9}$$

Since $\boldsymbol{T}$ does not appear in the rate constraint, the optimal $\boldsymbol{T}$ can be viewed as the Linear Least Squares Estimate (LLSE) of a random $\boldsymbol{x}$ given $\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{\varepsilon}$. Therefore, the optimal decoder, $\boldsymbol{T}^*$ for a given encoder $\boldsymbol{W}$ is (e.g. [37]):

$$\boldsymbol{T}^* = \boldsymbol{K}\boldsymbol{W}(\boldsymbol{W}^\top \boldsymbol{K}\boldsymbol{W} + \sigma^2 \boldsymbol{I})^{-1}. \tag{4.10}$$

Substituting for $\boldsymbol{T}$ in (4.9) yields an optimization problem over only $\boldsymbol{W}$

$$\begin{aligned} \inf_{\boldsymbol{W}} \quad & \operatorname{tr}(\boldsymbol{K}) - \operatorname{tr}(\boldsymbol{K}\boldsymbol{W}(\boldsymbol{W}^\top \boldsymbol{K}\boldsymbol{W} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{W}^\top \boldsymbol{K}) \\ \text{subject to} \quad & R \geq \rho\left(\left\{\boldsymbol{w}_j^\top \boldsymbol{K}\boldsymbol{w}_j\right\}_{j=1}^d\right). \end{aligned} \tag{4.11}$$

This problem is nonconvex in general. In the following subsection, we prove a structural result about the problem for a Schur-concave ρ. Namely, we show that the nonzero rows of $\boldsymbol{W}$ must be eigenvectors of $\boldsymbol{K}$. In Section 4.2, we solve the problem for the specific choice of ρ in (4.7). We also show how this generalizes conventional linear autoencoders.

4.1.1 Optimal Autoencoding with a Schur-Concave Constraint

The following is the main theoretical result of this chapter.

Theorem 46. *For Schur-concave $\rho : \mathbb{R}^d_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $R > 0$, the set of matrices whose nonzero columns are eigenvectors of the covariance matrix $\boldsymbol{K}$ is optimal for* (4.11)*. If ρ is strictly Schur-concave and $\boldsymbol{K}$ contains distinct eigenvalues, this set contains all optimal solutions of* (4.11)*.*

Proof. Let the eigenvalues of $\boldsymbol{K}$ be $\{\sigma_i^2\}_{i=1}^d$ with $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_d^2$. Let the eigendecomposition of $\boldsymbol{K}$ be given by $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$ where $\boldsymbol{U}$ is an orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{K}$ and $\boldsymbol{\Sigma}$ is a diagonal matrix with entries $\left\{\sigma_i^2\right\}_{i=1}^d$.

We first prove that the optimal value of (4.11) can be achieved by a $\boldsymbol{W}$ such that $\boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W}$ is a diagonal matrix. Let $\widetilde{\boldsymbol{W}} = \boldsymbol{W}\boldsymbol{Q}$ where $\boldsymbol{Q}$ is the orthogonal matrix obtained from the eigendecomposition of $\boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W}$ i.e.,

$$\boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W} = \boldsymbol{Q} \boldsymbol{\Lambda} \boldsymbol{Q}^\top,$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix formed from the eigenvalues of $\boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W}$. Note that

$$\begin{aligned}\operatorname{tr}\left(\boldsymbol{K}\widetilde{\boldsymbol{W}}\left(\widetilde{\boldsymbol{W}}^\top \boldsymbol{K}\widetilde{\boldsymbol{W}} + \sigma^2 \boldsymbol{I}\right)^{-1} \widetilde{\boldsymbol{W}}^\top \boldsymbol{K}\right) &= \operatorname{tr}\left(\boldsymbol{K}\boldsymbol{W}\boldsymbol{Q}\left(\boldsymbol{\Lambda} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{Q}^\top \boldsymbol{W}^\top \boldsymbol{K}\right) \\ &= \operatorname{tr}\left(\boldsymbol{K}\boldsymbol{W}\left(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top + \sigma^2 \boldsymbol{Q}\boldsymbol{Q}^\top\right)^{-1} \boldsymbol{W}^\top \boldsymbol{K}\right).\end{aligned}$$

Since $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top = \boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W}$ and $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}$, the objective remains the same. We now show that the constraint is only improved. Denoting the eigenvalues of $\boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W}$ by $\left\{\nu_j\right\}_{j=1}^d$, we have

$$\rho\left(\left\{\widetilde{\boldsymbol{w}}_j^\top \boldsymbol{K} \widetilde{\boldsymbol{w}}_j\right\}_{j=1}^d\right) = \rho\left(\left\{\boldsymbol{q}_j^\top \boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W} \boldsymbol{q}_j\right\}_{j=1}^d\right) = \rho\left(\left\{\nu_j\right\}_{j=1}^d\right).$$

Now since the eigenvalues of a Hermitian matrix majorize its diagonal elements by the Schur-Horn theorem [34, Theorem 4.3.45],

$$\left\{\boldsymbol{w}_j^\top \boldsymbol{K} \boldsymbol{w}_j\right\}_{j=1}^d \prec \left\{\nu_j\right\}_{j=1}^d.$$

Since ρ is Schur-concave, this implies

$$\rho\left(\left\{\boldsymbol{w}_j^\top \boldsymbol{K} \boldsymbol{w}_j\right\}_{j=1}^d\right) \geq \rho\left(\left\{\nu_j\right\}_{j=1}^d\right) = \rho\left(\left\{\widetilde{\boldsymbol{w}}_j^\top \boldsymbol{K} \widetilde{\boldsymbol{w}}_i\right\}_{j=1}^d\right).$$

Therefore, if ρ is Schur-concave, the rate constraint can only improve. This implies an optimal solution can be attained when $\boldsymbol{W}$ is such that $\boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W}$ is diagonal. If ρ is strictly Schur-concave, the rate constraint strictly improves implying that the optimal $\boldsymbol{W}$ must be such that $\boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W}$ is diagonal. This implies that

$$\begin{aligned}\operatorname{tr}\left(\boldsymbol{K}\boldsymbol{W}\left(\boldsymbol{W}^\top \boldsymbol{K}\boldsymbol{W} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{K}\right) &= \operatorname{tr}\left(\boldsymbol{W}^\top \boldsymbol{K}^2 \boldsymbol{W}\left(\boldsymbol{W}^\top \boldsymbol{K}\boldsymbol{W} + \sigma^2 \boldsymbol{I}\right)^{-1}\right) \\ &= \sum_{i=1}^d \frac{\boldsymbol{w}_i^\top \boldsymbol{K}^2 \boldsymbol{w}_i}{\sigma^2 + \boldsymbol{w}_i^\top \boldsymbol{K} \boldsymbol{w}_i}.\end{aligned}$$

Note that minimizing the objective in (4.11) is equivalent to maximizing the above expression. Perform the change of variable

$$\begin{aligned}\boldsymbol{w}_j \mapsto &\begin{cases}\left(\frac{\boldsymbol{K}^{1/2}\boldsymbol{w}_j}{\|\boldsymbol{K}^{1/2}\boldsymbol{w}_j\|}, \|\boldsymbol{K}^{1/2}\boldsymbol{w}_j\|^2\right) & \text{if } \boldsymbol{K}^{1/2}\boldsymbol{w}_j \neq \boldsymbol{0} \\ (\boldsymbol{0}, 0) & \text{if } \boldsymbol{K}^{1/2}\boldsymbol{w}_j = \boldsymbol{0}\end{cases} \\ &= (y_j, y_j).\end{aligned}$$

The assumption that $\boldsymbol{W}^\top \boldsymbol{K}\boldsymbol{W}$ is diagonal and the normalization in the definition of y_j implies that

$$\boldsymbol{Y} = [y_1 y_2, \cdots, y_d]$$

is a matrix whose nonzero columns form an orthonormal set. Rewriting the objective in terms of the (y_j, y_j), we have

$$\sum_{i=1}^{d} \frac{\boldsymbol{w}_i^\top \boldsymbol{K}^2 \boldsymbol{w}_i}{\sigma^2 + \boldsymbol{w}_i^\top \boldsymbol{K}\boldsymbol{w}_i} = \sum_{i=1}^{d} y_i^\top \boldsymbol{K} y_i \frac{y_i}{\sigma^2 + y_i} = \sum_{i=1}^{d} y_i^\top \boldsymbol{K} y_i m_i, \tag{4.12}$$

where $m_i = \frac{y_i}{\sigma^2 + y_i}$. Observe that under this new parametrization, the constraint only depends on $\{y_i\}_{i=1}^d$. Without loss of generality, we assume that $y_1 \geq y_2 \geq \cdots \geq y_d$, implying that $m_1 \geq m_2 \geq \cdots \geq m_d$. We now prove that for given $\{y_i\}_{i=1}^d$, choosing the y_i along the eigenvectors of $\boldsymbol{K}$ is optimal.

Denote the diagonal elements of $\boldsymbol{Y}^\top \boldsymbol{K}\boldsymbol{Y}$ by $\{\lambda_i^2\}_{i=1}^d$ and let $\{\lambda_{i,\downarrow}^2\}_{i=1}^d$ denote the same diagonal elements arranged in descending order. Denote the eigenvalues of $\boldsymbol{Y}^\top \boldsymbol{K}\boldsymbol{Y}$ by $\{\mu_i^2\}_{i=1}^d$ where $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_d$. Again invoking the Schur-Horn theorem, the eigenvalues of $\boldsymbol{Y}^\top \boldsymbol{K}\boldsymbol{Y}$ majorize its diagonal entries

$$\left\{\lambda_i^2\right\}_{i=1}^d \prec \left\{\mu_i^2\right\}_{i=1}^d. \tag{4.13}$$

Substituting $\lambda_i^2 = y_i^\top \boldsymbol{K} y_i$ in (4.12), we have

$$\begin{aligned}
\sum_{i=1}^{d} \lambda_i^2 m_i \overset{(a)}{\leq} \sum_{i=1}^{d} \lambda_{i,\downarrow}^2 m_i &= \lambda_{1,\downarrow}^2 m_1 + \sum_{i=2}^{d} \left(\sum_{j=1}^{i} \lambda_{j,\downarrow}^2 - \sum_{j=1}^{i-1} \lambda_{j,\downarrow}^2 \right) m_i \\
&= \lambda_{1,\downarrow}^2 m_1 + \sum_{i=2}^{d} m_i \sum_{j=1}^{i} \lambda_{j,\downarrow}^2 - \sum_{i=2}^{d} m_i \sum_{j=1}^{i-1} \lambda_{j,\downarrow}^2 \\
&= \lambda_{1,\downarrow}^2 (m_1 - m_2) + m_d \left(\sum_{j=1}^{d} \lambda_{j,\downarrow}^2 \right) + \sum_{i=2}^{d-1} (m_i - m_{i+1}) \sum_{j=1}^{i} \lambda_{j,\downarrow}^2 \\
&\overset{(b)}{\leq} \mu_1^2 (m_1 - m_2) + m_d \left(\sum_{j=1}^{d} \mu_j^2 \right) + \sum_{i=2}^{d-1} (m_i - m_{i+1}) \sum_{j=1}^{i} \mu_j^2 \\
&\overset{(c)}{\leq} \sigma_1^2 (m_1 - m_2) + m_d \left(\sum_{j=1}^{d} \sigma_j^2 \right) + \sum_{i=2}^{d-1} (m_i - m_{i+1}) \sum_{j=1}^{i} \sigma_j^2 \\
&= \sum_{i=1}^{d} \sigma_i^2 m_i,
\end{aligned}$$

where inequality (a) follows from the assumption that $m_1 \geq m_2 \geq \cdots \geq m_d$, and (b) from the definition in (4.13). Since $\boldsymbol{Y}$'s nonzero columns form an orthonormal set, the eigenvalues of $\boldsymbol{Y}^\top \boldsymbol{K} \boldsymbol{Y}$, when arranged in descending order, are at most the eigenvalues of $\boldsymbol{K}$ from Corollary 4.3.37 in [34], and therefore (c) follows.

This upper bound is attained when $y_i = \boldsymbol{u}_i$ for nonzero y_i, where $\boldsymbol{u}_i$ is the normalized eigenvector of $\boldsymbol{K}$ corresponding to eigenvalue σ_i^2. To see this, note that when $y_i = \boldsymbol{u}_i$, $\lambda_i^2 = \mu_i^2 = \sigma_i^2$. From the definition of $y_i, \boldsymbol{w}_i = \boldsymbol{K}^{-1/2} \boldsymbol{u}_i \sqrt{y_i} = \boldsymbol{u}_i \frac{\sqrt{y_i}}{\sigma_i}$. Therefore, for a Schur-concave ρ, the set of matrices whose nonzero columns are eigenvectors of $\boldsymbol{K}$ is optimal. We now prove that for a strictly Schur-concave ρ, if $\boldsymbol{K}$ has distinct eigenvalues, this set contains all of the optimal solutions $\boldsymbol{W}$.

We know that for a fixed $y_1 \geq y_2 \geq \cdots \geq y_d$, (implying a fixed $m_1 \geq m_2 \geq \cdots \geq m_d$) the upper bound $\sum_{i=1}^{d} \sigma_i^2 m_i$ is attained by the previous choice of y_i. Note that if all nonzero m_i are distinct, equality in (b) and (c) is attained if and only if the nonzero diagonal elements of $\boldsymbol{Y}^\top \boldsymbol{K} \boldsymbol{Y}$ equal

the corresponding eigenvalues of $\boldsymbol{K}$. This implies that, if all nonzero m_i are distinct, the upper bound is attained if and only if $y_i = \boldsymbol{u}_i$ for nonzero y_i. Therefore, it is sufficient to prove that for the following optimization problem

$$\begin{aligned} \sup_{\{y_i \geq 0\}} \quad & \sum_{i=1}^{d} \sigma_i^2 \frac{y_i}{\sigma^2 + y_i} \\ \text{subject to} \quad & R \geq \rho\left(\{y_i\}_{i=1}^d\right), \end{aligned} \tag{4.14}$$

any optimal $\{y_i\}$ must be such that the nonzero y_i are distinct. Firstly, note that since $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_d^2$, we must have $y_1 \geq y_2 \geq \cdots \geq y_d$. Assume to the contrary that for an optimal $\{y_i\}_{i=1}^d$ there exists $1 \leq j, \ell < d$ such that $y_{j-1} > y_j = y_{j+1} = y_{j+2} = \cdots = y_{j+\ell} > y_{j+\ell+1} \geq 0$, where y_0 is chosen to be any real number strictly greater than y_1 and $y_{d+1} = 0$. Take $\delta > 0$ small. Denote a new sequence $\{y_i'\}_{i=1}^d$ where $y_j' = y_j + \delta, y_{j+\ell}' = y_{j+\ell} - \delta$ and $y_i' = y_i$ for $1 \leq i \leq d$ with $i \neq j$ and $j + \ell$. Since ρ is strictly Schur-concave, the constraint is strictly improved,

$$\rho\left(\{y_i'\}_{i=1}^d\right) < \rho\left(\{y_i\}_{i=1}^d\right).$$

Since $\sigma_j^2 > \sigma_{j+\ell}^2$, the objective is strictly improved for sufficiently small δ,

$$\sum_{i=1}^{d} \sigma_i^2 \frac{y_i}{\sigma^2 + y_i} < \sum_{i=1}^{d} \sigma_i^2 \frac{y_i'}{\sigma^2 + y_i'},$$

as desired. □

As a consequence of Theorem 46, encoding via an optimal $\boldsymbol{W}$ can be viewed as a projection along the eigenvectors of $\boldsymbol{K}$, followed by different scalings applied to each component, i.e. $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{S}$ where $\boldsymbol{S}$ is a diagonal matrix with entries $s_i \geq 0$ and $\boldsymbol{U}$ is the normalized eigenvector matrix. Only $\boldsymbol{S}$ remains to be determined, and to this end, we may assume that $\boldsymbol{K}$ is diagonal with nonincreasing diagonal entries, implying $\boldsymbol{U} = \boldsymbol{I}$. In subsequent sections, our choice of ρ will be

of the form $\sum_{i=1}^{d} \rho_{sl}$, where $\rho_{sl} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$[1] is (strictly) concave, making ρ (strictly) Schur-concave (see Proposition 54 in Appendix A). Therefore, (4.11) reduces to

$$\begin{aligned} \inf_{\boldsymbol{S}} \quad & \operatorname{tr}(\boldsymbol{K}) - \operatorname{tr}(\boldsymbol{K}\boldsymbol{S}(\boldsymbol{S}^\top \boldsymbol{K}\boldsymbol{S} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{S}^\top \boldsymbol{K}) \\ \text{subject to} \quad & R \geq \rho_{sl}\left(\{s_i^2 \sigma_i^2\}\right), \end{aligned} \tag{4.15}$$

where the infimum is over diagonal matrices $\boldsymbol{S}$. To handle situations for which

$$\lim_{s \to \infty} \rho_{sl}(s) < \infty, \tag{4.16}$$

we allow the diagonal entries of $\boldsymbol{S}$ to be ∞, with the objective for such cases defined via its continuous extension.

In the next section, we will solve (4.15) for several specific choices of ρ_{sl}.

4.2 Explicit Solutions: Conventional Linear Autoencoders and PBA

4.2.1 Conventional Linear Autoencoders

Given a centered dataset $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n \in \mathbb{R}^d$, consider a linear autoencoder optimization problem where the encoder and decoder, $\boldsymbol{W}$ and $\boldsymbol{T}$, respectively, are d-by-k matrices where $k \leq d$ is a parameter. The goal is to minimize the mean squared error as given by (4.5). PCA corresponds to the global optimal solution of this optimization problem, where $\boldsymbol{W} = \boldsymbol{T} = \boldsymbol{U}_k$, where $\boldsymbol{U}_k \in \mathbb{R}^{d \times k}$ is a matrix whose columns are the k eigenvectors corresponding to the k largest eigenvalues of $\boldsymbol{K}$. However, there are multiple global optimal solutions, given by any encoder-decoder pair of the form $(\boldsymbol{U}_k \boldsymbol{V}, \boldsymbol{U}_k \boldsymbol{V})$, where $\boldsymbol{V}$ is an orthogonal matrix [6].

[1] "sl" stands for single-letter

We now recover linear autoencoders through our framework in Section 4.1. Consider the optimization problem in (4.15) where $\rho_{sl} : \mathbb{R}_{\geq 0} \to \{0, 1\}$ is a concave function defined as

$$\rho_{sl}(x) = \mathbf{1}\left[x > 0\right]. \tag{4.17}$$

Note that this penalizes the dimension of the latents, as desired. Note also that this cost is Schur-concave but not strictly so. The fact that PCA solves conventional linear autoencoding, but is not necessarily the unique solution, follows immediately from Theorem 46.

Theorem 47. *If $\rho_{sl}(\cdot)$ is given by* (4.17)*, then an optimal solution for* (4.15) *is given by a diagonal matrix $\boldsymbol{S}$ whose top* $\min(\lfloor R \rfloor, d)$ *diagonal entries are equal to ∞ and the remaining entries are* 0*.*

Proof. Let $\mathcal{F} \stackrel{\text{def}}{=} \{i \in [d] : s_i > 0\}$, implying $|\mathcal{F}| \leq R$. Since $\boldsymbol{K}$ and $\boldsymbol{S}$ are diagonal, the optimization problem in (4.15) can be written as

$$\begin{aligned} &\inf_{\{s_\ell\}} \sum_{j \in [d] \setminus \mathcal{F}} \sigma_j^2 + \sum_{\ell \in \mathcal{F}} \frac{\sigma^2 \sigma_\ell^2}{\sigma^2 + \sigma_\ell^2 s_\ell^2} \\ &\text{subject to} \quad R \geq \sum_{i=1}^{d} \mathbf{1}\left[s_i > 0\right]. \end{aligned} \tag{4.18}$$

Since the value of $s_\ell, \ell \in \mathcal{F}$ does not affect the rate constraint, each of the s_ℓ can be made as large as possible without changing the rate constraint. Therefore, the infimum value of the objective is $\sum_{j \in [d] \setminus \mathcal{F}} \sigma_j^2$. Since we seek to minimize the distortion, the optimal $\mathcal{F}$ is the set of indices with the largest $|\mathcal{F}|$ eigenvalues. Since the number of these eigenvalues cannot exceed R, we choose $|\mathcal{F}| = \min(\lfloor R \rfloor, d)$. □

Unlike the conventional linear autoencoder framework, in Section 4.1, the latent variables $\boldsymbol{W}^\top \boldsymbol{x}$ are quantized, which we model with additive white noise of fixed variance. Therefore,

an infinite value of s_i indicates sending $\boldsymbol{u}_i^\top \boldsymbol{x}$ with full precision where $\boldsymbol{u}_i$ is the eigenvector corresponding to the i^{th} largest eigenvalue. This implies that PCA with parameter k corresponds to $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{S}$, where $\boldsymbol{S}$ is a diagonal matrix whose top k diagonal entries are equal to ∞ and the $d-k$ remaining diagonal entries are 0. Therefore, for any R such that $\lfloor R \rfloor = k$, an optimal solution to (4.15) corresponds to linearly projecting the data along the top k eigenvectors, which is the same as PCA. Note that, like [6], we only prove that projecting along the eigenvectors is one of possibly other optimal solutions. However, even a slight amount of curvature in ρ would make it strictly Schur-concave, thus recovering the principal directions. We next turn to a specific cost function with curvature, namely the PBA cost function that was our original motivation.

4.2.2 Principal Bit Analysis (PBA)

Consider the choice of $\rho_{sl} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ that provided the original impetus for Theorem 46. For $\gamma > \frac{2}{\sigma^2}$,

$$\rho_{sl}(x) = \frac{1}{2}\log(\gamma x + 1). \tag{4.19}$$

The nature of the optimization problem depends on the value of γ. For $1 \leq \gamma\sigma^2 \leq 2$, the problem can be made convex with a simple change of variable. For $\gamma\sigma^2 = 1$, the problem coincides with the classical waterfilling procedure in rate-distortion theory, in fact. For $\gamma\sigma^2 > 2$, the problem is significantly more challenging. Since we are interested in relatively large values of γ for our compression application (see Section 4.4 to follow), we focus on the case $\gamma > 2/\sigma^2$.

Theorem 48. *If $\rho_{sl}(\cdot)$ is given by* (4.19) *for $\gamma > \frac{2}{\sigma^2}$, then for any $\lambda > 0$, the pair $\bar{R}_{opt}, \bar{D}_{opt}$ obtained*

Algorithm 1 Principal Bit Analysis (PBA)

Require: $\lambda > 0$, $\alpha > 2$,

$$\boldsymbol{K} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix} \succ 0, \tag{4.20}$$

such that $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_d^2$.

If $\lambda \geq \sigma_1^2/(4(\alpha-1))$, Output $\bar{R}_{\text{opt}} = 0, \bar{D}_{\text{opt}} = \sum_{i=1}^{d} \sigma_i^2$.

Set $\bar{d} = \max\left\{i : \lambda < \sigma_i^2/4(\alpha-1)\right\}$.

Set $\bar{R}$, $\bar{D}$ to zero arrays of size $2\bar{d}$.

for $r \in \left\{1, 2, \cdots \bar{d}\right\}$ **do**

$$\bar{D}(2r-1) = \sum_{i=1}^{r} \frac{\sigma_i^2}{2(\alpha-1)}\left(1 - \sqrt{1 - \frac{4\lambda(\alpha-1)}{\sigma_i^2}}\right) + \sum_{i=r+1}^{d} \frac{\sigma_i^2}{\alpha},$$

$$\bar{R}(2r-1) = \sum_{i=1}^{r} \frac{1}{2}\log\left(\frac{\sigma_i^2}{4\lambda}\right) + \log\left(1 + \sqrt{1 - \frac{4\lambda(\alpha-1)}{\sigma_i^2}}\right).$$

$$\bar{D}(2r) = \left(\sum_{i=1}^{r-1} \frac{\sigma_i^2}{2(\alpha-1)}\left(1 - \sqrt{1 - \frac{4\lambda(\alpha-1)}{\sigma_i^2}}\right) + \frac{\sigma_r^2}{2(\alpha-1)}\left(1 + \sqrt{1 - \frac{4\lambda(\alpha-1)}{\sigma_r^2}}\right) + \sum_{i=r+1}^{d} \frac{\sigma_i^2}{\alpha}\right).$$

$$\bar{R}(2r) = \sum_{i=1}^{r} \frac{1}{2}\log\left(\frac{\sigma_i^2}{4\lambda}\right) + \sum_{i=1}^{r-1} \log\left(1 + \sqrt{1 - \frac{4\lambda(\alpha-1)}{\sigma_i^2}}\right) + \log\left(1 - \sqrt{1 - \frac{4\lambda(\alpha-1)}{\sigma_r^2}}\right).$$

end for

$r^* \leftarrow \arg\min_{j \in \left[2\bar{d}\right]} \bar{D}(j) + \lambda \bar{R}(j)$.

Output $\bar{R}_{\text{opt}} = \bar{R}(r^*), \bar{D}_{\text{opt}} = \bar{D}(r^*)$.

from the output of Algorithm 1 satisfies

$$\bar{D}_{opt} + \lambda \bar{R}_{opt} = \inf_{\boldsymbol{S}} \quad tr(\boldsymbol{K}) - tr(\boldsymbol{K}\boldsymbol{S}(\boldsymbol{S}^\top \boldsymbol{K}\boldsymbol{S} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{S}^\top \boldsymbol{K}) + \lambda \sum_{i=1}^{d} \rho_{sl}\left(\{s_i^2 \sigma_i^2\}\right), \tag{4.21}$$

Proof. Since $\boldsymbol{K}$ and $\boldsymbol{S}$ are diagonal, the optimization problem in (4.21) can be written as

$$\inf_{\{s_i\}} \quad \sum_{i=1}^{d} \frac{\sigma^2 \sigma_i^2}{\sigma^2 + s_i^2 \sigma_i^2} + \lambda \cdot \frac{1}{2} \sum_{i=1}^{d} \log\left(1 + \gamma s_i^2 \sigma_i^2\right). \tag{4.22}$$

With the following change of variables $\alpha = \gamma\sigma^2$, $s_i \mapsto s_i'^2 = \alpha \frac{s_i'^2}{\sigma^2}$, we obtain

$$\inf_{\{s_i'\}} \quad \sum_{i=1}^{d} \alpha \frac{\sigma_i^2}{\alpha + s_i'^2 \sigma_i^2} + \lambda \cdot \frac{1}{2} \sum_{i=1}^{d} \log\left(1 + s_i'^2 \sigma_i^2\right). \tag{4.23}$$

Ignoring the constant factor in the objective, perform the change of variable $s_i' \mapsto D_i = \frac{\sigma_i^2}{\alpha + s_i'^2 \sigma_i^2}$ to obtain

$$\begin{aligned} \inf_{\{D_i\}} \quad & \sum_{i=1}^{d} D_i + \frac{\lambda}{2} \sum_{i=1}^{d} \log\left(\frac{\sigma_i^2}{D_i} - (\alpha - 1)\right), \\ \text{subject to} \quad & D_i \le \frac{\sigma_i^2}{\alpha} \qquad \text{for all } i \in [d]\,. \end{aligned} \tag{4.24}$$

This optimization problem is nonconvex since the function $\log\left(\frac{\sigma_i^2}{D_i} - (\alpha - 1)\right)$ is convex for $0 \le D_i \le \frac{\sigma_i^2}{2(\alpha-1)}$ but concave for $\frac{\sigma_i^2}{2(\alpha-1)} < D_i \le \frac{\sigma_i^2}{\alpha}$ and the latter interval is nonempty since $\alpha > 2$.

Any optimizing $\{D_i\}$ must be a stationary point of

$$\mathcal{L}\left(\{D_i\}_{i=1}^d, \lambda, \{\mu_i\}_{i=1}^d\right) = \sum_{i=1}^{d} D_i + \lambda\left(\sum_{i=1}^{d} \log\left(\frac{\sigma_i^2}{D_i} - (\alpha - 1)\right)\right) + \sum_{i=1}^{d} \mu_i \left(D_i - \frac{\sigma_i^2}{\alpha}\right). \tag{4.25}$$

for some $\{\mu_i\}_{i=1}^d$ with $\mu_i \ge 0$ for all $i \in [d]$ and satisfying the complementary slackness condition [52, Prop. 3.3.1]. The stationary points satisfy, for each i

$$\frac{\partial \mathcal{L}}{\partial D_i} = 1 - \lambda\left(\frac{\frac{\sigma_i^2}{D_i^2}}{\left(\frac{\sigma_i^2}{D_i} - (\alpha - 1)\right)}\right) + \mu_i = 0. \tag{4.26}$$

Let $\mathcal{F} = \left\{i : D_i < \frac{\sigma_i^2}{\alpha}\right\}$. For $i \in \mathcal{F}$, $\mu_i = 0$ due to complementary slackness. Substituting in (4.26) we obtain a quadratic equation in D_i

$$(\alpha - 1)D_i^2 - \sigma_i^2 D_i + \lambda \sigma_i^2 = 0.$$

which gives

$$D_i = \frac{\sigma_i^2}{2(\alpha - 1)}\left(1 \pm \sqrt{1 - \frac{4\lambda(\alpha - 1)}{\sigma_i^2}}\right).$$

Let $c_i = \sqrt{1 - \frac{4\lambda(\alpha-1)}{\sigma_i^2}}$. Note that $\frac{\sigma_i^2}{2(\alpha-1)}(1 + c_i)$ is always in the concave region and $\frac{\sigma_i^2}{2(\alpha-1)}(1 - c_i)$ is always in the convex region for a λ chosen such that D_i is a real number strictly less than $\frac{\sigma_i^2}{\alpha}$.

Therefore the optimal set of distortions are contained in the following set of 3^d points

$$\prod_{i=1}^{d}\left\{\frac{\sigma_i^2}{2(\alpha-1)}\left(1+\sqrt{1-\frac{4\lambda(\alpha-1)}{\sigma_i^2}}\right),\frac{\sigma_i^2}{2(\alpha-1)}\left(1-\sqrt{1-\frac{4\lambda(\alpha-1)}{\sigma_i^2}}\right),\frac{\sigma_i^2}{\alpha}\right\}.$$

We now reduce the size of the above set by making a two observations:

(1). $\mathcal{F}$ is contiguous.

Lemma 49. *There exists an optimal* $\left\{D_i^*\right\}_{i=1}^d$ *for* (4.24) *such that (a)* $\frac{\sigma_i^2}{D_i^*}$ *is a nonincreasing sequence and (b)* $\mathcal{F}=\{1,2,\cdots|\mathcal{F}|\}$.

Proof. Substitute $x_i=\frac{\sigma_i^2}{D_i}$ in (4.24). This gives us

$$\begin{aligned}\inf_{\{x_i\}}\quad & \sum_{i=1}^{d}\frac{\sigma_i^2}{x_i}+\frac{\lambda}{2}\sum_{i=1}^{d}\log\left(x_i-(\alpha-1)\right),\\ \text{subject to}\quad & x_i\geq\alpha\qquad\text{for all } i\in[d].\end{aligned}\tag{4.27}$$

Let $\{x_i^*\}_{i=1}^d$ be an optimal solution for (4.27). If, for $i>j$, $x_i^*>x_j^*\geq\alpha$, then exchanging the values provides a solution that has the same rate and lower distortion since $\frac{\sigma_i^2}{x_i^*}+\frac{\sigma_j^2}{x_j^*}\geq\frac{\sigma_i^2}{x_j^*}+\frac{\sigma_j^2}{x_i^*}$. This proves (a). Part (b) follows immediately. □

(2). No two solutions are concave.

Lemma 50. *For* $R>0$, *let* $\left\{D_i^*\right\}_{i=1}^d$ *be an optimal solution for* (4.24). *There exists at most one* D_i^* *such that* $\frac{\sigma_i^2}{2(\alpha-1)}<D_i^*<\frac{\sigma_i^2}{\alpha}$.

Proof. Let D_i^*, D_j^* be such that $\frac{\sigma_i^2}{2(\alpha-1)}<D_i^*<\frac{\sigma_i^2}{\alpha}$ and $\frac{\sigma_j^2}{2(\alpha-1)}<D_j^*<\frac{\sigma_j^2}{\alpha}$. Without loss of generality, assume $D_i^*<D_j^*$. Denote the individual rate constraint function by $r(D_i)\triangleq\log\left(\frac{\sigma_i^2}{D_i}-(\alpha-1)\right)$. Since r is concave in $\left(\frac{\sigma_i^2}{2(a-1)},\frac{\sigma_i^2}{\alpha}\right)$, there exist an $\varepsilon>0$ such that

$$r\left(D_i^* - \varepsilon\right) + r\left(D_j^* + \varepsilon\right) = r\left(D_i^*\right) - \varepsilon r'\left(D_i^*\right) + O(\varepsilon^2) + r\left(D_j^*\right) + \varepsilon r'\left(D_j^*\right) + O(\varepsilon^2) \tag{4.28}$$

$$< r\left(D_i^*\right) + r\left(D_j^*\right) \tag{4.29}$$

The last inequality follows from concavity of r. Therefore, replacing $\left(D_i^*, D_j^*\right)$ with $\left(D_i^* - \varepsilon, D_j^* + \varepsilon\right)$, the rate constraint can be improved while keeping the objective in (4.24) constant, contradicting the optimality assumption of $\left\{D_i^*\right\}$. □

There is at most one D_i^* such that $D_i^* = \frac{\sigma_i^2}{2(\alpha-1)}\left(1 + c_i\right)$. Assuming such an i exists, $x_i = \frac{2(\alpha-1)}{1+c_i} < 2(\alpha - 1)$. For the convex roots, $x_i = \frac{2(\alpha-1)}{1-c_i} > 2(a - 1)$. Therefore from Lemma 49, all the convex roots are contiguous. Therefore, the set of potentially optimal solutions reduces to cardinality $2d$, where each solution is characterized by the number of components that send non-zero rate and whether or not a concave root is sent. PBA, detailed in Algorithm 1 finds the minimum value of the Lagrangian across these $2d$ solutions for a fixed λ. □

Note that by sweeping $\lambda > 0$, one can compute the lower convex envelope of the (D, R) curve. Since every Pareto optimal (D, R) must be a stationary point of (4.21), one can also use Algorithm 1 to compute the (D, R) curve itself by sweeping λ and retaining all those stationary points that are not Pareto dominated.

4.3 Application to Variable-Rate Compression

We have seen that an autoencoder formulation inspired by data compression succeeds in providing guaranteed recovery the principal source components. Conversely, a number of successful multimedia compressors have recently been proposed that are either related to, or directly inspired by,

autoencoders [67, 65, 9, 64, 63, 56, 31, 3, 10, 74, 5, 7]. In particular, Ballé *et al.* [10] show that the objective minimized by their compressor coincides with that of variational autoencoders. Following [7], we refer to this objective as *nonlinear transform coding (NTC)*. We next use Theorem 46 to show that any minimizer of the NTC objective is guaranteed to recover the principal source components if (1) the source is Gaussian, (2) the transforms are restricted to be linear, and (3) the entropy model is *factorized*, as explained below.

Let $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{K})$, where $\boldsymbol{K}$ is a positive semidefinite covariance matrix. As before, we consider an autoencoder defined by its encoder-decoder pair (f, g), where for $k \leq d$, $f : \mathbb{R}^d \to \mathbb{R}^k$ and $g : \mathbb{R}^k \to \mathbb{R}^d$ are chosen from prespecified classes C_f and C_g. The NTC framework assumes dithered quantization during training, as in Section 4.1 and [4, 20], and seeks to minimize the Lagrangian

$$\inf_{f \in C_f, g \in C_g} \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon}} \left[\|\boldsymbol{x} - g\left(Q\left(f(\boldsymbol{x}) + \boldsymbol{\varepsilon}\right) - \boldsymbol{\varepsilon}\right)\|_2^2 \right] + \lambda H\left(Q\left(f(\boldsymbol{x}) + \boldsymbol{\varepsilon}\right) - \boldsymbol{\varepsilon} | \boldsymbol{\varepsilon}\right). \tag{4.30}$$

where $\lambda > 0$ and $\boldsymbol{\varepsilon}$ has i.i.d. Unif$[-0.5, 0.5]$ components. NTC assumes variable-length compression, and the quantity

$$H\left(Q\left(f(\boldsymbol{x}) + \boldsymbol{\varepsilon}\right) - \boldsymbol{\varepsilon} | \boldsymbol{\varepsilon}\right)$$

is an accurate estimate of minimum expected codelength length for the discrete random vector $Q\left(f(\boldsymbol{x}) + \boldsymbol{\varepsilon}\right)$. As we noted in Section 4.1, [73] showed that for any random variable $\boldsymbol{x}$, $Q\left(\boldsymbol{x} + \boldsymbol{\varepsilon}\right) - \boldsymbol{\varepsilon}$ and $\boldsymbol{x} + \boldsymbol{\varepsilon}$ have the same joint distribution with $\boldsymbol{x}$. They also showed that $H\left(Q\left(\boldsymbol{x} + \boldsymbol{\varepsilon}\right) - \boldsymbol{\varepsilon} | \boldsymbol{\varepsilon}\right) = I\left(\boldsymbol{x} + \boldsymbol{\varepsilon}; \boldsymbol{x}\right) = h(\boldsymbol{x} + \boldsymbol{\varepsilon})$, where $h(\cdot)$ denotes differential entropy. Therefore, the objective can be written as

$$\inf_{f \in C_f, g \in C_g} \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon}} \left[\|\boldsymbol{x} - g\left(f\left(\boldsymbol{x}\right) + \boldsymbol{\varepsilon}\right)\|_2^2 \right] + \lambda h\left(f\left(\boldsymbol{x}\right) + \boldsymbol{\varepsilon}\right). \tag{4.31}$$

(Compare eq.(13) in [7]).

We consider the case in which C_f, C_g are the class of linear functions. Let $\boldsymbol{W}, \boldsymbol{T}$ be d-by-d

matrices. Define $f(\boldsymbol{x}) = \boldsymbol{W}^\top \boldsymbol{x}$, $g(\boldsymbol{x}) = \boldsymbol{T}\boldsymbol{x}$. Substituting this in the above equation, we obtain

$$\inf_{\boldsymbol{W},\boldsymbol{T}} \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon}} \left[\left\| \boldsymbol{x} - \boldsymbol{T}(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{\varepsilon}) \right\|_2^2 \right] + \lambda h(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{\varepsilon}). \tag{4.32}$$

Since $\boldsymbol{T}$ does not appear in the rate constraint, the optimal $\boldsymbol{T}$ can be chosen to be the minimum mean squared error estimator of $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{K})$ given $\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{\varepsilon}$, as in Section 4.1. This gives

$$\inf_{\boldsymbol{W}} \operatorname{tr}(\boldsymbol{K}) - \operatorname{tr}(\boldsymbol{K}\boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W} + \frac{\boldsymbol{I}}{12} \right)^{-1} \boldsymbol{W}^\top \boldsymbol{K}) + \lambda h(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{\varepsilon}). \tag{4.33}$$

As noted earlier, the rate term $h(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{\varepsilon})$ is an accurate estimate for the minimum expected length of the compressed representation of $Q(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{\varepsilon})$. This assumes that the different components of this vector are encoded jointly, however. In practice, one often encodes them separately, relying on the transform $\boldsymbol{W}$ to eliminate redundancy among the components. Accordingly, we replace the rate term with

$$\sum_{i=1}^{d} h(\boldsymbol{w}_i^\top \boldsymbol{x} + [\boldsymbol{\varepsilon}]_i),$$

to arrive at the optimization problem

$$\inf_{\boldsymbol{W}} \operatorname{tr}(\boldsymbol{K}) - \operatorname{tr}(\boldsymbol{K}\boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W} + \frac{\boldsymbol{I}}{12} \right)^{-1} \boldsymbol{W}^\top \boldsymbol{K}) + \lambda \cdot \sum_{i=1}^{d} h(\boldsymbol{w}_i^\top \boldsymbol{x} + [\boldsymbol{\varepsilon}]_i). \tag{4.34}$$

Theorem 51. *Suppose $\boldsymbol{K}$ has distinct eigenvalues. Then any $\boldsymbol{W}$ that achieves the infimum in (4.34) has the property that all of its nonzero rows are eigenvectors of $\boldsymbol{K}$.*

Proof. Since the distribution of $\boldsymbol{\varepsilon}$ is fixed, by the Gaussian assumption on $\boldsymbol{x}$, $h\left(\boldsymbol{w}_j^\top \boldsymbol{x} + [\boldsymbol{\varepsilon}]_j\right)$ only depends on $\boldsymbol{w}_j$ through $\boldsymbol{w}_j^\top \boldsymbol{K} \boldsymbol{w}_j$. Thus we may write

$$h(\boldsymbol{w}_j^\top \boldsymbol{x} + \epsilon) = \rho_{sl}(\boldsymbol{w}_j^\top \boldsymbol{K} \boldsymbol{w}_j). \tag{4.35}$$

By Theorem 46, it suffices to show that $\rho_{sl}(\cdot)$ is strictly concave. Let Z be a standard Normal random variable and let ϵ be uniformly distributed over $[-1/2, 1/2]$, independent of Z. Then we

have

$$\rho_{sl}(s) = h(\sqrt{s} \cdot Z + \epsilon). \tag{4.36}$$

Thus by de Bruijn's identity [23],

$$\rho'_{sl}(s) = \frac{1}{2} J(\epsilon + \sqrt{s} \cdot Z), \tag{4.37}$$

where $J(\cdot)$ is the Fisher information. To show that $\rho'_{sl}(\cdot)$ is strictly concave, it suffices to show that $J(\epsilon + \sqrt{s} \cdot Z)$ is strictly decreasing in s.[2] To this end, let $t > s > 0$ and let Z_1 and Z_2 be i.i.d. standard Normal random variables, independent of ϵ. Then

$$J(\epsilon + \sqrt{t} \cdot Z) = J(\epsilon + \sqrt{s} \cdot Z_1 + \sqrt{t-s} \cdot Z_2) \tag{4.38}$$

and by the convolution inequality for Fisher information [43],

$$\frac{1}{J(\epsilon + \sqrt{s} \cdot Z_1 + \sqrt{t-s} \cdot Z_2)} > \frac{1}{J(\epsilon + \sqrt{s} \cdot Z_1)} + \frac{1}{J(\sqrt{t-s} \cdot Z_2)} > \frac{1}{J(\epsilon + \sqrt{s} \cdot Z_1)}, \tag{4.39}$$

where the first inequality is strict because $\epsilon + \sqrt{s} \cdot Z_1$ is not Gaussian distributed. □

4.4 Compression Experiments

We validate the PBA algorithm experimentally by comparing the performance of a PBA-derived fixed-rate compressor against the performance of baseline fixed-rate compressors. The code of our implementation can be found at `https://github.com/SourbhBh/PBA`. As we noted in the previous section, although variable-rate codes are more commonplace in practice, fixed-rate codes do offer some advantages over their more general counterparts:

[2] If $g'(\cdot)$ is strictly decreasing then for all $t > s$, $g(t) = g(s) + \int_s^t g'(u)du < g(s) + g'(s)(t-s)$ and likewise for $t < s$. That $g(\cdot)$ is strictly concave then follows from the standard first-order test for concavity [19].

1. In applications where a train of source realizations are compressed sequentially, fixed-rated coding allows for simple concatenation of the compressed representations. Maintaining synchrony between the encoder and decoder is simpler than with variable-rate codes.
2. In applications where a dataset of source realizations are individually compressed, fixed-rate coding allows for random access of data points from the compressed representation.
3. In streaming in which a sequence of realizations will be streamed, bandwidth provisioning is simplified when the bit-rate is constant over time.

Fixed-rate compressors exist for specialized sources such as speech [48, 58] and audio more generally [68]. We consider a general-purpose, learned, fixed-rate compressor derived from PBA and the following two quantization operations. The first, $Q_{CD}(a, \sigma^2, U, x)$[3] accepts the hyperparameter a, a variance estimate σ^2, a dither realization U, and the scalar source realization to be compressed, x, and outputs (a binary representation of) the nearest point to x in the set

$$\left\{ i + U : i \in \mathbb{Z} \text{ and } i + U \in \left(-\frac{\Gamma}{2}, \frac{\Gamma}{2} \right] \right\}, \tag{4.40}$$

where

$$\Gamma = 2^{\lfloor \frac{1}{2} \log_2(4a^2\sigma^2 + 1) \rfloor}. \tag{4.41}$$

This evidently requires $\log_2 \Gamma$ bits. The second function, $Q'_{CD}(a^2, \sigma^2, U, b)$, where b is a binary string of length $\log_2 \Gamma$, maps the binary representation b to the point in (4.40). These quantization routines are applied separately to each latent component. The σ^2 parameters are determined during training. The dither U is chosen uniformly over the set $[-1/2, 1/2]$, independently for each component. We assume that U is chosen pseudorandomly from a fixed seed that is known to both the encoder and the decoder. As such, it does not need to be explicitly communicated. For our experiments, we fix the a parameter at 15 and hard code this both at the encoder and at the decoder.

[3] "CD" stands for "clamped dithered."

We found that this choice balances the dual goals of minimizing the excess distortion due to the clamping quantized points to the interval $(\Gamma/2, \Gamma/2]$ and minimizing the rate.

PBA compression proceeds by applying Algorithm 1 to a training set to determine the matrices $\boldsymbol{W}$ and $\boldsymbol{T}$. The variance estimates $\sigma_1^2, \ldots, \sigma_d^2$ for the d latent variances are chosen as the empirical variances on the training set and are hard-coded in the encoder and decoder. Given a data point $\boldsymbol{x}$, the encoded representation is the concatenation of the bit strings $b_1, \ldots, b_d$, where

$$b_i = Q_{CD}(a^2, \sigma_i^2, U_i, \boldsymbol{w}_i^\top x),$$

The decoder parses the received bits into $b_1, \ldots, b_d$. and computes the latent reconstruction $\hat{\boldsymbol{y}}$, where

$$\hat{\boldsymbol{y}}_i = Q'_{CD}(a^2, \sigma_i^2, U_i, b_i),$$

The reconstruction is then $\boldsymbol{T}\hat{\boldsymbol{y}}$.

We evaluate the PBA compressor on MNIST [40], CIFAR-10 [39], MIT Faces Dataset, Free Spoken Digit Dataset (FSDD) [36] and a synthetic Gaussian dataset. The synthetic Gaussian dataset is generated from a diagonal covariance matrix obtained from the eigenvalues of the Faces Dataset. We compare our algorithms primarily using mean-squared error since our theoretical analysis uses mean squared error as the distortion metric. Our plots display Signal-to-Noise ratios (SNRs) for ease of interpretation. For image datasets, we also compare our algorithms using the Structural Similarity (SSIM) or the Multi-scale Strctural Similarity (MS-SSIM) metrics when applicable [72]. We also consider errors on downstream tasks, specifically classification, as a distortion measure.

For all datasets, we compare the performance of the PBA compressor against baseline scheme derived from PCA that uses Q_{CD} and Q'_{CD}. The PCA-based scheme sends some of the principal components essentially losslessly, and no information about the others. Specifically, in the context

of our framework, for any given k, we choose the first k columns of $\boldsymbol{W}$ to be aligned with the first k principal components of the dataset; the remaining columns are zero. Each nonzero column is scaled such that its Euclidean length multiplied by the eigenvalue has all the significant digits. This is done so that at high rates, the quantization procedure sends the k principal components losslessly. The quantization and decoder operations are as in the PBA-based scheme; in particular the a^2 parameter is as specified above. By varying k, we trade off rate and distortion.

4.4.1 SNR Performance

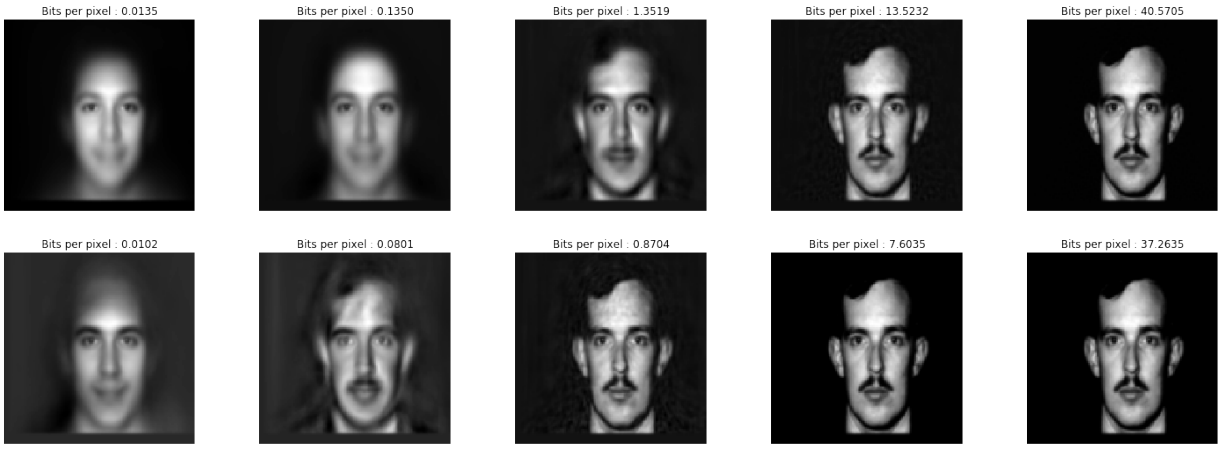


Figure 4.2: Reconstructions at different bits/pixel values for PCA (top) and PBA (bottom)

We begin by examining compression performance under mean squared error, or equivalently, the SNR, defined as

$$\text{SNR} = 10 \cdot \log_{10}\left(\frac{P}{\text{MSE}}\right).$$

where P is the empirical second moment of the dataset. This was the objective that PBA (and PCA) is designed to minimize.

In Figure 4.2, we display reconstructions for a particular image in the Faces Dataset under PBA and PCA. Figure 4.3 shows the tradeoff for PBA and PCA against JPEG and JPEG2000 (for the

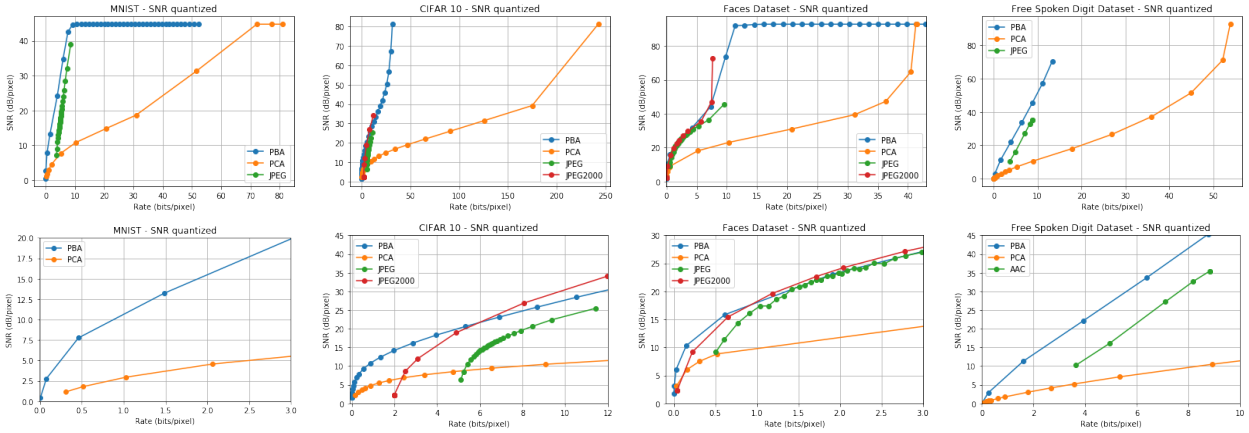


Figure 4.3: SNR/pixel vs Rate (bits/pixel) for MNIST, CIFAR-10, Faces, FSDD datasets. Figures in the bottom row are zoomed-in.

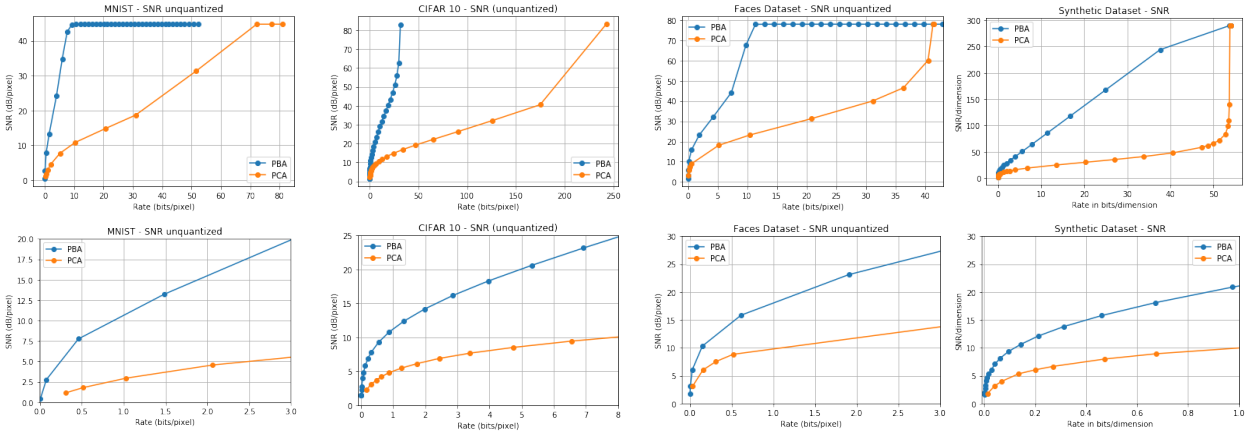


Figure 4.4: SNR/pixel vs Rate (bits/pixel) for MNIST, CIFAR-10, Faces and Synthetic dataset. Reconstructions are not rounded to integers from 0 to 255. The bottom four plots are zoomed-in versions of the top four plots.

image datasets) and AAC (for the audio dataset). All of the image datasets have integer pixel values between 0 and 255. Accordingly, we round the reconstuctions of PBA and PCA to the nearest integer in this range. Figure 4.4 shows the same tradeoff for PBA and PCA when reconstructions are not rounded off to the nearest integer. We see that PBA consistently outperforms PCA and JPEG, and is competitive with JPEG2000, even though the JPEG and JPEG2000 are variable-rate. [4]

[4]It should be noted, however, that JPEG and JPEG2000 aim to minimize subjective distortion, not MSE, and they do not allow for training on sample images, as PBA and PCA do. A similar caveat applies to AAC.

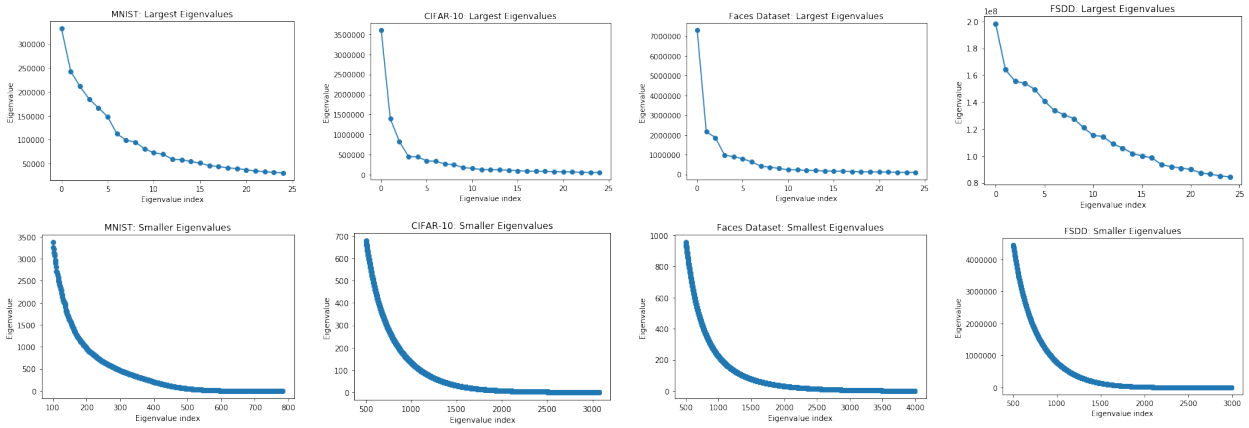


Figure 4.5: Eigenvalue distribution of the datasets. The top three plots are the largest 25 eigenvalues for MNIST, CIFAR-10, Faces and FSDD dataset. The bottom four figures plot the remaining eigenvalues except the largest 500.

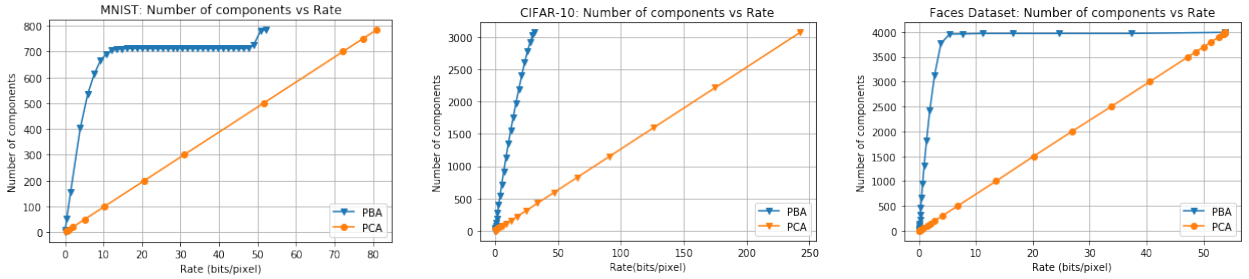


Figure 4.6: Plots of number of components sent vs rate (bits/pixel) for PBA and PCA.

We estimate the size of the JPEG header by compressing an empty image and subtract this estimate from all the compression sizes produced by JPEG. We do not plot JPEG2000 performance for MNIST since it requires at least a 32x32 image. For audio data, we observe that PBA consistently outperforms PCA and AAC. Since the image data all use 8 bits per pixel, one can obtain infinite SNR at this rate via the trivial encoding that communicates the raw bits. PCA and PBA do not find this solution because they quantize in the transform domain, where the lattice-nature of the pixel distribution is not apparent. Determining how to leverage lattice structure in the source distribution for purposes of compression is an interesting question that transcends the PBA and PCA algorithms and that we will not pursue here.

The reason that PCA performs poorly is that it favors sending the less significant bits of the most significant components over the most significant bits of less significant components, when the latter are more valuable for reconstructing the source. Arguably, it does not identify the "principal bits." Figure 4.5 shows the eigenvalue distribution of the different datasets, and Figure 4.6 shows the number of distinct components about which information is sent as a function of rate for both PBA and PCA. We see that PBA sends information about many more components for a given rate than does PCA. We discuss the ramifications of this for downstream tasks, such as classification, in Section 4.4.3.

4.4.2 SSIM Performance

Structural similarity (SSIM) and Multi-Scale Structural similarity (MS-SSIM) are metrics that are tuned to perceptual similarity. Given two images, the SSIM metric outputs a real value between 0 and 1 where a higher value indicates more similarity between the images. We evaluate the performance of our algorithms on these metrics as well in Figure 4.7. We see that PBA consistently dominates PCA, and although it was not optimized for this metric, beats JPEG at low rates as well.

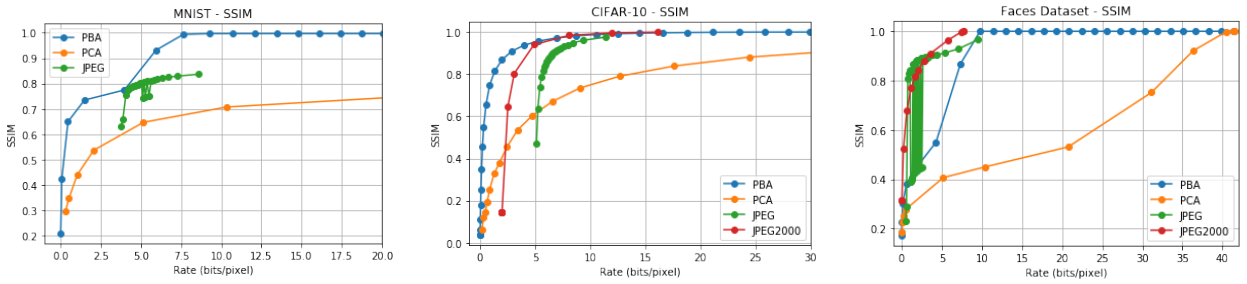


Figure 4.7: SSIM vs Rate (bits/pixel) for MNIST, CIFAR-10, Faces Dataset

4.4.3 Performance on Downstream tasks

Lastly, we compare the impact of using PBA and PCA on an important downstream task, namely classification. We evaluate the algorithms on MNIST and CIFAR-10 datasets and use neural networks for classification. Our hyperparameter and architecture choices are given in Table 4.1. We divide the dataset into three parts. From the first part, we obain the covariance matrix that we use for PCA and to obtain the PBA compressor. The second and third part are used as training and testing data for the purpose of classification. For a fixed rate, reconstructions are passed to the neural networks for training and testing respectively. Since our goal is to compare classification accuracy across the compressors, we fix both the architecture and hyperparameters, and do not perform any additional tuning for the separate algorithms.

Figure 4.8 shows that PBA outperforms PCA in terms of accuracy. The difference is especially significant for low rates; all algorithms attain roughly the same performance at higher rates.

Hyperparameter	MNIST	CIFAR-10
Architecture	2-layer fully connected NN	Convolutional Neural Network with 2 convolutional layers, pooling and three fully connected layers
# Hidden Neurons	100	NA
Optimization Algorithm	Adam	SGD with momentum
Loss	Cross-entropy	Cross-entropy
Learning Rate	0.0005	0.01

Table 4.1: Hyperparameter Choices and Architecture for Classification

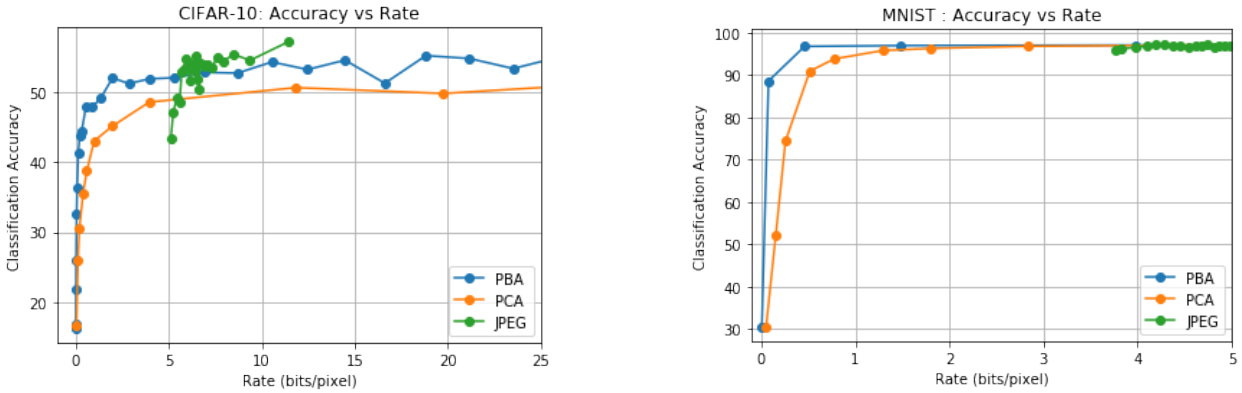


Figure 4.8: Accuracy vs Rate (bits/pixel) for MNIST, CIFAR-10

APPENDIX A

REVIEW OF SCHUR-CONVEXITY

In this section, we review the key definitions and theorems related to Schur-convexity that we use in the proof of Theorem 46.

Definition 52. *(**Majorization**)[34] For a vector $\boldsymbol{v} \in \mathbb{R}^d$, let $\boldsymbol{v}^{\downarrow}$ denote the vector with the same components arranged in descending order. Given vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, we say $\boldsymbol{a}$ majorizes $\boldsymbol{b}$ and denote $\boldsymbol{a} \succ \boldsymbol{b}$, if*

$$\sum_{i=1}^{d} [\boldsymbol{a}]_i = \sum_{i=1}^{d} [\boldsymbol{b}]_i ,$$

and for all $k \in [d-1]$,

$$\sum_{i=1}^{k} \left[\boldsymbol{a}^{\downarrow}\right]_i \geq \sum_{i=1}^{k} \left[\boldsymbol{b}^{\downarrow}\right]_i .$$

Definition 53. *(**Schur-convexity**) A function $f : \mathbb{R}^d \to \mathbb{R}$ is Schur-convex if for any vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, such that $\boldsymbol{a} \succ \boldsymbol{b}$,*

$$f(\boldsymbol{a}) \geq f(\boldsymbol{b}) .$$

f is strictly Schur-convex if the above inequality is a strict inequality for any $\boldsymbol{a} \succ \boldsymbol{b}$ that are not permutations of each other. f is Schur-concave if the direction of the inequality is reversed and is strictly Schur concave if the direction of the inequality is reversed and it is a strict inequality for any $\boldsymbol{a} \succ \boldsymbol{b}$ that are not permutations of each other.

Proposition 54. *[46] If $f : \mathbb{R} \to \mathbb{R}$ is convex, then $\phi : \mathbb{R}^d \to \mathbb{R}$ given by*

$$\phi(\boldsymbol{v}) = \sum_{i=1}^{d} f([v]_i)$$

is Schur-convex. If f is concave, then ϕ is Schur-concave. Likewise if f is strictly convex, ϕ is strictly Schur-convex and if f is strictly concave, ϕ is strictly Schur-concave.

BIBLIOGRAPHY

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

[2] E. Abaya and G. Wise. "Some remarks on optimal quantization". In: *Proc. Conf. on Inf. Sci. and Sys.* Mar. 1982.

[3] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool. "Soft-to-hard Vector Quantization for End-to-end Learning Compressible Representations". In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 1141–1151.

[4] E. Agustsson and L. Theis. "Universally Quantized Neural Compression". In: *Advances in Neural Information Processing Systems 33*. 2020, pp. 12367–12376.

[5] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool. "Generative Adversarial Networks for Extreme Learned Image Compression". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 221–231.

[6] P. Baldi and K. Hornik. "Neural networks and principal component analysis: Learning from examples without local minima". In: *Neural Networks* 2.1 (1989), pp. 53–58.

[7] J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici. "Nonlinear Transform Coding". In: *IEEE Journal of Selected Topics in Signal Processing* 15.2 (2020), pp. 339–353.

[8] J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici. "Nonlinear Transform Coding". In: *IEEE Trans. on Special Topics in Signal Proc.* (2020). to appear. DOI: 10.1109/JSTSP.2020.3034501.

[9] J. Ballé, V. Laparra, and E. Simoncelli. "End-to-end Optimization of Nonlinear Transform Codes for Perceptual Quality". In: *2016 Picture Coding Symposium (PCS)*. 2016.

[10] 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*. 2018.

[11] J. Batson, C. G. Haaf, Y. Kahn, and D. A. Roberts. "Topological obstructions to autoencoding". In: *Journal of High Energy Physics* 2021.4 (2021), pp. 1–43.

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

[13] S. Bhadane and A. B. Wagner. "On One-Bit Quantization". In: *2022 IEEE International Symposium on Information Theory (ISIT)*. 2022, pp. 584–589. DOI: 10.1109/ISIT50566.2022.9834435.

[14] S. Bhadane, A. B. Wagner, and J. Acharya. "Principal Bit Analysis: Autoencoding with Schur-Concave Loss". In: *International Conference on Machine Learning*. 2021, pp. 852–862.

[15] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738.

[16] Y. Blau and T. Michaeli. "Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 675–685.

[17] G. Boente, M. S. Barrera, and D. E. Tyler. "A characterization of elliptical distributions and some optimality properties of principal components for functional data". In: *Journal of Multivariate Analysis* 131 (2014), pp. 254–264.

[18] H. Bourlard and Y. Kamp. "Auto-association by multilayer perceptrons and singular value decomposition". In: *Biological Cybernetics* 59 (1988), pp. 291–294.

[19] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge, 2004.

[20] Y. Choi, M. El-Khamy, and J. Lee. "Variable Rate Deep Image Compression With a Conditional Autoencoder". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[21] P. A. Chou, T. Lookabaugh, and R. M. Gray. "Entropy-constrained vector quantization". In: *IEEE Transactions on acoustics, speech, and signal processing* 37.1 (1989), pp. 31–42.

[22] *Cisco Visual Networking Index: Forecast and Trends, 2017–2022*. Accessed: 2023-07-09.

[23] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2006.

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

[25] M. Cule, R. Samworth, and M. Stewart. "Maximum likelihood estimation of a multidimensional log-concave density". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72.5 (2010), pp. 545–607.

[26] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies. "Data Compression and Harmonic Analysis". In: *IEEE Trans. Inf. Theory* 44.6 (1998). DOI: 10.1109/18.720544.

[27] P. Fleischer. "Sufficient Conditions for Achieving Minimum Distortion in a Quantizer". In: *IEEE International Convention Record* 12.1 (1964), pp. 104–111.

[28] Gerald B. Folland. *Real Analysis*. 2nd. Wiley Interscience, 1999.

[29] H. Gish and J. Pierce. "Asymptotically efficient quantizing". In: *IEEE Transactions on Information Theory* 14.5 (1968), pp. 676–683.

[30] A. György and T. Linder. "Optimal Entropy-Constrained Scalar Quantization of a Uniform Source". In: *IEEE Trans. on Inf. Theory* 46.7 (2000), pp. 2704–2711. DOI: 10.1109/18.887885.

[31] A. Habibian, T. v. Rozendaal, J. M. Tomczak, and T. S. Cohen. “Video Compression with Rate-Distortion Autoencoders”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 7033–7042.

[32] 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.

[33] G. E. Hinton and R. R. Salakhutdinov. “Reducing the dimensionality of data with neural networks”. In: *Science* 313.5786 (2006), pp. 504–507.

[34] R. A. Horn and C. R. Johnson, eds. *Matrix Analysis*. USA: Cambridge University Press, 2013.

[35] J. Huang and P. Schultheiss. “Block Quantization of Correlated Gaussian Random Variables”. In: *IEEE Transactions on Communications Systems* 11.3 (1963), pp. 289–296.

[36] Z. Jackson. *Free Spoken Digit Dataset (FSDD)*. https://github.com/Jakobovski/free-spoken-digit-dataset.

[37] S. M. Kay. *Estimation Theory*. Prentice Hall PTR, 1998.

[38] J. Kieffer. “Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function”. In: *IEEE Transactions on Information Theory* 29.1 (1983), pp. 42–47.

[39] A. Krizhevsky. “Learning Multiple Layers of Features from Tiny Images”. In: *Master’s Thesis, Department of Computer Science, University of Toronto* (2009).

[40] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-based Learning Applied to Document Recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[41] L. Liao, X. Zhang, X. Wang, S. Lin, and X. Liu. “Generalized Image Reconstruction over T-Algebra”. In: *arXiv:2101.06650* (2021).

[42] S. Lloyd. “Least Squares Quantization in PCM”. In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

[43] N. M. Blachman. “The Convolution Inequality for Entropy Powers”. In: *IEEE Trans. Inf. Theory* 11.2 (Apr. 1965), pp. 267–271.

[44] A. Magnani, A. Ghosh, and R. M. Gray. “Optimal one-bit quantization”. In: *2005 Data Compression Conference (DCC)*. 2005, pp. 270–278.

[45] D. Marco and D. Neuhoff. “Low-resolution scalar quantization for Gaussian sources and squared error”. In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1689–1697. DOI: 10.1109/TIT.2006.871610.

[46] A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and its Applications*. Vol. 143. Springer, 2011.

[47] J. Max. “Quantizing for Minimum Distortion”. In: *IRE Transactions on Information Theory* 6.1 (1960), pp. 7–12. DOI: 10.1109/TIT.1960.1057548.

[48] A. V. McCree and T. P. Barnwell. “A mixed excitation LPC vocoder model for low bit rate speech coding”. In: *IEEE Transactions on Speech and Audio Processing* 3.4 (1995), pp. 242–250.

[49] Y. Meyer. “Wavelets and Applications”. In: *Lecture at CIRM Luminy meeting*. Mar. 1992.

[50] J. Mo and R. W. Heath. “Capacity Analysis of One-Bit Quantized MIMO Systems With Transmitter Channel State Information”. In: *IEEE Transactions on Signal Processing* 63.20 (2015), pp. 5498–5512. DOI: 10.1109/TSP.2015.2455527.

[51] J. R. Munkres. *Topology*. Vol. 2. Prentice Hall Upper Saddle River, 2000.

[52] D. P. Bertsekas. *Nonlinear Programming*. 2nd. Athena Scientific, 1999.

[53] W. A. Pearlman and A. Said. *Digital signal compression: principles and practice*. Cambridge university press, 2011.

[54] A. Prékopa. “On logarithmic concave measures and functions”. In: *Acta Scientiarum Mathematicarum* 34 (1973), pp. 335–343.

[55] A. Rahimi and B. Recht. “Random features for large-scale kernel machines”. In: *Advances in neural information processing systems* 20 (2007).

[56] O. Rippel and L. Bourdev. “Real-Time Adaptive Image Compression”. In: *International Conference on Machine Learning*. 2017, pp. 2922–2930.

[57] R. d. E. Santo. “Principal Component Analysis Applied to Digital Image Compression”. en. In: *Einstein (Sáo Paulo)* 10 (June 2012), pp. 135–139. ISSN: 1679-4508. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1679-45082012000200004&nrm=iso.

[58] M. Schroeder and B. Atal. “Code-excited Linear Prediction(CELP): High-quality speech at very low bit rates”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 10. 1985, pp. 937–940.

[59] C. E. Shannon. “A Mathematical Theory of Communication.” In: *Bell Syst. Tech. J.* 27.3 (1948), pp. 379–423.

[60] G. Sullivan. “Efficient scalar quantization of exponential and Laplacian random variables”. In: *IEEE Transactions on Information Theory* 42.5 (1996), pp. 1365–1374. DOI: 10.1109/18.532878.

[61] W. Szpankowski and S. Verdú. “Minimum Expected Length of Fixed-to-Variable Lossless Compression Without Prefix Constraints”. In: *IEEE Trans. on Inf. Theory* 57.7 (2011). DOI: 10.1109/TIT.2011.2145590.

[62] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng. “Fourier features let networks learn high frequency

functions in low dimensional domains". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 7537–7547.

[63] L. Theis, W. Shi, A. Cunningham, and F. Huszár. "Lossy Image Compression with Compressive Autoencoders". In: *International Conference on Learning Reperesentations*. 2017.

[64] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar. "Variable Rate Image Compression with Recurrent Neural Networks". In: *International Conference on Learning Representations*. 2016.

[65] G. Toderici, D. Vincent, N. Johnston, S. J. Hwang, D. Minnen, J. Shor, and M. Covell. "Full Resolution Image Compression with Recurrent Neural Networks". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5435–5443.

[66] A. Trushkin. "Sufficient Conditions for Uniqueness of a Locally Optimal Quantizer for a Class of Convex Error Weighting Functions". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 187–198. DOI: 10.1109/TIT.1982.1056480.

[67] M. Tschannen, E. Agustsson, and M. Lucic. "Deep Generative Models for Distribution-Preserving Lossy Compression". In: *Advances in Neural Information Processing Systems 31*. 2018, pp. 5929–5940.

[68] *Vorbis Audio Compression*. https://xiph.org/vorbis/. Accessed: 2021-01-26.

[69] A. B. Wagner and J. Ballé. "Neural Networks Optimally Compress the Sawbridge". arXiv:2011.05065. 2020.

[70] A. B. Wagner and J. Ballé. "Neural Networks Optimally Compress the Sawbridge". In: *2021 Data Compression Conference (DCC)*. 2021, pp. 143–152.

[71] F. Wang, J. Fang, H. Li, Z. Chen, and S. Li. "One-Bit Quantization Design and Channel Estimation for Massive MIMO Systems". In: *IEEE Transactions on Vehicular Technology* 67.11 (2018), pp. 10921–10934. DOI: 10.1109/TVT.2018.2870580.

[72] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. “Image Quality Assessment: From Error Visibility to Structural Similarity”. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.

[73] R. Zamir and M. Feder. “On Universal Quantization by Randomized Uniform/Lattice Quantizers”. In: *IEEE Trans. Inf. Theory* 38 (1992), pp. 428–436.

[74] L. Zhou, C. Cai, Y. Gao, S. Su, and J. Wu. “Variational Autoencoder for Low Bit-rate Image Compression”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 2617–2620.