Principal Bit Analysis: Autoencoding with Schur-Concave Loss

Sourbh Bhadane [1] **Aaron B. Wagner** [1] **Jayadev Acharya** [1]

Abstract

We consider a linear autoencoder in which the latent variables are quantized, or corrupted by noise, and the constraint is Schur-concave in the set of latent variances. Although finding the optimal encoder/decoder pair for this setup is a nonconvex optimization problem, we show that decomposing the source into its principal components is optimal. If the constraint is strictly Schur-concave and the empirical covariance matrix has only simple eigenvalues, then any optimal encoder/decoder must decompose the source in this way. As one application, we consider a strictly Schur-concave constraint that estimates the number of bits needed to represent the latent variables under fixed-rate encoding, a setup that we call *Principal Bit Analysis (PBA)*. This yields a practical, general-purpose, fixed-rate compressor that outperforms existing algorithms. As a second application, we show that a prototypical autoencoder-based variable-rate compressor is guaranteed to decompose the source into its principal components.

1. Introduction

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 $\mathcal{C}_f$ and $\mathcal{C}_g$ 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 $\mathcal{C}_f$ and $\mathcal{C}_g$ 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 (Hinton & Salakhutdinov, 2006). 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 (Bourlard & Kamp, 1988; Baldi & Hornik, 1989). That is, the components of $f(\boldsymbol{x})$ are not necessarily proportional to projections of $\boldsymbol{x}$ against the eigenvectors of the covariance matrix

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

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 sandpont, 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 (Bishop, 2006; Bourlard & Kamp, 1988; Liao et al., 2021; do Espírito Santo, 2012), 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 projec-

[1]Cornell University. Correspondence to: Sourbh Bhadane <snb62@cornell.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

tions 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 (Tschannen et al., 2018; Toderici et al., 2017; Ballé et al., 2016; Toderici et al., 2016; Theis et al., 2017; Rippel & Bourdev, 2017; Habibian et al., 2019; Agustsson et al., 2017; Ballé et al., 2018; Zhou et al., 2018; Agustsson et al., 2019; Ballé et al., 2021). 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 the paper 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.

Related Work. Several recent works have examined how linear autoencoders can be modified to guarantee recovery of PCA. Most solutions involve eliminating the invariant global optimal solutions by introducing regularization of some kind. (Oftadeh et al., 2020) propose a loss function which adds k penalties to recover the k principal directions, each corresponding to recovering up to the first $i \leq k$ principal directions. (Kunin et al., 2019) show that ℓ_2 regularization helps reduce the symmetry group to the orthogonal group. (Bao et al., 2020) further break the symmetry by considering non-uniform ℓ_2 regularization and deterministic dropout. (Ladjal et al., 2019) consider a nonlinear autoencoder with a covariance loss term to encourage finding orthogonal directions. Recovering PCA is an important problem even in the stochastic counterpart of autoencoders. (Lucas et al., 2019) analyze linear variational autoencoders (VAEs) and show that the global optimum of its objective is identical to the global optimum of log marginal likelihood of probabilistic PCA (pPCA). (Rolínek et al., 2019) analyze an approximation to the VAE loss function and show that the linear approximation to the decoder is orthogonal.

Our result on variable-rate compressors is connected to the sizable recent literature on compression using autoencoder-like architectures. Representative contributions to the literature were noted above. Those works focus mostly on the empirical performance of deep, nonlinear networks, with a particular emphasis on finding a differentiable proxy for quantization so as to train with stochastic gradient descent. In contrast, this work considers provable properties of the compressors when trained perfectly. Learned, neural fixed-rate compressors have been considered in (Li et al., 2018; Toderici et al., 2016). However, we don't compare against these since ours is a linear scheme.

Notation. We denote matrices by bold capital letters e.g. $\boldsymbol{M}$, and vectors by bold small, e.g. $\boldsymbol{v}$. The j^{th} column of a matrix $\boldsymbol{M}$ is denoted by $\boldsymbol{m}_j$ and the j^{th} entry of a vector $\boldsymbol{v}$ by $[\boldsymbol{v}]_j$. We denote the set $\{1, 2, \cdots d\}$ by $[d]$. A sequence $a_1, a_2, \cdots a_n$ is denoted by $\{a_i\}_{i=1}^n$. We denote the zero column by $\mathbf{0}$. Logarithms without specified bases denote natural logarithms.

Organization. The balance of the paper is organized as follows. We describe our constrained linear autoencoder framework in Section 2. This results in an optimization problem that we solve for any Schur-concave constraint in Section 2.1. In Section 3, we recover linear autoencoders and PBA under our framework. We apply the PBA solution to a problem in variable-rate compression of Gaussian sources in Section 4. Section 5 contains experiments comparing the performance of the PBA-based fixed-rate compressor against existing fixed-rate linear compressors on

image and audio datasets. Complete proofs of all theorems can be found in the supplementary.

2. Linear Autoencoding with a Schur-Concave Constraint

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

$$f(\boldsymbol{x}) = \boldsymbol{W}^\top \boldsymbol{x}, \text{ and } g(\boldsymbol{x}) = \boldsymbol{T}\boldsymbol{x}. \tag{2}$$

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 to 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

$$\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{\varepsilon} \tag{3}$$

and our objective is to minimize the mean squared error

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\left\| \boldsymbol{x}_i - \boldsymbol{T}\left(\boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{\varepsilon}\right) \right\|_2^2 \right]. \tag{4}$$

This is equivalent to quantizing the latents, in the following sense. Let $Q(\cdot)$ be the function that maps any real number to its nearest integer and ε be a random variable unformly distributed over$[-1/2, 1/2]$. Then for X independent of ε, the quantities $Q(X+\varepsilon) - \varepsilon$ and $X + \varepsilon$ have the same distribution (Zamir & Feder, 1992). Thus (4) 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. 1.

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 (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

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

We arrive at our optimization problem:

$$\begin{aligned} \inf_{\boldsymbol{W},\boldsymbol{T}} \quad & \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\left\| \boldsymbol{x}_i - \boldsymbol{T}\left(\boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{\varepsilon}\right) \right\|_2^2 \right] \\ \text{subject to} \quad & R \geq \sum_{j=1}^{d} \frac{1}{2}\log\left(4a^2 \boldsymbol{w}_i^\top \boldsymbol{K} \boldsymbol{w}_i + 1\right). \end{aligned} \tag{5}$$

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 supplementary 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

$$\begin{aligned} \inf_{\boldsymbol{W},\boldsymbol{T}} \quad & \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\left\| \boldsymbol{x}_i - \boldsymbol{T}\left(\boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{\varepsilon}\right) \right\|_2^2 \right] \\ \text{subject to} \quad & R \geq \rho\left(\left\{ \boldsymbol{w}_j^\top \boldsymbol{K} \boldsymbol{w}_j \right\}_{j=1}^d \right) \end{aligned} \tag{6}$$

where $\rho(\cdot)$ is any Schur-concave function. 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. (Kay, 1998)):

$$\boldsymbol{T}^* = \boldsymbol{K}\boldsymbol{W}(\boldsymbol{W}^\top \boldsymbol{K}\boldsymbol{W} + \sigma^2 \boldsymbol{I})^{-1}. \tag{7}$$

Substituting for $\boldsymbol{T}$ in (6) yields an optimization problem over only $\boldsymbol{W}$. Using standard linear algebra, we rewrite the objective in terms of $\boldsymbol{K}$ and $\boldsymbol{W}$ as

$$\begin{aligned} \inf_{\boldsymbol{W}} \quad & \text{tr}(\boldsymbol{K}) - \text{tr}(\boldsymbol{K}\boldsymbol{W}(\boldsymbol{W}^\top \boldsymbol{K}\boldsymbol{W} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{W}^\top \boldsymbol{K}) \\ & \text{subject to} \quad R \geq \rho\left(\left\{ \boldsymbol{w}_j^\top \boldsymbol{K} \boldsymbol{w}_j \right\}_{j=1}^d \right). \end{aligned} \tag{8}$$

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 3, we solve the problem for the specific choice of ρ in (5). We also show how this generalizes conventional autoencoders.

2.1. Optimal Autoencoding with a Schur-Concave Constraint

The following is the main theoretical result of the paper.

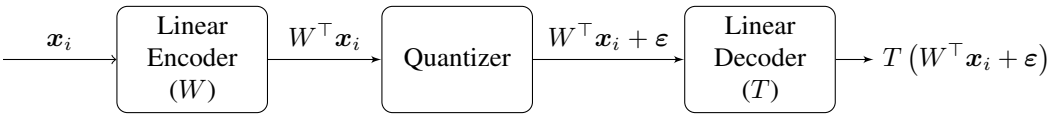

Figure 1. Compression Block Diagram

Theorem 1. *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 (8). If ρ is strictly Schur-concave and $\boldsymbol{K}$ contains distinct eigenvalues, this set contains all optimal solutions of (8).*

Proof Sketch. Let the eigenvalues of $\boldsymbol{K}$ be $\{\sigma_i^2\}_{i=1}^d$ with $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_d^2$, and let $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$ where $\boldsymbol{\Sigma}$ is a diagonal matrix with nonincreasing diagonal entries and $\boldsymbol{U}$ is an orthogonal matrix whose columns are the corresponding normalized eigenvectors of $\boldsymbol{K}$.

We first prove that the optimal value of (8) can be achieved by a $\boldsymbol{W}$ such that $\boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W}$ is a diagonal matrix. Indeed, for a given $\boldsymbol{W}$, consider $\widetilde{\boldsymbol{W}} = \boldsymbol{W}\boldsymbol{Q}$ where $\boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}$. Under this transformation, the objective is unchanged; but since the eigenvalues of $\boldsymbol{W}^\top \boldsymbol{K} \boldsymbol{W}$ majorize its diagonal elements, the rate constraint improves. Using this fact, we prove an upper bound on the objective that can be achieved by choosing the nonzero columns of $\boldsymbol{W}$ as eigenvectors of $\boldsymbol{K}$. We then prove that for a strictly Schur-concave ρ and for $\boldsymbol{K}$ whose eigenvalues are distinct, this is the only way to attain the upper bound. □

As a consequence of Theorem 1, 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 the 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. Therefore, (8) reduces to

$$\begin{aligned} \inf_{\boldsymbol{S}} \quad & \operatorname{tr}(\boldsymbol{K}) - \operatorname{tr}(\boldsymbol{K}\boldsymbol{S}(\boldsymbol{S}^\top \boldsymbol{K}\boldsymbol{S} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{S}^\top \boldsymbol{K}) \\ \text{subject to} \quad & R \geq \rho_{sl}\left(\{s_i^2 \sigma_i^2\}\right), \end{aligned} \tag{9}$$

where the infimum is over diagonal matrices $\boldsymbol{S}$. To handle situations for which $\lim_{s\to\infty} \rho_{sl}(s) < \infty$, we allow the diagonal entries of $\boldsymbol{S}$ to be ∞, with the objective for such cases defined via its continuous extension.

[1] "sl" stands for single-letter

In the next section, we will solve (9) for several specific choices of ρ_{sl}.

3. Explicit Solutions: Conventional Linear Autoencoders and PBA

3.1. 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). 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 (Baldi & Hornik, 1989).

We now recover linear autoencoders through our framework in Section 2. Consider the optimization problem in (9) where $\rho_{sl} : \mathbb{R}_{\geq 0} \to \{0, 1\}$ is a concave function defined as

$$\rho_{sl}(x) = \mathbf{1}\left[x > 0\right]. \tag{10}$$

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 the conventional linear autoencoding, but is not necessarily the unique, solution, follows immediately from Theorem 1.

Theorem 2. *If $\rho_{sl}(\cdot)$ is given by (10), then an optimal solution for (9) is given by a diagonal matrix $\boldsymbol{S}$ whose top $\min(\lfloor R \rfloor, d)$ diagonal entries are equal to ∞ and the remaining diagonal 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 (9) can be written as

$$\begin{aligned} \inf_{\{s_\ell\}} \quad & \sum_{j\in[d]\setminus\mathcal{F}} \sigma_j^2 + \sum_{\ell\in\mathcal{F}} \frac{\sigma^2 \sigma_\ell^2}{\sigma^2 + \sigma_\ell^2 s_\ell^2} \\ \text{subject to} \quad & R \geq \sum_{i=1}^{d} \mathbf{1}\left[s_i > 0\right]. \end{aligned} \tag{11}$$

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 of the largest $|\mathcal{F}|$ eigenvalues. Since the number of these eigenvalues cannot exceed R, we choose $|\mathcal{F}| = \min(\lfloor R \rfloor, d)$. $\square$

Unlike the conventional linear autoencoder framework, in Section 2, 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 (9) corresponds to linearly projecting the data along the top k eigenvectors, which is the same as PCA. Note that, like (Baldi & Hornik, 1989), 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.

3.2. 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 1. For $\gamma > \frac{2}{\sigma^2}$,

$$\rho_{sl}(x) = \frac{1}{2}\log(\gamma x + 1). \tag{12}$$

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 5 to follow), we focus on the case $\gamma > 2/\sigma^2$.

Theorem 3. *If $\rho_{sl}(\cdot)$ is given by (12), then for any $\lambda > 0$, the pair $\bar{R}_{opt}, \bar{D}_{opt}$ obtained from the output of Algorithm 1 satisfies*

$$\begin{aligned}\bar{D}_{opt} + \lambda \bar{R}_{opt} = \\ \inf_{\boldsymbol{S}} \quad tr(\boldsymbol{K}) - tr(\boldsymbol{K}\boldsymbol{S}(\boldsymbol{S}^\top \boldsymbol{K}\boldsymbol{S} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{S}^\top \boldsymbol{K}) \\ + \lambda \sum_{i=1}^{d} \rho_{sl}\left(\{s_i^2\sigma_i^2\}\right),\end{aligned} \tag{14}$$

Algorithm 1 Principal Bit Analysis (PBA)

Require: $\lambda > 0, \alpha = \gamma\sigma^2 > 2$,

$$\boldsymbol{K} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix} \succ 0, \tag{13}$$

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**
  $\bar{D}(2r-1) = \sum_{i=1}^{r} \frac{\sigma_i^2}{2(\alpha-1)}(1-c_i) + \sum_{i=r+1}^{d} \frac{\sigma_i^2}{\alpha}$,
  $\bar{R}(2r-1) = \sum_{i=1}^{r} \frac{1}{2}\log\left(\frac{\sigma_i^2}{4\lambda}\right) + \log(1+c_i)$.
  $\bar{D}(2r) = \sum_{i=1}^{r-1} \frac{\sigma_i^2}{2(\alpha-1)}(1-c_i) + \frac{\sigma_r^2}{2(\alpha-1)}(1+c_r) + \sum_{i=r+1}^{d} \frac{\sigma_i^2}{\alpha}$.
  $\bar{R}(2r) = \sum_{i=1}^{r} \frac{1}{2}\log\left(\frac{\sigma_i^2}{4\lambda}\right) + \sum_{i=1}^{r-1} \log(1+c_i) + \log(1-c_r)$.
end for
$r^* \leftarrow \arg\min_{j\in\lfloor 2\bar{d}\rfloor} \bar{D}(j) + \lambda\bar{R}(j)$.
Output $\bar{R}_{\text{opt}} = \bar{R}(r^*), \bar{D}_{\text{opt}} = \bar{D}(r^*)$.

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 (14), 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. 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 (Ballé et al., 2021; 2016; Toderici et al., 2017; 2016). In particular, Ballé *et al.* (Ballé et al., 2018) show that the objective minimized by their compressor coincides with that of variational autoencoders. Following (Ballé et al., 2021), we refer to this as the *nonlinear transform coding (NTC)* objective. We next use Theorem 1 to show that any minimizer of the NTC objective is guaranteed to recover the principal source components if (1) the source is Gaussian

distributed, (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 $\mathcal{C}_f$ and $\mathcal{C}_g$. The NTC framework assumes dithered quantization (Agustsson & Theis, 2020; Choi et al., 2019) as in Section 2, and seeks to minimize the Lagrangian

$$\inf_{f\in\mathcal{C}_f, g\in\mathcal{C}_g} \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon}}\left[\left\|\boldsymbol{x} - g\left(Q\left(f(\boldsymbol{x})+\boldsymbol{\varepsilon}\right)-\boldsymbol{\varepsilon}\right)\right\|_2^2\right] + \lambda H\left(Q\left(f(\boldsymbol{x})+\boldsymbol{\varepsilon}\right)-\boldsymbol{\varepsilon}|\boldsymbol{\varepsilon}\right),$$

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 of the discrete random vector $Q\left(f(\boldsymbol{x})+\boldsymbol{\varepsilon}\right)$. As we noted in Section 2, (Zamir & Feder, 1992) 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

$$\inf_{f\in\mathcal{C}_f, g\in\mathcal{C}_g} \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon}}\left[\left\|\boldsymbol{x} - g\left(f\left(\boldsymbol{x}\right)+\boldsymbol{\varepsilon}\right)\right\|_2^2\right] + \lambda h\left(f\left(\boldsymbol{x}\right)+\boldsymbol{\varepsilon}\right). \quad (15)$$

(Compare eq.(13) in (Ballé et al., 2021)).

We consider the case where $\mathcal{C}_f, \mathcal{C}_g$ are the class of linear functions. Let $\boldsymbol{W}, \boldsymbol{T}$ be d-by-d matrices. Define $f\left(\boldsymbol{x}\right) = \boldsymbol{W}^\top\boldsymbol{x}$, $g\left(\boldsymbol{x}\right) = \boldsymbol{T}\boldsymbol{x}$. Substituting this in the above equation, we obtain

$$\inf_{\boldsymbol{W},\boldsymbol{T}} \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon}}\left[\left\|\boldsymbol{x} - \boldsymbol{T}\left(\boldsymbol{W}^\top\boldsymbol{x}+\boldsymbol{\varepsilon}\right)\right\|_2^2\right] + \lambda h\left(\boldsymbol{W}^\top\boldsymbol{x}+\boldsymbol{\varepsilon}\right). \quad (16)$$

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 2. This gives

$$\inf_{\boldsymbol{W}} \mathrm{tr}(\boldsymbol{K}) - \mathrm{tr}(\boldsymbol{K}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{K}\boldsymbol{W} + \frac{\boldsymbol{I}}{12}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{K}) + \lambda h\left(\boldsymbol{W}^\top\boldsymbol{x}+\boldsymbol{\varepsilon}\right). \quad (17)$$

As noted earlier, the rate term $h\left(\boldsymbol{W}^\top\boldsymbol{x}+\boldsymbol{\varepsilon}\right)$ is an accurate estimate for the minimum expected length of the compressed representation of $Q\left(\boldsymbol{W}^\top\boldsymbol{x}+\boldsymbol{\varepsilon}\right)$. 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

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

to arrive at the following optimization problem

$$\inf_{\boldsymbol{W}} \mathrm{tr}(\boldsymbol{K}) - \mathrm{tr}(\boldsymbol{K}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{K}\boldsymbol{W} + \frac{\boldsymbol{I}}{12}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{K}) + \lambda\cdot\sum_{i=1}^{d} h\left(\boldsymbol{w}_i^\top\boldsymbol{x} + [\boldsymbol{\varepsilon}]_i\right). \quad (18)$$

Theorem 4. *Suppose $\boldsymbol{K}$ has distinct eigenvalues. Then any $\boldsymbol{W}$ that achieves the infimum in (18) 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

$$h(\boldsymbol{w}_j^\top\boldsymbol{x}+\epsilon) = \rho_{sl}(\boldsymbol{w}_j^\top\boldsymbol{K}\boldsymbol{w}_j). \quad (19)$$

By Theorem 1, 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

$$\rho_{sl}(s) = h(\sqrt{s}\cdot Z + \epsilon). \quad (20)$$

Thus by de Bruijn's identity (Cover & Thomas, 2006),

$$\rho'_{sl}(s) = \frac{1}{2}J(\epsilon + \sqrt{s}\cdot Z), \quad (21)$$

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

$$J(\epsilon + \sqrt{t}\cdot Z) = J(\epsilon + \sqrt{s}\cdot Z_1 + \sqrt{t-s}\cdot Z_2) \quad (22)$$

and by the convolution inequality for Fisher information (Blachman, 1965),

$$\frac{1}{J(\epsilon + \sqrt{s}\cdot Z_1 + \sqrt{t-s}\cdot Z_2)} > \frac{1}{J(\epsilon + \sqrt{s}\cdot Z_1)} + \frac{1}{J(\sqrt{t-s}\cdot Z_2)} > \frac{1}{J(\epsilon + \sqrt{s}\cdot Z_1)}, \quad (23)$$

[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 (Boyd & Vandenberghe, 2004).

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

5. 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. Although variable-rate codes are more commonplace in practice, fixed-rate codes do offer some advantages over their more general counterparts:

1. In applications where a train of source realizations will be compressed sequentially, fixed-rated coding allows for simple concatenation of the compressed representations, which also helps in maintaining synchrony.
2. In applications where a dataset of source realizations are individually compressed, fixed-rate coding allows for random access of the data points from the compressed representation.
3. In streaming, bandwidth provisioning is simplified when the bit-rate is constant over time.

Fixed-rate compressors exist for specialized sources such as speech (McCree & Barnwell, 1995; Schroeder & Atal, 1985) and audio more generally (Vor). 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

$$\left\{ i + U : i \in \mathbb{Z} \text{ and } i + U \in \left(-\frac{\Gamma}{2}, \frac{\Gamma}{2} \right] \right\}. \tag{24}$$

where

$$\Gamma = 2^{\lfloor \frac{1}{2} \log_2(4a^2\sigma^2+1) \rfloor}. \tag{25}$$

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 (24). 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. For our experiments, we fix the a parameter at 15 and hard code this in both the encoder and the decoder. We found that this choice balances the dual goals of minimizing the excess distortion due to the clamping quantized points to the interval $(\Gamma/2, \Gamma/2]$ while 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, as is the parameter a^2. Given a data point $\boldsymbol{x}$, the encoded representation is the concatenation of the bit strings $b_1, \ldots, b_d$, where

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

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 (LeCun et al., 1998), CIFAR-10 (Krizhevsky, 2009), MIT Faces Dataset (Fac), Free Spoken Digit Dataset (FSDD) (Jackson). We compare our algorithms mainly 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 Structural Similarity (MS-SSIM) metrics when applicable (Wang et al., 2004). We also consider errors on downstream tasks, specifically classification, as a distortion measure. We plot results from only selected datasets here and defer the rest to the supplementary.

For all datasets, we compare the performance of the PBA compressor against baseline scheme derived from PCA. The PCA-based scheme sends some of the principal components essentially losslessly, and sends no information about the others. Specifically, 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 projections on the column contain all 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 parameter is as specified above. By varying k, we trade off rate and distortion.

5.1. SNR Performance

We examine compression performance under mean squared error, or equivalently, the SNR, defined as

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

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

Figure 2. Reconstructions at different bits/pixel values for PCA (top) and PBA (bottom)

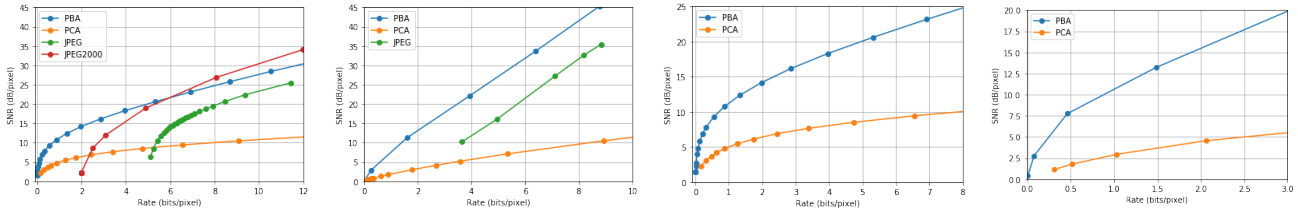

Figure 3. Plot of SNR/pixel vs Rate (bits/pixel). Left to right: CIFAR-10, FSDD, CIFAR-10, MNIST. In the last two figures reconstructions are not rounded to integers from 0 to 255. All figures are zoomed-in. Zoomed-out versions are in the supplementary.

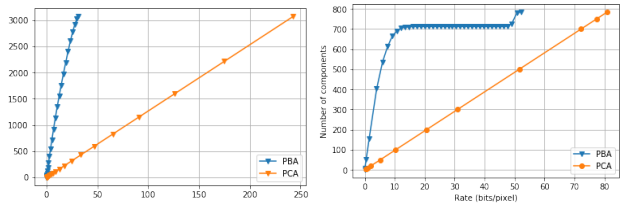

Figure 4. Number of components sent vs rate (bits/pixel). Left: CIFAR-10, Right: MNIST.

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

In Figure 2, we display reconstructions for a particular image in the Faces Dataset under PBA and PCA. The first two figures in Figure 3 show the tradeoff for PBA and PCA against JPEG and JPEG2000 (for CIFAR-10) and AAC (for FSDD). 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. The last two figures in Figure 3 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 JPEG and JPEG2000 are variable-rate. [4] We estimate the size of the JPEG header by compressing an empty header and subtract this estimate from all the compression sizes produced by JPEG. 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 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.

PCA performs poorly because 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 shows the number of distinct components about which information is sent as a function of rate for both PBA and

[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.

PCA. We see that PBA sends information about many more components for a given rate than does PCA.

5.2. SSIM Performance

Structural similarity (SSIM) and Multi-Scale Structural similarity (MS-SSIM) are metrics that are attuned 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 5. We see that PBA consistently dominates PCA, and although it was not optimized for this metric, beats JPEG at low rates as well.

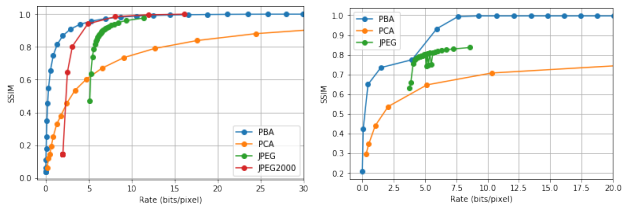

Figure 5. SSIM vs Rate (bits/pixel). Left: CIFAR-10, Right: MNIST.

5.3. Performance on Downstream tasks

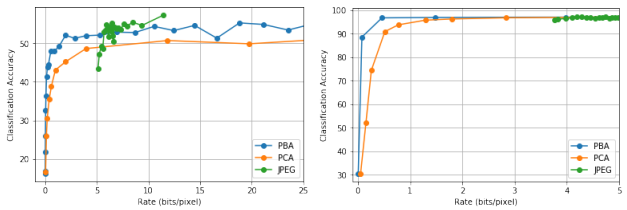

Figure 6. Accuracy vs Rate (bits/pixel). Left: CIFAR-10, Right: MNIST.

Lastly, we compare the impact of using PBA and PCA on an important downstream task, namely classification performance. We evaluate the algorithms on MNIST and CIFAR-10 datasets and use neural networks for classification. Our hyperparameter and architecture choices are in the supplementary. We divide the dataset into three parts. From the first part, we obain the covariance matrix that we use for PCA and 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 don't perform any additional tuning for the algorithms separately.

Figure 6 shows that PBA outperforms PCA in terms of accuracy. The difference is especially significant for low rates and all algorithms attain roughly the same performance at higher rates.

6. Acknowledgements

This research was supported by the US National Science Foundation under grants CCF-2008266, CCF-1934985, CCF-1617673, CCF-1846300, CCF-1815893 and the US Army Research Office under grant W911NF-18-1-0426.

References

Faces dataset. https://courses.media.mit.edu/2004fall/mas622j/04.projects/faces/. Accessed: 2021-02-03.

Vorbis audio compression. https://xiph.org/vorbis/. Accessed: 2021-01-26.

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

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

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

Baldi, P. and Hornik, K. Neural networks and Principal Component Analysis: Learning from Examples Without Local Minima. *Neural Networks*, 2(1):53–58, 1989.

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

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. Variational Image compression with a Scale Hyperprior. In *International Conference on Learning Representations*, 2018.

Ballé, J., Chou, P. A., Minnen, D., Singh, S., Johnston, N., Agustsson, E., Hwang, S. J., and Toderici, G. Nonlinear Transform Coding. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):339–353, 2021. doi: 10.1109/JSTSP.2020.3034501.

Bao, X., Lucas, J., Sachdeva, S., and Grosse, R. B. Regularized Linear Autoencoders Recover the Principal Components, Eventually. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6971–6981, 2020.

Bishop, C. M. *Pattern Recognition and Machine Learning* . Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Blachman, N. M. The convolution inequality for entropy powers. *IEEE Trans. Inf. Theory*, 11(2):267–271, April 1965.

Bourlard, H. and Kamp, Y. Auto-association by Multilayer Perceptrons and Singular Value Decomposition. *Biological Cybernetics*, 59:291–294, 1988.

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

Choi, Y., El-Khamy, M., and Lee, J. Variable rate deep image compression with a conditional autoencoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

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

do Espírito Santo, R. Principal Component Analysis Applied to Digital Image Compression. *Einstein (Sáo Paulo)*, 10:135–139, June 2012. ISSN 1679-4508.

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

Hinton, G. E. and Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313 (5786):504–507, 2006.

Jackson, Z. Free spoken digit dataset (fsdd). https://github.com/Jakobovski/free-spoken-digit-dataset.

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

Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. *Master's Thesis, Department of Computer Science, University of Toronto*, 2009.

Kunin, D., Bloom, J., Goeva, A., and Seed, C. Loss Landscapes of Regularized Linear Autoencoders. In *International Conference on Machine Learning*, volume 97, pp. 3560–3569, 2019.

Ladjal, S., Newson, A., and Pham, C. A PCA-like Autoencoder. *arXiv preprint arXiv:1904.01277*, 2019.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, M., Zuo, W., Gu, S., Zhao, D., and Zhang, D. Learning Convolutional Networks for Content-Weighted Image Compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3214–3223, 2018. doi: 10.1109/CVPR.2018.00339.

Liao, L., Zhang, X., Wang, X., Lin, S., and Liu, X. Generalized Image Reconstruction over T-Algebra. *arXiv:2101.06650*, 2021.

Lucas, J., Tucker, G., Grosse, R. B., and Norouzi, M. Don't Blame the ELBO! A Linear VAE Perspective on Posterior Collapse. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9408–9418, 2019.

McCree, A. V. and Barnwell, T. P. A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding. *IEEE Transactions on Speech and Audio Processing*, 3 (4):242–250, 1995.

Oftadeh, R., Shen, J., Wang, Z., and Shell, D. Eliminating the Invariance on the Loss Landscape of Linear Autoencoders. In *International Conference on Machine Learning*, pp. 7405–7413, 2020.

Rippel, O. and Bourdev, L. Real-Time Adaptive Image Compression. In *International Conference on Machine Learning*, pp. 2922–2930, 2017.

Rolínek, M., Zietlow, D., and Martius, G. Variational Autoencoders Pursue PCA Directions (by Accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Schroeder, M. and Atal, B. Code-excited Linear Prediction(CELP): High-quality speech at very low bit rates. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 10, pp. 937–940, 1985.

Theis, L., Shi, W., Cunningham, A., and Huszár, F. Lossy Image Compression with Compressive Autoencoders. In *International Conference on Learning Reperesentations*, 2017.

Toderici, G., O'Malley, S. M., Hwang, S. J., Vincent, D., Minnen, D., Baluja, S., Covell, M., and Sukthankar, R. Variable Rate Image Compression with Recurrent Neural Networks. In *International Conference on Learning Representations*, 2016.

Toderici, G., Vincent, D., Johnston, N., Hwang, S. J., Minnen, D., Shor, J., and Covell, M. Full Resolution Image Compression with Recurrent Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5435–5443, 2017.

Tschannen, M., Agustsson, E., and Lucic, M. Deep Generative Models for Distribution-Preserving Lossy Compression. In *Advances in Neural Information Processing Systems 31*, pp. 5929–5940. 2018.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Zamir, R. and Feder, M. On Universal Quantization by Randomized Uniform/Lattice Quantizers. *IEEE Trans. Inf. Theory*, 38:428–436, 1992.

Zhou, L., Cai, C., Gao, Y., Su, S., and Wu, J. Variational autoencoder for low bit-rate image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2617–2620, 2018.