Bridging Linear Algebra and Autoencoders

Matthias Chung

Abstract

In recent years *autoencoders* – mappings $A_\theta : \mathcal{X} \to \mathcal{X}$ parameterized by $\theta \in \mathbb{R}^\ell$ – have emerged as a cornerstone of machine learning and data science, playing a pivotal role in numerous applications. Their ability to learn efficient low-dimensional representations of data has led to significant advancements in fields such as image and natural language processing, anomaly detection, and generative modeling.

While universal approximation theorems provide a general theoretical foundation of autoencoder, various analytical aspects such as interpretability, robustness, network design, and hyperparameter selection remain relatively unexplored. *Numerical linear algebra* has played a fundamental and crucial role in the development of modern science and technology and its impact on autoencoders remains under-utilized.

The connection between linear autoencoder and singular value decomposition/principal component analysis has been laid out in various works. Recognizing the connection between linear autoencoders and singular value decomposition has sparked novel research utilizing autoencoders in fields such as matrix factorizations, model reduction, denoising, spectral clustering, and low-rank approximations to name a few.

In this work, we aim to investigate and initiate discussions on how tools from the numerical linear algebra community may provide fundamental and novel results for autoencoders, scientific machine learning, and beyond. We will discuss fundamental connections between matrix factorizations, classical inverse problems, and autoencoders in the field of signal compression and inverse problems. In the following, we provide details on the formulation of linear autoencoders through the Bayes risk formulation and the linear algebra involved in its analysis.

Linear autoencoder. Autoencoders are neural networks that learn to encode input data x into a compressed representation (latent representation) and then decode it back to reconstruct the original data $x \in \mathbb{R}^n$. Let us consider a linear autoencoder $A \in \mathbb{R}^{n\times n}$, where each element in A represents a trainable parameter. Assuming we have an ℓ-dimensional *latent space* we may compute a generic optimal autoencoder by minimizing the *Bayes risk*, i.e.,

$$\min_{\operatorname{rank}(A)\leq \ell} \; f(A) = \mathbb{E}\, \|(A-I)x\|_2^2\,, \tag{1}$$

given a distribution of the random variable x and where $\mathbb{E}$ denotes the expectation and I the identity mapping. Assuming the random variable x has symmetric positive definite second moment $\mathbb{E}\; xx^\top = \Gamma$ with Cholesky decomposition $\Gamma = BB^\top$, then

$$\mathbb{E}\; \|(A-I)x\|_2^2 = \operatorname{tr}((A-I)\Gamma(A^\top - I)) = \|AB - B\|_{\mathrm{F}}^2 \tag{2}$$

and (1) is equivalent to

$$\min_{\operatorname{rank}(A)\leq \ell} \; \|AB - B\|_{\mathrm{F}}^2\,. \tag{3}$$

For $\ell = n$ the identity mapping $A = I$ is an optimal solution. For rank constraint problems $\ell < n$ an optimal low-rank solution can be found using the following generalization of the Eckart–Young–Mirsky theorem.

Theorem 1. *Let matrix $B \in \mathbb{R}^{n\times n}$ have full row rank with SVD given by $B = U\Sigma V^\top$. Then*

$$\widehat{A} = U_\ell U_\ell^\top$$

is a solution to the minimization problem

$$\min_{rank(A)\le \ell} \|AB - B\|_F^2,$$

having a minimal Frobenius norm $\|\widehat{A}\|_F = \sqrt{\ell}$ and $\|\widehat{A}B - B\|_F^2 = \sum_{k=\ell+1}^{n} \sigma_k(B)$. This solution is unique if and only if either $\ell = n$ or $1 \le \ell < n$ and $\sigma_\ell(B) > \sigma_{\ell+1}(B)$.

Following this result, the natural choice for the autoencoder $\widehat{A}$ to be decomposed into an encoder and a decoder is $\widehat{A} = \widehat{D}\widehat{E}$, with encoder and decoder being $\widehat{E} = U_\ell^\top$ and $\widehat{D} = U_\ell$, respectively. Note that this decomposition is not unique, e.g., let K be any $n \times n$ invertible matrix then $\widehat{E} = U_\ell^\top K$ and $\widehat{D} = K^{-1}U_\ell$, are valid choices.

Sparse autoencoder. While for small latent spaces $\ell \ll n$ one obtains a low-rank approximation and a compressed approximation on the original signal x. However, compression can also be obtained utilizing a compressed sensing framework. Let us consider the problem of finding an optimal linear autoencoder A with the decomposition $A = DE$ into encoder $E \in \mathbb{R}^{\ell\times n}$ and $D \in \mathbb{R}^{n\times \ell}$ where $\ell > n$ by minimizing L^1-regularized optimization problem

$$\min_{D\in\mathbb{R}^{n\times\ell}, E\in\mathbb{R}^{\ell\times n}} \mathbb{E}\, \|(DE - I)x\|^2 + \lambda\|Ex\|_1 \tag{4}$$

with $\lambda > 0$. Autoencoders with $\ell > n$ are referred to as overcomplete autoencoders. Such sparsity-promoting overcomplete autoencoders were first been introduced in the 2010s with pioneering work from various research groups but are not commonly utilized. The generalized lasso approach (4) may generate sparse vectors Ex while maintaining the same expected squared error as an undercomplete linear autoencoder where $\ell < n$.

Numerical results. We present our analytical findings and confirm them through numerical examples. We approach linear inverse problems using linear autoencoder approximations with theoretical guarantees. Here, we illustrate this with medical tomography, deblurring, and a classic heat equation. Furthermore, we analyze small angle scattering (SAS) data – a technique from material science to obtain information about the size, shape, and arrangement of material – via the proposed sparse autoencoder. We are able to obtain superior compression rates compared to state-of-the-art approaches.