Eliminating the Invariance on the Loss Landscape of Linear Autoencoders

Reza Oftadeh [1] **Jiayi Shen** [1] **Zhangyang Wang** [1] **Dylan Shell** [1]

Abstract

This paper proposes a new loss function for linear autoencoders (LAEs) and analytically identifies the structure of the associated loss surface. Optimizing the conventional Mean Square Error (MSE) loss results in a decoder matrix that spans the principal subspace of the sample covariance of the data, but, owing to an invariance that cancels out in the global map, it will fail to identify the exact eigenvectors. We show here that our proposed loss function eliminates this issue, so the decoder converges to the exact ordered unnormalized eigenvectors of the sample covariance matrix. We characterize the full structure of the new loss landscape by establishing an analytical expression for the set of all critical points, showing that it is a subset of critical points of MSE, and that all local minima are still global. Specifically, the invariant global minima under MSE are shown to become saddle points under the new loss. Additionally, the computational complexity of the loss and its gradients are the same as MSE and, thus, the new loss is not only of theoretical importance but is of practical value, e.g., for low-rank approximation.

1. Introduction

Promising performance in deep learning has spurred active research to establish a formal understanding of the method's empirical results. Two important and complimentary lines of research have been brought to bear to analyze the behavior of various architectures of linear/non-linear neural networks: (i) global function approximation theorems describe structural aspects of networks (Leshno et al., 1993; Lu et al., 2017); while the dynamics of learning (under SGD) have been examined via (ii) approaches for analyzing properties of the loss landscape. Within the latter, a direction centered on the question of local vs. global minima, or more generally the categorization of extreme points, has flourished. Much work extends the results of Baldi & Hornik (1989) for LAEs to more complex networks (Kunin et al., 2019; Pretorius et al., 2018; Frye et al., 2019). Most notably, Zhou & Liang (2018) generalize the LAE results to deep linear networks and shallow RELU networks, and Laurent & Brecht (2018) prove global optimality for arbitrary convex differentiable loss under slightly different conditions.

Related Works Though the aforementioned works have been very successful in addressing the problem of local vs. global minima, not all global minima "are created equal". While for an LAE with MSE loss all local minima are global minima, Baldi & Hornik (1989) further show that at these minima the decoder's columns and the principal components of the covariance matrix of the data are not the necessarily the same but only span the same subspace. In other words, the LAE fails (almost surely) to identify the exact principal directions. This is due to the loss possessing a symmetry under the action of a group of invertible matrices, so that directions (and orderings/permutations thereto) will not be discriminated. (For a further elaboration and a more detailed algebraic characterization of the invariance, see Remark 1. Fig 1 also provides a visual demonstration.)

Several methods for neural networks compute the exact eigenvectors (Rubner & Tavan, 1989; Xu, 1993; Kung & Diamantaras, 1990; Oja et al., 1992), but they depend on either particular network structures or special optimization methods. More recent related works, mainly concerned with regularization, form two separate line of studies: one explores the effects of implicit regularization and the other investigates the consequence of adding a weight regularizer.

For implicit regularization, Gidel et al. (2019) extended the results of Saxe et al. (2019), and show that under some assumptions discrete Gradient Descent (GD) dynamics solves "a reduced-rank regression with a gradually increasing rank". In this approach, with *vanishing initialization* and time rescaling of the gradient dynamics, the GD optimizer learns the eigenvectors sequentially. However, not only this approach only works for GD optimization, it is approximate as the exact solution is achieved only when the initialization converges to zero.

In the case of weight regularization, it was observed by Plaut (2018), and further explored by Kunin et al. (2019)

[1]Department of Computer Science and Engineering, Texas A&M University, Texas, USA. Correspondence to: Reza Oftadeh <reza.oftadeh@tamu.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

that adding a weight regulizer to the MSE loss causes the left singular vectors of the decoder to become the exact eigenvectors of the sample covariance matrix. Recovering them, however, still requires an extra decomposition step. As Plaut (2018) points out, no existing method recovers the eigenvectors from an LAE in an optimization-independent way on a standard linear network — the present paper fills that lacuna.

Our Contributions This work proposes a loss function and shows for the first time that under this loss the decoder converges to the exact ordered unnormalized eigenvectors of the sample covariance matrix. The idea is simple: for identifying p principal directions we build up a total loss function as a sum of p squared error losses, where the i^{th} loss function identifies only the first i principal directions. This approach breaks the symmetry since minimizing the first loss results in the first principal direction, forcing the second loss to find the first and the second. This constraint is propagated through the rest of the losses, resulting in all p principal components being identified. For the new loss, we prove that all local minima are global minima.

Remarkably, the proposed loss function has both theoretical and practical implications. From a theoretical point of view, it provides a better understanding of the loss surface. Specifically, any critical point of our loss L is a critical point of the original MSE loss but not vice versa. We thus conclude that L eliminates those undesirable global minima of the original loss (i.e., exactly those which suffer from the invariance).

As for practical consequences, we show that the loss and its gradients can be compactly vectorized so that their computational complexity is no different from the MSE loss. Therefore, the loss L can be used (without needing any additional post hoc processing) to perform low rank approximation on large datasets via any method of optimization, where SGD is but one instance. In other words, the loss L enables low rank decomposition as a single optimization layer, akin to an instance of a fully differentiable building block in a larger NN pipeline (Amos & Kolter, 2017).

Another promising area where low rank approximation is of particular interest is in analysis and control of dynamical systems (Markovsky, 2014). Recently, research has shown how to conduct spectral analysis of such problems from an operator point of view by applying deep autoencoders (Lusch et al., 2018). Recovering the exact eigenfunctions of the dynamic operator is important in these contexts, which we address.

Organization of the Paper In the next section we define the loss and review the overall results. In Section 3, we provide compact expressions for the gradients, present the analytical structure of the critical points, and use them to analyze the loss landscape. Along the way we compare these results with that of MSE loss to further delineate the advantages of the new loss. We provide proof sketches and intuitions, postponing detailed proofs to the *supplementary document*. Further, to add concreteness and aid visualization, Section 4 presents some experimental results.

2. Main Results

Notation In this paper, the underlying field is always $\mathbb{R}$, and positive semidefinite matrices are symmetric by definition. We shall denote the transpose of matrix $\boldsymbol{M}$ by $\boldsymbol{M}'$. The Frobenius inner product and norm are denoted as $\langle\cdot,\cdot\rangle_F$, and $\|\cdot\|_F$, respectively. $\boldsymbol{I}_{i;p}$ is a $p \times p$ matrix with all elements zero except the first i diagonal elements being one. (Or, equivalently, the matrix obtained by setting the last $p-i$ diagonal elements of a $p \times p$ identity matrix to zero.)

The Linear Autoencoder (LAE) The LAE we consider here is a neural network consisting of n-dimensional input and output with a single hidden layer of width $p < n$. The network is linear in the sense that all activations are identity functions. The constraints on dimension and requiring only a single hidden layer are mainly for simplicity and can be relaxed without major impact on the results. Remark 9 further elaborates on configurations with multiple hidden linear layers, and dimensions that differ.

The Loss Let $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$ be the input and output matrices, where m centered sample points, each n-dimensional, are stacked column-wise. Let $\boldsymbol{x}_j \in \mathbb{R}^n$ and $\boldsymbol{y}_j \in \mathbb{R}^n$ be the j^{th} sample input and output (i.e. the j^{th} column of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively). Define the loss function $L(\boldsymbol{A},\boldsymbol{B})$ as

$$\begin{aligned} L(\boldsymbol{A},\boldsymbol{B}) &\coloneqq \sum_{i=1}^{p}\sum_{j=1}^{m} \|\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{x}_j\|_2^2 \\ &= \sum_{i=1}^{p} \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\|_F^2 , \end{aligned} \tag{1}$$

where, the matrices $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, and $\boldsymbol{B} \in \mathbb{R}^{p\times n}$ are the weights of the decoder and encoder of an LAE, respectively.

The results are based on the following standard assumptions that hold generically:

Assumption 1. *For an input $\boldsymbol{X}$ and output $\boldsymbol{Y}$, let $\boldsymbol{\Sigma}_{xx} \coloneqq \boldsymbol{X}\boldsymbol{X}'$, $\boldsymbol{\Sigma}_{xy} \coloneqq \boldsymbol{X}\boldsymbol{Y}'$, $\boldsymbol{\Sigma}_{yx} \coloneqq \boldsymbol{\Sigma}_{xy}'$ and $\boldsymbol{\Sigma}_{yy} \coloneqq \boldsymbol{Y}\boldsymbol{Y}'$ be their corresponding covariance matrices. We assume:*

- *The input and output data are centered (zero mean).*
- *$\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{xy}$, $\boldsymbol{\Sigma}_{yx}$ and $\boldsymbol{\Sigma}_{yy}$ are positive definite (of full rank and invertible).*
- *The sample covariance matrix $\boldsymbol{\Sigma} \coloneqq \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ is of full rank with n distinct eigenvalues denoted as $\lambda_1 > \lambda_2 > \cdots > \lambda_n$.*
- *The decoder matrix $\boldsymbol{A}$ has no zero columns.*

Remark 1 (The Invariance Problem). Using the notation just introduced, the invariance problem is easily observed. Let $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$ be the MSE loss, defined as

$$\tilde{L}(\boldsymbol{A}, \boldsymbol{B}) \coloneqq \|\boldsymbol{Y} - \boldsymbol{ABX}\|_F^2 . \tag{2}$$

Under Assumption 1, the classical result of Baldi & Hornik (1989) is that a global minima for the MSE loss, denoted as $(\boldsymbol{A}^*, \boldsymbol{B}^*)$, is given by

$$\boldsymbol{A}^* = \boldsymbol{U}_{1:p}, \text{ and } \boldsymbol{B}^* = \boldsymbol{U}'_{1:p}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1},$$

where the i^{th} column of $\boldsymbol{U}_{1:p}$ is a unit eigenvector of $\boldsymbol{\Sigma}$ that corresponds to the i^{th} largest eigenvalue. However, for any invertible $\boldsymbol{C} \in \mathbb{R}^{n\times n}$, the point $(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*)$ is another global minima since the MSE loss is invariant under the group action of $\mathrm{GL}_n(\mathbb{R})$: the general linear group of degree n (i.e. $\tilde{L}(\boldsymbol{A}^*, \boldsymbol{B}^*) = \tilde{L}(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*)$). Therefore, any optimization method will converge to some $(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*)$ with the $\boldsymbol{C}$ being dependant on initialization, ordering of data, and other factors particular to the specific method. There is no way to recover the exact eigenvectors expressed in $\boldsymbol{A}^*$ a posteriori. As presented in the following theorem, the loss L defined in Eq. 1 eliminates this problem by reducing the space of those invariant matrices from $\mathrm{GL}_n(\mathbb{R})$ to a subspace of $\mathrm{GL}_n(\mathbb{R})$ consisting of only the diagonal matrices.

Theorem 1. *Let $\boldsymbol{A}^* \in \mathbb{R}^{n\times p}$ and $\boldsymbol{B}^* \in \mathbb{R}^{p\times n}$. Under the conditions provided in Assumption 1, $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ define a local minima of the proposed loss function iff they are of the form*

$$\boldsymbol{A}^* = \boldsymbol{U}_{1:p}\boldsymbol{D}_p, \tag{3}$$

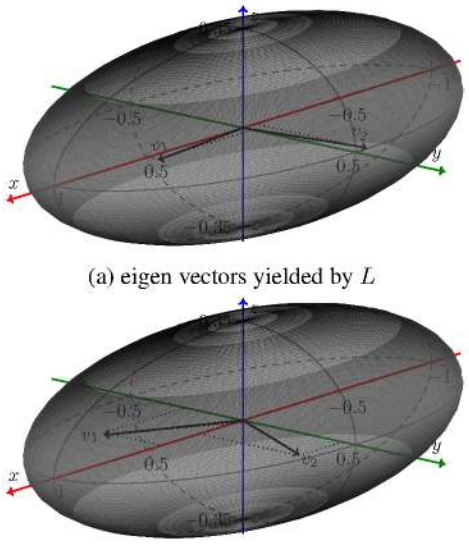
$$\boldsymbol{B}^* = \boldsymbol{D}_p^{-1}\boldsymbol{U}'_{1:p}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}, \tag{4}$$

where the i^{th} column of $\boldsymbol{U}_{1:p}$ is a unit eigenvector of $\boldsymbol{\Sigma} \coloneqq \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ corresponding the i^{th} largest eigenvalue and $\boldsymbol{D}_p$ is a diagonal matrix with nonzero diagonal elements. In other words, $\boldsymbol{A}^*$ contains ordered unnormalized eigenvectors of $\boldsymbol{\Sigma}$ corresponding to the p largest eigenvalues. Moreover, all the local minima are global minima with the value of the loss function at those global minima being

$$L(\boldsymbol{A}^*, \boldsymbol{B}^*) = p\ \mathrm{Tr}(\boldsymbol{\Sigma}_{yy}) - \sum_{i=1}^{p} (p - i + 1)\,\lambda_i, \tag{5}$$

where λ_i is the i^{th} largest eigenvalue of $\boldsymbol{\Sigma}$.

Remark 2. While $L(\cdot, \cdot)$ in the given form contains $O(p)$ matrix products, we will show that it can be evaluated with constant (fewer than 7) matrix products independent of the value p, and more importantly, component-wise scaling of gradients of MSE loss yields the gradients of L. Finally, the requirements given by Assumption 1 can be relaxed in several ways. We elaborate on these relaxations in the next section after necessary notation and definitions are given.



(a) eigen vectors yielded by L

(b) eigen vectors yielded by MSE loss

Figure 1: Visualization in low-dimensional case of the eigenvectors yielded by two loss functions when the training process of LAE converges. Top row: using L. Bottom row: using MSE loss. The newly proposed L yields the exact desired eigenvectors while the MSE loss fails to do so. The data are drawn from a zero mean multivariate Gaussian distribution with diagonal covariance matrix. The shaded ellipsoid represents the covariance.

Before we get to the theoretical analysis, we also provide an illustrative example in three dimensions, that is set to be reduced to two; i.e. $n = 3, p = 2$, and $m = 1000$. The samples are drawn from a multivariate Gaussian distribution with zero mean and a diagonal covariance matrix with reducing diagonal elements. As shown in Figure 1, the proposed loss yields the desired eigenvectors (that are x and y axes) while the MSE loss fails to do so and only projects onto the eigenspace spanned by the principal eigenvectors (that is the xy plane.)

3. Theoretical Analysis

The following constant matrices are used extensively throughout. The matrices $\boldsymbol{T}_p \in \mathbb{R}^{p\times p}$ and $\boldsymbol{S}_p \in \mathbb{R}^{p\times p}$ are defined as

$$\boldsymbol{T}_p = \mathrm{diag}\,(p, p-1, \cdots, 1)\,, \tag{6}$$

$$(\boldsymbol{S}_p)_{ij} = p - \max(i, j) + 1, \text{ e.g. } S_3 = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{7}$$

Another matrix that will appear in the formulation is $\hat{\boldsymbol{S}}_p \coloneqq \boldsymbol{T}_p^{-1}\boldsymbol{S}_p\boldsymbol{T}_p^{-1}$. Clearly, the diagonal matrix $\boldsymbol{T}_p$ is positive definite. As shown in the supplementary document, $\boldsymbol{S}_p$

and $\hat{\boldsymbol{S}}_p$ are positive definite as well.

Detailed proofs of claims appear in the supplementary document. Here, we provide proof sketches and remark on the implications of the claims. The general strategy to prove Theorem 1 is as follows. First the analytical gradients of the loss are derived in a matrix form in Propositions 1 and 2. We compare the gradients with that of the original Mean Squared Error (MSE) loss. Then we analyze the loss surface by solving the gradient equations which gives the general structure of critical points based on the rank of the decoder matrix $\boldsymbol{A}$. Thereafter, we describe several interesting properties of the critical points analytically. Notably, any critical point of the loss is also a critical point for the MSE loss but not the other way around. Finally, by performing second order analysis on the loss in Theorem 1 the exact equations for the local minima are derived which are shown to be global minima as claimed.

Remember the definition of the Loss $L(\boldsymbol{A}, \boldsymbol{B})$ and MSE loss $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$ from Eqs. 1, and 2, respectively. The first step is to calculate the gradients with respect to $\boldsymbol{A}$ and $\boldsymbol{B}$ and set them to zero to derive the implicit expressions for the critical points. To do so, first, as shown in the the supplementary document for a fixed $\boldsymbol{A}$, we derive the directional (Gateaux) derivative of the loss with respect to $\boldsymbol{B}$ along an arbitrary direction $\boldsymbol{W} \in \mathbb{R}^{p\times n}$, denoted as $d_{\boldsymbol{B}}L(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{W}$, i.e.

$$d_{\boldsymbol{B}}L(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{W} = \lim_{\|\boldsymbol{W}\|_F \to \boldsymbol{0}} \frac{L(\boldsymbol{A}, \boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A}, \boldsymbol{B})}{\|\boldsymbol{W}\|_F}.$$

As shown in the proof of the lemma, $d_{\boldsymbol{B}}L(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{W}$ is derived by writing the norm in the loss as an inner product, opening it up using linearity of inner product, disposing second order terms in $\boldsymbol{W}$(i.e. $O(\|\boldsymbol{W}\|^2)$) and rearranging the result as the inner product between the gradient with respect to $\boldsymbol{B}$, and the direction $\boldsymbol{W}$, which yields

$$d_{\boldsymbol{B}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W} = -2\langle \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx} - (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\boldsymbol{B}\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F, \tag{8}$$

where, $\circ$ is the (element-wise) Hadamard product. Second, the same process is followed to derive $d_{\boldsymbol{A}}L(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{V}$; the derivative of L with respect to $\boldsymbol{A}$ in an arbitrary direction $\boldsymbol{V} \in \mathbb{R}^{n\times p}$, for a fixed $\boldsymbol{B}$, which is then set to zero to derive the implicit expressions for the critical points. The results are formally stated in the two following propositions.

Proposition 1. *For any fixed matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ the function $L(\boldsymbol{A}, \boldsymbol{B})$ is convex in the coefficients of $\boldsymbol{B}$ and attains its minimum for any $\boldsymbol{B}$ satisfying the equation*

$$(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\boldsymbol{B}\boldsymbol{\Sigma}_{xx} = \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}, \tag{9}$$

where $\boldsymbol{T}_p$ and $\boldsymbol{S}_p$ are constant matrices defined by Eqs 6 and 7. Further, if $\boldsymbol{A}$ has no zero column, then $L(\boldsymbol{A}, \boldsymbol{B})$ is strictly convex in $\boldsymbol{B}$ and has a unique minimum when the critical $\boldsymbol{B}$ is

$$\boldsymbol{B} = \hat{\boldsymbol{B}}(\boldsymbol{A}) = (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}, \tag{10}$$

and in the autoencoder case it becomes

$$\boldsymbol{B} = \hat{\boldsymbol{B}}(\boldsymbol{A}) = (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}'. \tag{10'}$$

Remark 3. Note that as long as $\boldsymbol{A}$ has no zero column, $\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})$ is nonsingular (the reasoning appears below). In practice, $\boldsymbol{A}$ with zero columns can always be avoided by nudging the zero columns of $\boldsymbol{A}$ during the gradient decent process.

Proposition 2. *For any fixed matrix $\boldsymbol{B} \in \mathbb{R}^{p\times n}$ the function $L(\boldsymbol{A}, \boldsymbol{B})$ is a convex function in $\boldsymbol{A}$. Moreover, for a fixed $\boldsymbol{B}$, the matrix $\boldsymbol{A}$ that satisfies*

$$\boldsymbol{A}\,(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')) = \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p \tag{11}$$

is a critical point of $L(\boldsymbol{A}, \boldsymbol{B})$.

The pair $(\boldsymbol{A}, \boldsymbol{B})$ is a critical point of L if they make $d_{\boldsymbol{B}}L(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{W}$ and $d_{\boldsymbol{A}}L(\boldsymbol{A}, \boldsymbol{B})V$ zero for any pair of directions $(\boldsymbol{V}, \boldsymbol{W})$. Therefore, the implicit equations for critical points are given below, next to the ones derived by Baldi & Hornik (1989) for $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$.

- For $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$: $\begin{cases} \boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx} &= \boldsymbol{A}'\boldsymbol{\Sigma}_{yx}, \\ \boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' &= \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'. \end{cases}$
- For $L(\boldsymbol{A}, \boldsymbol{B})$: $\begin{cases} (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\boldsymbol{B}\boldsymbol{\Sigma}_{xx} &= \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}, \\ \boldsymbol{A}\,(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')) &= \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p. \end{cases}$

Remark 4. Notice similar structure with the only difference being the presence of the Hadamard product by $\boldsymbol{S}_p$ on the left and by diagonal $\boldsymbol{T}_p$ on the right. Therefore, practically, the added computational cost of evaluating the gradients is negligible compared to that of MSE loss.

The next step is to determine the structure of $(\boldsymbol{A}, \boldsymbol{B})$ that satisfies the above equations, and find the subset of those solutions that account for local minima. For the MSE loss, the first expression ($\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx} = \boldsymbol{A}'\boldsymbol{\Sigma}_{yx}$) is used to solve for $\boldsymbol{B}$ and is substituted into the second expression to derive an expression solely of $\boldsymbol{A}$. To solve the first expression for $\boldsymbol{B}$, two cases are considered separately: the case where $\boldsymbol{A}$ is of full rank p, so $\boldsymbol{A}'\boldsymbol{A}$ is invertible, and the case of $\boldsymbol{A}$ being of rank $r < p$. Here we do the same but for us there is only one case. As long as the (not necessarily full rank) matrix $\boldsymbol{A}$ has no zero column, $\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})$ is positive definite and hence, is invertible. We give only a brief discussion here, with a detailed explanation in the first lemma of the supplementary document. As shown in the lemma, $\boldsymbol{S}_p$ is positive definite and by the Shur product theorem for any $\boldsymbol{A}$ (of any rank), $\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})$ is positive semidefinite. However, as a result of the Oppenheim inequality (see Horn & Johnson, 2012, Thm 7.8.16), which in our case

becomes $\det(\boldsymbol{S}_p)\prod_i(\boldsymbol{A}'\boldsymbol{A})_{ii} \leq \det(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))$, as long as $\boldsymbol{A}$ has no zero column, $\prod_i(\boldsymbol{A}'\boldsymbol{A})_{ii} > 0$ and therefore $\det(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})) > 0$. Here, we assume $\boldsymbol{A}$ of any rank $r \leq p$ has no zero column (since this can be easily avoided in practice) and consider $\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})$ to be always invertible. Therefore, $(\boldsymbol{A}, \boldsymbol{B})$ define a critical point of losses $\tilde{L}$ and L if the following equations for critical points hold:

- For $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$ and *full rank* $\boldsymbol{A}$:

$$\begin{cases} \boldsymbol{B} = \hat{\boldsymbol{B}}(\boldsymbol{A}) = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}, \\ \boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' = \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'. \end{cases}$$

- For $L(\boldsymbol{A}, \boldsymbol{B})$ and *no zero column* $\boldsymbol{A}$:

$$\begin{cases} \boldsymbol{B} = \hat{\boldsymbol{B}}(\boldsymbol{A}) = (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}, \\ \boldsymbol{A}\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right) = \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p. \end{cases}$$

Before we state the main theorem we need the following definitions and notation. First, a rectangular permutation matrix $\boldsymbol{\Pi}_r \in \mathbb{R}^{r\times p}$ is a matrix where each column consists of at most one nonzero element with the value 1. If the rank of $\boldsymbol{\Pi}_r$ is r with $r < p$ then clearly, $\boldsymbol{\Pi}_r$ has $p - r$ zero columns. Also when those zero columns are discarded the resulting $r \times r$ submatrix of $\boldsymbol{\Pi}_r$ is a standard square permutation matrix.

Second, under the conditions provided in Assumption 1, the matrix $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ has an eigenvalue decomposition $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$, where the i^{th} column of $\boldsymbol{U}$, denoted as $\boldsymbol{u}_i$, is an eigenvector of $\boldsymbol{\Sigma}$ corresponding to the i^{th} largest eigenvalue of $\boldsymbol{\Sigma}$, namely λ_i. Also $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$ is the diagonal vector of ordered eigenvalues of $\boldsymbol{\Sigma}$, with $\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0$. We use the following notation to organize a subset of the eigenvectors of $\boldsymbol{\Sigma}$ into a rectangular matrix. For any $r \leq p$, let $\mathbb{I}_r = \{i_1, \cdots, i_r\}(1 \leq i_1 < \cdots < i_r < n)$ be any *ordered* r-index set. Define $\boldsymbol{U}_{\mathbb{I}_r} \in \mathbb{R}^{n\times p}$ as $\boldsymbol{U}_{\mathbb{I}_r} = [\boldsymbol{u}_{i_1}, \cdots, \boldsymbol{u}_{i_r}]$. That is the columns of $\boldsymbol{U}_{\mathbb{I}_r}$ are the ordered orthonormal eigenvectors of $\boldsymbol{\Sigma}$ associated with eigenvalues $\lambda_{i_1} < \cdots < \lambda_{i_r}$. Clearly when $r = p$, we have $\boldsymbol{U}_{\mathbb{I}_r} = [\boldsymbol{u}_{i_1}, \cdots, \boldsymbol{u}_{i_p}]$ corresponding to a p-index set $\mathbb{I}_p = \{i_1, \cdots, i_p\}(1 \leq i_1 < \cdots < i_p < n)$. Similarly, we define $\boldsymbol{\Lambda}_{\mathbb{I}_r} \in \mathbb{R}^{p\times p}$ as $\boldsymbol{\Lambda}_{\mathbb{I}_r} = \operatorname{diag}(\lambda_{i_1}, \cdots, \lambda_{i_r})$.

Theorem 2. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{B} \in \mathbb{R}^{p\times n}$ be such that $\boldsymbol{A}$ is of rank $r \leq p$. Under the conditions of Assumption 1, the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ define a critical point of $L(\boldsymbol{A}, \boldsymbol{B})$ if and only if for any given r-index set $\mathbb{I}_r$, and a nonsingular diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{r\times r}$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are of the form*

$$\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}, \tag{12}$$

$$\boldsymbol{B} = \boldsymbol{D}^{-1}\boldsymbol{\Pi}_{\boldsymbol{C}}\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}, \tag{13}$$

where, $\boldsymbol{C} \in \mathbb{R}^{r\times p}$ is of full rank r with nonzero and normalized columns such that $\boldsymbol{\Pi}_{\boldsymbol{C}} := (\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'$ is a rectangular permutation matrix of rank r and $\boldsymbol{C}\boldsymbol{\Pi}_{\boldsymbol{C}} = \boldsymbol{I}_r$. For all $1 \leq r \leq p$, such $\boldsymbol{C}$ always exists. In particular, if matrix $\boldsymbol{A}$ is of full rank p, i.e. $r = p$, the two given conditions on $\boldsymbol{\Pi}_{\boldsymbol{C}}$ are satisfied iff the invertible matrix $\boldsymbol{C}$ is any squared $p \times p$ permutation matrix $\boldsymbol{\Pi}$. In this case $(\boldsymbol{A}, \boldsymbol{B})$ define a critical point of $L(\boldsymbol{A}, \boldsymbol{B})$ iff they are of the form

$$\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}, \tag{14}$$

$$\boldsymbol{B} = \boldsymbol{D}^{-1}\boldsymbol{\Pi}'\boldsymbol{U}'_{\mathbb{I}_p}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}. \tag{15}$$

Remark 5. The above theorem provides explicit equations for the critical points of the loss surface in terms of the rank of the decoder matrix $\boldsymbol{A}$ and the eigenvectors of $\boldsymbol{\Sigma}$. This explicit structure allows us to further analyze the loss surface and its local/global minima.

Here, we provide a proof sketch for the above theorem to clarify the claims. Recall the EVD of $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ is $\boldsymbol{\Sigma} = U\boldsymbol{\Lambda}U'$. For both $\tilde{L}$ and L, $\boldsymbol{B}$ on the RHS of critical point equations is replaced by the corresponding $\hat{\boldsymbol{B}}(\boldsymbol{A})$. For the L, as shown in the proof, the simplification yields

$$\boldsymbol{U}'\boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\hat{\boldsymbol{B}}\boldsymbol{\Sigma}_{xx}\hat{\boldsymbol{B}}'\right)\right)\boldsymbol{A}'\boldsymbol{U} = \boldsymbol{\Lambda}\boldsymbol{\Delta}, \tag{16}$$

where $\boldsymbol{\Delta} := \boldsymbol{U}'\boldsymbol{A}\boldsymbol{T}_p(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{U}$ is symmetric and positive semidefinite. The LHS of eq. (16) is symmetric so the RHS is symmetric too, hence $\boldsymbol{\Lambda}\boldsymbol{\Delta} = (\boldsymbol{\Lambda}\boldsymbol{\Delta})' = \boldsymbol{\Delta}'\boldsymbol{\Lambda}' = \boldsymbol{\Delta}\boldsymbol{\Lambda}$. Therefore $\boldsymbol{\Delta}$ commutes with the diagonal matrix of eigenvalues $\boldsymbol{\Lambda}$. Since the eigenvalues are assumed to be distinct, $\boldsymbol{\Delta}$ has to be diagonal as well. The matrix $\boldsymbol{T}_p(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p$ is positive definite and $\boldsymbol{U}$ is an orthogonal matrix. Therefore, $r = \operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{\Delta}) = \operatorname{rank}(\boldsymbol{U}'\boldsymbol{\Delta}\boldsymbol{U})$, which implies that the diagonal matrix $\boldsymbol{\Delta}$, has r nonzero and *positive* diagonal entries. There exists an r-index set $\mathbb{I}_r$ corresponding to the nonzero diagonal elements of $\boldsymbol{\Delta}$. Forming a diagonal matrix $\boldsymbol{\Delta}_{\mathbb{I}_r} \in \mathbb{R}^{r\times r}$ by filling its diagonal entries (in order) by the nonzero diagonal elements of $\boldsymbol{\Delta}$, we have

$$\boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}' = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{\Delta}_{\mathbb{I}_r}\boldsymbol{U}'_{\mathbb{I}_r} \xRightarrow{\text{Def of } \boldsymbol{\Delta}}$$

$$\boldsymbol{A}\boldsymbol{T}_p(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}' = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{\Delta}_{\mathbb{I}_r}\boldsymbol{U}'_{\mathbb{I}_r}, \tag{17}$$

which indicates that the matrix $\boldsymbol{A}$ has the same column space as $\boldsymbol{U}_{\mathbb{I}_r}$. Hence there exists a full rank matrix $\bar{\boldsymbol{C}} \in \mathbb{R}^{r\times p}$ such that $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_r}\bar{\boldsymbol{C}}$. Since $\boldsymbol{A}$ has no zero column, $\bar{\boldsymbol{C}}$ has no zero column. Further, by normalizing the columns of $\bar{\boldsymbol{C}}$ we can write $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}$, where $\boldsymbol{D} \in \mathbb{R}^{p\times p}$ is a diagonal matrix that contains the norms of columns of $\bar{\boldsymbol{C}}$.

Baldi & Hornik (1989) did something similar for full rank $\boldsymbol{A}$ for the loss $\tilde{L}$ to derive ($\boldsymbol{A}_{\tilde{L}} = \boldsymbol{U}_{\mathbb{I}_p}\tilde{\boldsymbol{C}}$). But their $\tilde{\boldsymbol{C}}$ can be any invertible $p \times p$ matrix. In our case, the matrix $\boldsymbol{C} \in \mathbb{R}^{r\times p}$ corresponding to rank $r \leq p$ matrix $\boldsymbol{A}$, has to satisfy eq. (17) by replacing $\boldsymbol{A}$ by $\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}$ and eq. (16) by replacing $\hat{\boldsymbol{B}}(\boldsymbol{A})$ by $\hat{\boldsymbol{B}}(\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D})$. For the original loss $\tilde{L}$, equations similar to eq. (17) and eq. (16) appear but

they are are satisfied trivially by any invertible matrix $\tilde{\boldsymbol{C}}$. However, in our case, simplifying those equations by using $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}$ results (after some algebraic manipulation) in the following two conditions for $\boldsymbol{C}$:

$$\boldsymbol{\Delta}_{\mathbb{I}_r} = \boldsymbol{C}\boldsymbol{T}_p \left(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C})\right)^{-1} \boldsymbol{T}_p\boldsymbol{C}', \text{ and} \tag{18}$$

$$\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Delta}_{\mathbb{I}_r} = \boldsymbol{C}\left(\boldsymbol{S}_p \circ \big(\ (\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p (\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\ \big)\right)\boldsymbol{C}'. \tag{19}$$

As detailed in proof of Theorem 2, solving for $\boldsymbol{C}$ leads to the specific structure given in the theorem statement.

Remark 6. When $\boldsymbol{A}$ is of rank $r < p$, and with no zero columns, the invariant matrix $\boldsymbol{C}$ is not necessarily a rectangular permutation matrix but $\boldsymbol{\Pi}_{\boldsymbol{C}} \coloneqq (\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'$ is a rectangular permutation matrix with $\boldsymbol{C}\boldsymbol{\Pi}_{\boldsymbol{C}} = \boldsymbol{I}_r$. It is only when $r = p$ that the invariant matrix $\boldsymbol{C}$ becomes a permutation matrix. Nevertheless, as we show in the following corollary, the global map is always $\forall r \leq p : \boldsymbol{G} = \boldsymbol{A}\boldsymbol{B} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}$. It is possible to find further structure (in terms of block matrices) for the invariant matrix $\boldsymbol{C}$ when $r < p$. However this is not necessary, as we will shortly show that all rank-deficient matrix $\boldsymbol{A}$s are saddle points for the loss and ideally should be passed by during the gradient decent process. Based on some numerical results our conjecture is that when $r < p$ the matrix $\boldsymbol{C}$ can only start with a $r \times k$ rectangular permutation matrix of rank r with $r \leq k \leq p$ and the remaining $p - k$ columns of $\boldsymbol{C}$ may be arbitrary but nonzero.

Corollary 1. *Let $(\boldsymbol{A}, \boldsymbol{B})$ be a critical point of $L(\boldsymbol{A}, \boldsymbol{B})$ under the conditions provided in Assumption 1 and rank$\boldsymbol{A} = r \leq p$. Then the following hold:*

1. *The matrix $\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'$ is a $p \times p$ diagonal matrix of rank r.*
2. *For all $1 \leq r \leq p$, for any critical pair $(\boldsymbol{A}, \boldsymbol{B})$, the global map $\boldsymbol{G} \coloneqq \boldsymbol{A}\boldsymbol{B}$ becomes*

$$\boldsymbol{G} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}. \tag{20}$$

   For the autoencoder case ($\boldsymbol{Y} = \boldsymbol{X}$) the global map is simply $\boldsymbol{G} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{U}'_{\mathbb{I}_r}$.
3. *$(\boldsymbol{A}, \boldsymbol{B})$ is also the critical point of the classical loss $\tilde{L}(\boldsymbol{A}, \boldsymbol{B}) = \sum_{i=1}^{p} \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}\|_F^2$.*

Remark 7. The above corollary implies that $L(\boldsymbol{A}, \boldsymbol{B})$ not only does not add any extra critical points compared to the original loss $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$, it provides the same global map $\boldsymbol{G} \coloneqq \boldsymbol{A}\boldsymbol{B}$. It only limits the structure of the invariance matrix $\boldsymbol{C}$ as described in Theorem 2 so that the decoder matrix $\boldsymbol{A}$ can recover the exact eigenvectors of $\boldsymbol{\Sigma}$.

Lemma 1. *The loss function $L(\boldsymbol{A}, \boldsymbol{B})$ can be written as*

$$L(\boldsymbol{A}, \boldsymbol{B}) = p\,\mathrm{Tr}(\boldsymbol{\Sigma}_{yy}) - 2\,\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xy}\right) + \mathrm{Tr}\left(\boldsymbol{B}'\left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\right). \tag{21}$$

Remark 8. The above identity shows that the number of matrix operations required for computing the loss $L(\boldsymbol{A}, \boldsymbol{B})$ is constant and thereby independent of the value of p.

Remark 9 (Relaxing the assumptions). The proposed loss is still applicable for LAEs with multiple hidden layers because, owing to linearity, any multilayer LAE may be reduced via matrix multiplication to a network with one hidden layer (that layer being no wider than the narrowest of the original hidden layers). The only modification being that the analytical form of the gradients should be evaluated for each layer separately via the procedure underlying Propositions 1 and 2.

The second and third conditions in Assumption 1 can be relaxed by requiring only $\boldsymbol{\Sigma}_{xx}$ to be full rank. The output data may have a different dimension than the input. That is $\boldsymbol{Y} \in \mathbb{R}^{n \times m}$ and $\boldsymbol{X} \in \mathbb{R}^{n' \times m}$, where $n \neq n'$. The reason is that the given loss function is very similar to MSE loss structurally and can be represented as a Frobenius norm on the space of $n \times m$ matrices. In this case the covariance matrix $\boldsymbol{\Sigma} \coloneqq \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ is still $n \times n$. Clearly, for under-constrained systems, with $n < n'$, the full rank assumption of $\boldsymbol{\Sigma}$ is still feasible. For the over-determined case, where $n' > n$, the second and third conditions in Assumption 1 can be relaxed: we only require $\boldsymbol{\Sigma}_{xx}$ to be full rank since this is the only matrix that is inverted in the theorems. Note that if $p > \min(n', n)$ then $\boldsymbol{\Lambda}_{\mathbb{I}_p}$: the $p \times p$ diagonal matrix of eigenvalues of $\boldsymbol{\Sigma}$ for a p-index-set $\mathbb{I}_p$ inevitably has some zeros. If it has rank r, then $r < p$, which in turn results in the decoder $\boldsymbol{A}$ having rank r. However, Theorem 2 holds for decoders of any rank $r \leq p$. Finally then, following Theorem 1, the first r columns of the decoder $\boldsymbol{A}$ converge to ordered eigenvectors of $\boldsymbol{\Sigma}$ while the $p - r$ remaining columns span the kernel space of $\boldsymbol{\Sigma}$.

Finally, $\boldsymbol{\Sigma}$ need not have distinct eigenvectors. In such cases $\boldsymbol{\Delta}_{\mathbb{I}_r}$ becomes a block diagonal matrix, where the blocks correspond to identical eigenvalues and the corresponding eigenvectors in $\boldsymbol{A}^*$ are not unique but they span the respective eigenspace.

4. Experiments

Comparison of the Two Losses We will verify the loss function $L(\boldsymbol{A}, \boldsymbol{B})$ defined in eq. (1) by setting the input matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ equal to the output matrix $\boldsymbol{Y} \in \mathbb{R}^{n \times m}$ ($\boldsymbol{Y} = \boldsymbol{X}$), where the linear autodecoder (LAE) becomes a solution to PCA. In order for comparison, we train another LAE using the MSE loss $\tilde{L}(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}})$ defined as

$$\tilde{L}(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}) = \left\|\boldsymbol{Y} - \tilde{\boldsymbol{A}}\tilde{\boldsymbol{B}}\boldsymbol{X}\right\|_F^2,$$

where $\boldsymbol{Y} = \boldsymbol{X}$ is also applied.

The weights of networks are initialized to random numbers with a small enough standard deviation (10^{-7} in our case).

We choose to use the Adam optimizer with a scheduled learning rate (starting from 10^{-3} and ending with 10^{-6}), which empirically benefits the optimization process. The two training processes are stopped at the same iteration at which one of the models firstly finds all of the principal directions. As a side note, we feed all data samples to the network at one time with batch size equal to m, although mini-batch implementations are apparently amendable.

Evaluation Metrics: We use the classical PCA approach to get the ground truth principal direction matrix $\boldsymbol{A}^* \in \mathbb{R}^{n\times p}$, by conducting Eigen Value Decomposition (EVD) to $\boldsymbol{X}\boldsymbol{X}' \in \mathbb{R}^{n\times n}$ or Singular Value Decomposition (SVD) to $\boldsymbol{X} \in \mathbb{R}^{n\times m}$. As a reminder, $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ stands for the decoder weight matrix of an trained LAE given a loss function L. To measure the distance between $\boldsymbol{A}^*$ and $\boldsymbol{A}$, we propose an absolute cosine similarity (**ACS**) matrix inspired by mutual coherence (Donoho et al., 2005), which is defined as:

$$\mathbf{ACS}_{ij} = \frac{|\langle \boldsymbol{A}_i^*, \boldsymbol{A}_j \rangle|}{\|\boldsymbol{A}_i^*\| \cdot \|\boldsymbol{A}_j\|}, \tag{22}$$

where $\boldsymbol{A}_i^* \in \mathbb{R}^{n\times 1}$ denotes the i^{th} ground truth principal direction, and $\boldsymbol{A}_j \in \mathbb{R}^{n\times 1}$ denotes the j^{th} column of the decoder $\boldsymbol{A}$, $i, j = 1, 2, \ldots, p$. The elements of $\mathbf{ACS} \in \mathbb{R}^{p\times p}$ in eq. (22) take values between [0,1], measuring pairwise similarity across two sets of vectors. The absolute value absorbs the sign ambiguity of principal directions.

The performances of LAEs are evaluated by defining the following metrics:

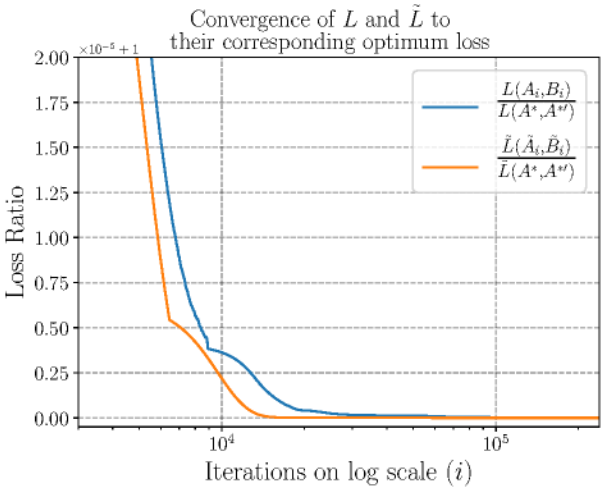
$$\mathbf{Ratio}_{TP} = \sum_{i=1}^{p} \boldsymbol{I}[\mathbf{ACS}_{ii} > 1 - \epsilon]/p \tag{23}$$

$$\mathbf{Ratio}_{FP} = \sum_{\substack{i,j=1 \\ i\neq j}}^{p} \boldsymbol{I}[\mathbf{ACS}_{ij} > 1 - \epsilon]/p, \text{ and} \tag{24}$$

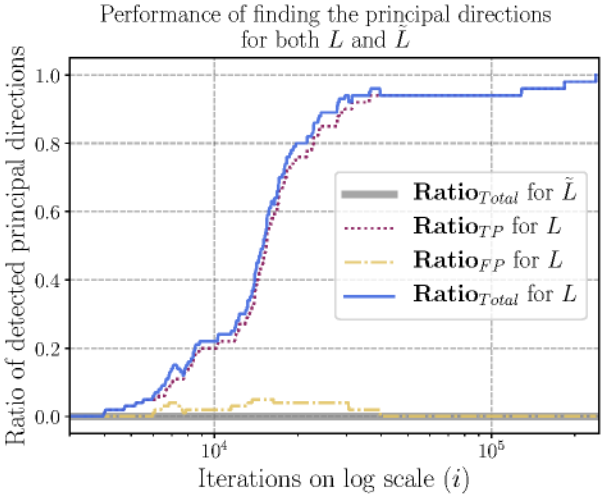
$$\mathbf{Ratio}_{Total} = \mathbf{Ratio}_{TP} + \mathbf{Ratio}_{FP}, \tag{25}$$

where $\boldsymbol{I}$ is the indicator function and ϵ is a manual tolerance threshold ($\epsilon = 0.01$ in our case). If two vectors have absolute cosine similarity over $1 - \epsilon$, they are deemed equal. Considering some columns of decoder may be correct principal directions but not in the right order, we introduce $\mathbf{Ratio}_{TP}$ and $\mathbf{Ratio}_{FP}$ in eqs. (23) and (24) to check the ratio of correct in-place and out-of-place principal directions respectively. Then $\mathbf{Ratio}_{Total}$ in eq. (25) measures the total ratio of the correctly obtained principal directions by the LAE regardless of the order.

Datasets: Although our experiments are only intended for our theories' proof-of-concept, we include both synthetic data and real data experiments. For the synthetic data, 2000 zero-centered data samples are generated from a 1000-dimension zero mean multivariate normal distribution with the covariance matrix being $\mathrm{diag}(\mathbb{N}_p)$. For the real data, we choose to use MNIST dataset (LeCun et al., 1998), which includes 60,000 grayscale handwritten digits images, each of $28 \times 28 = 784$ pixels.



(a)



(b)

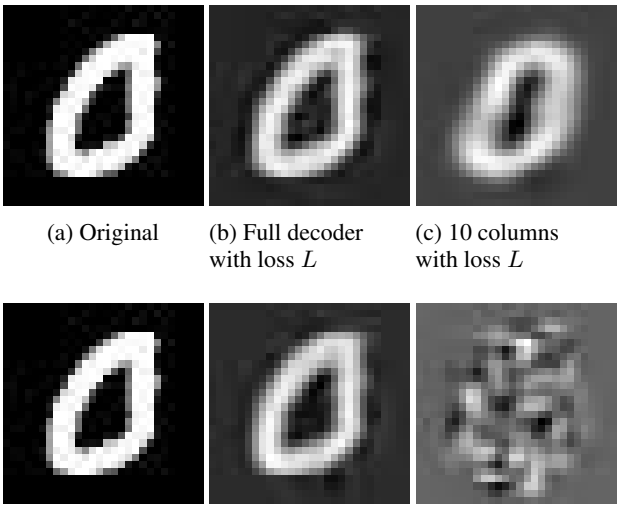
Figure 2: (a) Convergence of losses to their corresponding optimal loss. Note that the correct shift and scaling of the y-axis tick values is printed at the top left corner of the figure. (b) Performance of both losses L and $\tilde{L}$ in finding the principal directions at the columns of their respective decoders.

Synthetic Data Experiments: In our experiment, p, the number of desired principal components (PCs), is set to 100, i.e. the dimension is to be reduced from 1000 to 100. Figures 2a and 2b demonstrate a few conclusions. First, during the training process, the *loss ratio* of both losses continuously decreases to 1, i.e. they both converge to the optimal loss value. However, when both get close enough, L require more iterations since the optimizer is forced to find the right directions: it fully converges only after it has found all the principal directions in the right order.

Second, using the loss L results in finding more correct principal directions, with $\mathbf{Ratio}_{TP}$ continuously rising; and ultimately affords all correct and ordered principal directions, with $\mathbf{Ratio}_{TP}$ ending with 100%. Notice that occasionally and temporarily, some principal direction is found but not at their correct position, which is indicated by the small fluctuations of $\mathbf{Ratio}_{FP}$ in the plot. However, as optimization continues, they are shifted back to the right column, which results in $\mathbf{Ratio}_{FP}$ going back to zero, and $\mathbf{Ratio}_{TP}$ reaching one. As for $\tilde{L}$, we see that it fails to identify any principal direction correctly; both $\mathbf{Ratio}_{TP}$ and $\mathbf{Ratio}_{FP}$ for $\tilde{L}$ stay at 0, which indicates that none of the columns of the decoder $\tilde{\boldsymbol{A}}$, aligns with any principal direction.

Third, as shown in the figure, while the optimizer finds almost all the principal directions rather quickly, it requires much more iterations to find some final ones. This is because some eigenvalues in the empirical covariance matrix of the finite 2000 samples become very close (the difference becomes less than 1). Therefore, the loss has to get very close to the optimal loss, making the gradient of the loss hard to distinguish between the two.

Real Data Experiments On the MNIST dataset, we set the number of principal components (PCs) as 100, i.e., the dimension is to be reduced from 784 to 100. We also show reconstruction with the first 10 columns of the decoder that results from optimization of the respective losses. The results are depicted in Fig. 3: comparing (c) and (f), it is clear that the reconstruction performance of L is consistently better than $\tilde{L}$.



(a) Original (b) Full decoder with loss L (c) 10 columns with loss L

(d) Original (e) Full decoder with loss $\tilde{L}$ (f) 10 columns with loss $\tilde{L}$

Figure 3: Experimental comparison of reconstruction performance using real data from MNIST images. First column: original image. Second column: reconstructed image using all the decoder's columns. Third column: reconstructed image using the first 10 columns. Top row: using L. Bottom row: using $\tilde{L}$. For loss L, the first 10 columns of the decoder matrix are the most significant. For MSE loss, $\tilde{L}$, the columns do not represent principal directions and the matrix does not have them ordered in any way. The reconstruction in (c) is far superior to the one in (f).

The reason for this superiority in reconstruction is that optimization with L results in decoder weights converging to the ordered eigenvectors of the sample covariance matrix. Hence, the most significant 10 columns are simply the first 10 columns of the decoder matrix. For MSE loss, $\tilde{L}$, we again use the first 10 columns of the decoder. However, in this case, there is no way to determine which columns are more significant than others as none of them necessarily represent the exact eigenvectors, they merely span the principal eigenspace collectively. This experiment gives visual confirmation that $\tilde{L}$ does not identify PCs, while applying L performs PCA directly.

5. Conclusion

This paper introduces and analyzes a new loss function. We have proved that all local minima are global minima and that optimizing the given loss L results in a decoder matrix that converges to the exact ordered unnormalized eigenvectors of the sample covariance matrix. Given that the set of critical points of L was shown to be a subset of the critical points of the standard MSE loss, much prior work on loss surfaces of more complex networks likely extends as well. In light of the removal of undesirable global minima through L, examining more complex networks is certainly a very promising direction. For practitioners, the new loss function is valuable for low-rank approximation problems, for instance in performing principal component analysis and linear regression with linear autoencoders. There are several other possible generalizations of this approach we are currently working on. Informed by our experimental results on synthetic data, one promising thread is to improve the performance when the corresponding eigenvalues of two principal directions are very close and another is generalization of the loss for tensor decomposition.

Acknowledgments

We would like to sincerely thank Dr. Boris Hanin for introducing us to the invariance problem of LAEs. We are also grateful to the valuable comments made by the reviewers. This work is supported in part by NSF under grants ECCS-1637889 and IIS-1453652, and US Army Research Office Young Investigator Award W911NF2010240.

References

Amos, B. and Kolter, J. Z. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning (ICML)*, pp. 136–145, 2017.

Baldi, P. and Hornik, K. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

Donoho, D. L., Elad, M., and Temlyakov, V. N. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2005.

Frye, C. G., Wadia, N. S., DeWeese, M. R., and Bouchard, K. E. Numerically recovering the critical points of a deep linear autoencoder. *arXiv preprint arXiv:1901.10603*, 2019.

Gidel, G., Bach, F., and Lacoste-Julien, S. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems (NeuroIPS)*, pp. 3202–3211, 2019.

Horn, R. and Johnson, C. *Matrix Analysis*. Cambridge University Press, Cambridge, U.K., 2012. ISBN 9781139788885.

Kung, S.-Y. and Diamantaras, K. A neural network learning algorithm for adaptive principal component extraction (APEX). In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 861–864, 1990.

Kunin, D., Bloom, J. M., Goeva, A., and Seed, C. Loss landscapes of regularized linear autoencoders. *arXiv preprint arXiv:1901.08168*, 2019.

Laurent, T. and Brecht, J. Deep linear networks with arbitrary loss: All local minima are global. In *International Conference on Machine Learning (ICML)*, pp. 2902–2907, 2018.

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.

Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems (NeuroIPS)*, pp. 6231–6239, 2017.

Lusch, B., Kutz, J. N., and Brunton, S. L. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):1–10, 2018.

Markovsky, I. *Low Rank Approximation: Algorithms, Implementation, Applications*. Communications and Control Engineering. Springer London, 2014. ISBN 9781447158363.

Oja, E., Ogawa, H., and Wangviwattana, J. Principal component analysis by homogeneous neural networks, part 1: The weighted subspace criterion. *IEICE Transactions on Information and Systems*, 75(3):366–375, 1992.

Plaut, E. From principal subspaces to principal components with linear autoencoders. *arXiv preprint arXiv:1804.10253*, 2018.

Pretorius, A., Kroon, S., and Kamper, H. Learning dynamics of linear denoising autoencoders. In *International Conference on Machine Learning (ICML)*, pp. 4138–4147, 2018.

Rubner, J. and Tavan, P. A self-organizing network for principal-component analysis. *EPL (Europhysics Letters)*, 10(7):693, 1989.

Saxe, A. M., McClelland, J. L., and Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

Xu, L. Least mean square error reconstruction principle for self-organizing neural-nets. *Neural Networks*, 6(5): 627–648, 1993.

Zhou, Y. and Liang, Y. Critical points of linear neural networks: Analytical forms and landscape properties. In *International Conference on Learning Representations (ICLR)*, 2018.