Loss Landscapes of Regularized Linear Autoencoders

Daniel Kunin [* 1]   **Jonathan M. Bloom** [* 2]   **Aleksandrina Goeva** [2]   **Cotton Seed** [2]

Abstract

Autoencoders are a deep learning model for representation learning. When trained to minimize the distance between the data and its reconstruction, linear autoencoders (LAEs) learn the subspace spanned by the top principal directions but cannot learn the principal directions themselves. In this paper, we prove that L_2-regularized LAEs are symmetric at all critical points and learn the principal directions as the left singular vectors of the decoder. We smoothly parameterize the critical manifold and relate the minima to the MAP estimate of probabilistic PCA. We illustrate these results empirically and consider implications for PCA algorithms, computational neuroscience, and the algebraic topology of learning.

1. Introduction

Consider a data set consisting of points $x_1, \ldots, x_n$ in $\mathbb{R}^m$. Let $X \in \mathbb{R}^{m \times n}$ be the data matrix with columns x_i. We will assume throughout that $k \le \min\{m, n\}$ and that the singular values of X are positive and distinct.

An *autoencoder* consists of an encoder $f : \mathbb{R}^m \to \mathbb{R}^k$ and decoder $g : \mathbb{R}^k \to \mathbb{R}^m$; the latter maps the latent representation $f(x_i)$ to the reconstruction $\hat{x}_i = g(f(x_i))$ (Goodfellow et al., 2016). The full network is trained to minimize reconstruction error, typically the squared Euclidean distance between the dataset X and its reconstruction $\hat{X}$ (or equivalently, the Frobenius norm of $X - \hat{X}$). When the activations of the network are the identity, the model class reduces to that of one encoder layer $W_1 \in \mathbb{R}^{k \times m}$ and one decoder layer $W_2 \in \mathbb{R}^{m \times k}$. We refer to this model as a *linear autoencoder* (LAE) with loss function defined by

$$\mathcal{L}(W_1, W_2) = ||X - W_2 W_1 X||_F^2.$$

Parameterizing $\mathcal{L}$ by the product $W = W_2 W_1$, the Eckart-Young Theorem (Eckart & Young, 1936) states that the optimal W orthogonally projects X onto the subspace spanned by its top k *principal directions*[1].

Without regularization, LAEs learn this subspace but cannot learn the principal directions themselves due to the symmetry of $\mathcal{L}$ under the action of the group $\mathrm{GL}_k(\mathbb{R})$ of invertible $k \times k$ matrices defined by $(W_1, W_2) \mapsto (GW_1, W_2 G^{-1})$:

$$X - (W_2 G^{-1})(GW_1)X = X - W_2 W_1 X. \tag{1}$$

Indeed, $\mathcal{L}$ achieves its minimum value on a smooth submanifold of $\mathbb{R}^{k \times m} \times \mathbb{R}^{m \times k}$ diffeomorphic to $\mathrm{GL}_k(\mathbb{R})$; the learned latent representation is only defined up to deformation by invertible linear maps; and the k-dimensional eigenspace of W with eigenvalue one has no preferred basis.

In light of the above, the genesis of this work was our surprise at the theorem[2] and empirical observation in Plaut (2018) that the principal directions of X are recovered from a trained LAE as the left singular vectors of the decoder (or as the right singular vectors of the encoder). We realized by looking at the code that training was done with the common practice of L_2-regularization:

$$\mathcal{L}_\sigma(W_1, W_2) = \mathcal{L}(W_1, W_2) + \lambda \left(||W_1||_F^2 + ||W_2||_F^2 \right).$$

In this paper, we prove that LAEs *with* L_2-regularization do in fact learn the principal directions in this way, while shrinking eigenvalues in the manner of probabilistic PCA (Tipping & Bishop, 1999). The key idea is that regularization reduces the symmetry group from $\mathrm{GL}_k(\mathbb{R})$ to the orthogonal group $\mathrm{O}_k(\mathbb{R})$, which preserves the structure of SVD. We further prove that the encoder and decoder are transposes at all

---

[*]Equal contribution [1]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, California, USA [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. Correspondence to: Daniel Kunin <kunin@stanford.edu>, Jonathan M. Bloom <jbloom@broadinstitute.org>.

[1]The principal directions of X are the eigenvectors of the $m \times m$ covariance of X in descending order by eigenvalue, or equivalently the left singular vectors of the mean-centered X in descending order by (squared) singular values.

[2]The theorem was retracted following our correspondance.

critical points, with implications for whether the brain could plausibly implement error backpropagation.

1.1. Related work

Building on the original work of Eckart & Young (1936) on low-rank matrix approximation, Izenman (1975) demonstrated a connection between a rank-reduced regression model similar to an LAE and PCA. Bourlard & Kamp (1988) characterized the minima of an unregularized LAE; Baldi & Hornik (1989) extended this analysis to all critical points.

Several studies of the effect of regularization on LAEs have emerged of late. The rank-reduced regression model was extended in Mukherjee & Zhu (2011) to the study of rank-reduced ridge regression. A similar extension of the LAE model was given in Josse & Wager (2016). An in depth analysis of the linear denoising autoencoder was given in Pretorius et al. (2018) and most recently, Mianjy et al. (2018) explored the effect of dropout regularization on the minima of an LAE.

While L_2-regularization is a foundational technique in statistical learning, its effect on autoencoder models has not been fully characterized. Recent work of Mehta et al. (2018) on L_2-regularized deep linear networks applies techniques from algebraic geometry to highlight how algebraic symmetries result in "flat" critical manifolds and how L_2-regularization breaks these symmetries to produce isolated critical points. We instead apply elementary linear algebra, with intuition from algebraic topology, to completely resolve dynamics in the special case of LAEs.

1.2. Our contributions

The contributions of our paper are as follows.

- In Section 2 we consider LAEs with (i) no regularization, (ii) L_2-regularization of the composition of the encoder and decoder, and (iii) L_2-regularization of the encoder and decoder separately as in $\mathcal{L}_\sigma$. We build intuition by analyzing the scalar case, consider the relationship between regularization and orthogonality, and deduce that the encoder and decoder are transposes at all critical points of $\mathcal{L}_\sigma$.
- In Section 3, we realize all three LAE models as generative processes, most notably relating the minimum of $\mathcal{L}_\sigma$ and the MAP estimate of probabilistic PCA.
- In Section 4, we characterize all three loss landscapes. To build intuition, we first leave the overparameterized world of coordinate representations to think geometrically about the squared distance from a plane to a point cloud. We expand on this topological viewpoint in Appendix E.
- In Section 5, we illustrate these results empirically,

with all code and several talks available on GitHub[3].
- In Section 6, we discuss implications for eigendecomposition algorithms, computational neuroscience, and deep learning.

The connections we draw between regularization and orthogonality, LAEs and probabilistic PCA, and the topology of Grassmannians are novel and provide a deeper understanding of the loss landscapes of regularized linear autoencoders.

2. Regularized LAEs

In the Appendix A, we provide a self-contained derivation of the fact that an LAE with bias parameters is equivalent to an LAE without bias parameters trained on mean-centered data (Bourlard & Kamp, 1988). So without loss of generality, we assume X is mean centered and consider the following three LAE loss functions for fixed $\lambda > 0$:

$$\mathcal{L}(W_1, W_2) = ||X - W_2 W_1 X||_F^2$$
$$\mathcal{L}_\pi(W_1, W_2) = \mathcal{L}(W_1, W_2) + \lambda ||W_2 W_1||_F^2$$
$$\mathcal{L}_\sigma(W_1, W_2) = \mathcal{L}(W_1, W_2) + \lambda(||W_1||_F^2 + ||W_2||_F^2)$$

We call these the **unregularized**, **product**, and **sum** losses, respectively.

The product and sum losses mirror the loss functions of a linear *denoising autoencoder* (DAE) (Vincent et al., 2010) and linear *contractive autoencoder* (CAE) (Rifai et al., 2011) respectively. See Appendix C for details.

2.1. Visualizing LAE loss landscapes

We can visualize these loss functions directly in the case $n = m = k = 1$, as shown in Figure 1. In fact, working out the critical points in this scalar case led us to conjecture the general result in Section 4. We invite the reader to enjoy deriving the following results and experimenting with these loss landscapes using our online visualization tool.

For all three losses, the origin $w_1 = w_2 = 0$ is the unique rank-0 critical point. For $\mathcal{L}$ and $\mathcal{L}_\pi$, the origin is always a saddle point, while for $\mathcal{L}_\sigma$ the origin is either a saddle point or global minimum depending of the value of λ.

For $\mathcal{L}$, the global minima are rank-1 and consist of the hyperbola[4]

$$w_2 w_1 = 1.$$

For $\mathcal{L}_\pi$, the global minima are rank-1 and consist of this hyperbola shrunk toward the origin as in ridge regression,

$$w_2 w_1 = (1 + \lambda x^{-2})^{-1}.$$

For $\mathcal{L}_\sigma$ the critical points depend on the scale of λ relative to x^2. For $\lambda < x^2$, the origin is a saddle point and the global

---

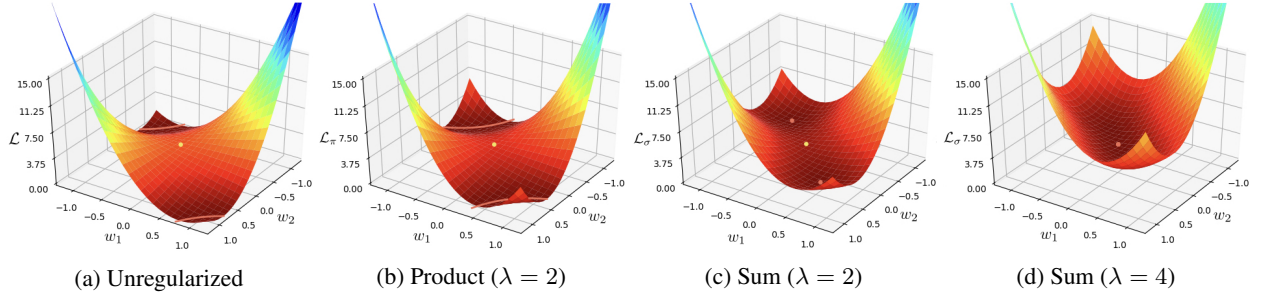| (a) Unregularized | (b) Product ($\lambda = 2$) | (c) Sum ($\lambda = 2$) | (d) Sum ($\lambda = 4$) |

Figure 1. Scalar loss landscapes with $x^2 = 4$. Yellow points are saddles and red curves and points are global minima.

minima are the two isolated rank-1 critical points[5] cut out by the equations

$$w_2 w_1 = 1 - \lambda x^{-2}, \quad w_1 = w_2.$$

As λ increases toward x^2, these minima move toward the origin, which remains a saddle point. As λ exceeds x^2, the origin becomes the unique global minimum. This loss of information was our first hint at the connection to probabilistic PCA formalized in Theorem 3.1.

2.2. Regularization and orthogonality

Adding L_2-regularization to the encoder and decoder separately reduces the symmetries of the loss from $\mathrm{GL}_k(\mathbb{R})$ to the orthogonal group $O_k(\mathbb{R})$. Two additional facts about the relationship between regularization and orthogonality have guided our intuition:

(a) Orthogonal matrices are the determinant ± 1 matrices of minimal Frobenius norm[6],

$$\underset{A}{\arg\min} ||A||_F^2 \quad \text{s.t.} \quad \det(A)^2 = 1.$$

(b) Orthogonal matrices are the inverse matrices of minimum total squared Frobenius norm,

$$\underset{A,B}{\arg\min} ||A||_F^2 + ||B||_F^2 \quad \text{s.t.} \quad AB = I,$$

and in particular $A = B^\mathsf{T}$ at all minima.

Both facts follow from the inequality of arithmetic and geometric means after casting the problems in terms of squared singular values[7]. While it was not immediately clear to us that the transpose relationship in (b) also holds at the minima of $\mathcal{L}_\sigma$, in fact, all critical points of an L$_2$-regularized linear autoencoder are symmetric:

---

[5]Identified with the components of $\mathrm{O}_1(\mathbb{R}) \cong \{\pm 1\}$.

[6]Geometrically: the unit-volume parallelotope of minimal total squared side length is the unit hypercube.

[7]The squared Frobenius norm is their sum and the squared determinate is their product.

Theorem 2.1 (Transpose Theorem). *All critical points of $\mathcal{L}_\sigma$ satisfy $W_1 = W_2^\mathsf{T}$.*

Our proof uses elementary properties of positive definite matrices as reviewed in Appendix B. Note that without regularization, all critical points are psuedoinverses, $W_1 = W_2^+$, as is clear in the scalar case and derived in Section 4.

Proof. Critical points of $\mathcal{L}_\sigma$ satisfy:

$$\frac{\partial \mathcal{L}_\sigma}{\partial W_1} = 2W_2^\mathsf{T}(W_2 W_1 - I)XX^\mathsf{T} + 2\lambda W_1 = 0,$$

$$\frac{\partial \mathcal{L}_\sigma}{\partial W_2} = 2(W_2 W_1 - I)XX^\mathsf{T}W_1^\mathsf{T} + 2\lambda W_2 = 0.$$

We first prove that the matrix

$$C = (I - W_2 W_1)XX^\mathsf{T}$$

is positive semi-definite[8]. Rearranging $\frac{\partial \mathcal{L}_\sigma}{\partial W_2} W_2^\mathsf{T}$ gives

$$XX^\mathsf{T}(W_2 W_1)^\mathsf{T} = (W_2 W_1)XX^\mathsf{T}(W_2 W_1)^\mathsf{T} + \lambda W_2 W_2^\mathsf{T}.$$

Both terms on the right are positive semi-definite, so their sum on the left is as well and therefore

$$XX^\mathsf{T}(W_2 W_1)^\mathsf{T} \succeq (W_2 W_1)XX^\mathsf{T}(W_2 W_1)^\mathsf{T}.$$

Cancelling $(W_2 W_1)^\mathsf{T}$ via Lemma B.1 gives $C \succeq 0$.

We now show the difference $A = W_1 - W_2^\mathsf{T}$ is zero. Rearranging terms using the symmetry of C gives

$$0 = \frac{\partial \mathcal{L}_\sigma}{\partial W_1} - \frac{\partial \mathcal{L}_\sigma}{\partial W_2}^\mathsf{T} = 2A(C + \lambda I).$$

Since $C \succeq 0$ and $\lambda > 0$ imply $C + \lambda I \succ 0$, we conclude from

$$A(C + \lambda I)A^T = 0$$

that $A = 0$. $\qquad\square$

---

[8]Intuitively, we expect this property so long as $W_2 W_1$ shrinks the principal directions of X, so that $I - W_2 W_1$ does as well.

3. Bayesian Models

In this section, we identify Bayesian counterparts of our three loss functions and derive a novel connection between (regularized) LAEs and (probabilistic) PCA. This connection enables the application of any LAE training method to Bayesian MAP estimation of the corresponding model.

Consider the rank-k (self-)regression model

$$x_i = W_2 W_1 x_i + \varepsilon_i = W x_i + \varepsilon_i$$

where W_1 and W_2 act through their product W and $\varepsilon_i \sim \mathcal{N}_m(0, 1)$.

- $\mathcal{L}$ is *rank-k regression*. The prior on W is the uniform distribution on $\mathbb{R}^{m \times m}$ restricted to rank-k matrices[9].
- $\mathcal{L}_\pi$ is *rank-k ridge regression*. The prior on W is $\mathcal{N}_{m \times m}(0, \lambda^{-1})$ restricted to rank-k matrices.
- $\mathcal{L}_\sigma$ is the model with W_1 and W_2^{T} independently drawn from $\mathcal{N}_{k \times m}(0, \lambda^{-1})$.

Theorem 4.2 shows that the minima of $\mathcal{L}_\sigma$, or equivalently the MAP of the Bayesian model, are such that W_2 is the orthogonal projection onto the top k principal directions followed by compression in direction i via multiplication by the factor $(1 - \lambda \sigma_i^{-2})^{\frac{1}{2}}$ for $\sigma_i^2 > \lambda$ and zero otherwise. Notably, for principal directions with eigenvalues dominated by λ, all information is lost no matter the number of data points. The same phenomenon occurs for pPCA with respect to eigenvalues dominated by the variance of the noise, σ^2. Let's consider these Bayesian models side by side, with $W_0 \in \mathbb{R}^{m \times k}$ the parameter of pPCA[10]:

Bayesian $\mathcal{L}_\sigma$	pPCA
$W_1, W_2^{\mathsf{T}} \sim \mathcal{N}_{k \times m}(0, \lambda^{-1})$	$z_i \sim \mathcal{N}_k(0, 1)$
$\varepsilon_i \sim \mathcal{N}_m(0, 1)$	$\varepsilon_i \sim \mathcal{N}_m(0, \sigma^2)$
$x_i = W_2 W_1 x_i + \varepsilon_i$	$x_i = W_0 z_i + \varepsilon_i$

Comparing the critical points of $\mathcal{L}_\sigma$ in Theorem 4.2,

$$W_1^T = W_2 = U_{\mathcal{I}}(I_\ell - \lambda \Sigma_{\mathcal{I}}^{-2})^{\frac{1}{2}} O^{\mathsf{T}}, \tag{2}$$

and pPCA (Tipping & Bishop, 1999),

$$W_0 = U_{\mathcal{I}} \Sigma_{\mathcal{I}} (I_\ell - \sigma^2 \Sigma_{\mathcal{I}}^{-2})^{\frac{1}{2}} O^{\mathsf{T}}, \tag{3}$$

where $O \in \mathbb{R}^{k \times \ell}$ has orthonormal columns, we see that λ corresponds to σ^2 (rather than the precision σ^{-2}) in the sense that principal directions with eigenvalues dominated by either are collapsed to zero. The critical points only differ in the factor by which the the remaining principal directions are shrunk. More precisely:

[9]Note that rank-(k-1) matrices are a measure zero subset of rank-k matrices.

[10]See Chapter 12.2 of Bishop (2006) for background on pPCA.

Theorem 3.1 (pPCA Theorem). *With $\sigma^2 = \lambda$, the critical points of*

$$\mathcal{L}_\sigma^0(W_1, W_2) = \mathcal{L}_\sigma(W_1(XX^{\mathsf{T}})^{-\frac{1}{2}}, (XX^{\mathsf{T}})^{-\frac{1}{2}} W_2)$$

coincide with the critical points of pPCA.

Proof. Multiplying the expression for W_2 in (2) on the left by $(XX^{\mathsf{T}})^{\frac{1}{2}}$ gives the expression for W_0 in (3). □

Interestingly, the generative model for $\mathcal{L}_\sigma^0$ differs from that of pPCA. For example, in the scalar case $\mathcal{L}_\sigma^0(w_1, w_2)$ is

$$x^2 \left(1 - x^{-2} w_2 w_1\right)^2 + \lambda x^{-2}(w_1^2 + w_2^2)$$

whereas the negative log likelihood of pPCA is

$$\frac{1}{2} \left(\ln(2\pi) + \ln(w_0^2 + \sigma^2) + x^2(w_0^2 + \sigma^2)^{-1}\right).$$

4. Loss Landscapes

Having contextualized LAE models in a Bayesian framework, we now turn to understanding their loss landscapes. Symmetries such as (1) exist because the model is expressed in an "overparameterized" coordinate form rooted in classical linear algebra. This results in "flat" critical manifolds rather than a finite number of critical points. In Section 4.1, we remove all symmetries by expressing the loss geometrically over a topological domain. This results in $\binom{m}{k}$ critical points, and in particular a unique minimum. This intuition will pay off in Sections 4.2 and 4.3, where we fully characterize the critical manifolds and local curvatures of all three LAE loss landscapes.

4.1. Critical points

We now consider reconstruction loss over the domain of k-dimensional planes through the origin in $\mathbb{R}^m$. This space has the structure of a $k(m - k)$-dimensional smooth, compact manifold called the Grassmannian of k-planes in $\mathbb{R}^m$ and denoted $\mathrm{Gr}_k(\mathbb{R}^m)$ (Hatcher, 2002). We'll build intuition with a few simple examples.

- $\mathrm{Gr}_1(\mathbb{R}^2)$ is the space of lines through the origin in the plane, which may be smoothly parameterized by a counterclockwise angle of rotation of the x-axis modulo the half turn that maps a line to itself.
- $\mathrm{Gr}_1(\mathbb{R}^3)$ is the space of lines through the origin in 3-space, also known as the real projective plane. We can visualize $\mathrm{Gr}_1(\mathbb{R}^3)$ as the northern hemisphere of the 2-sphere with equator glued by the antipodal map.
- $\mathrm{Gr}_2(\mathbb{R}^3)$ is identified with $\mathrm{Gr}_1(\mathbb{R}^3)$ by mapping a plane to its 1-dimensional orthogonal complement.

A point cloud X in $\mathbb{R}^m$ determines a smooth function

$$\mathcal{L}_X : \mathrm{Gr}_k(\mathbb{R}^m) \to \mathbb{R}$$

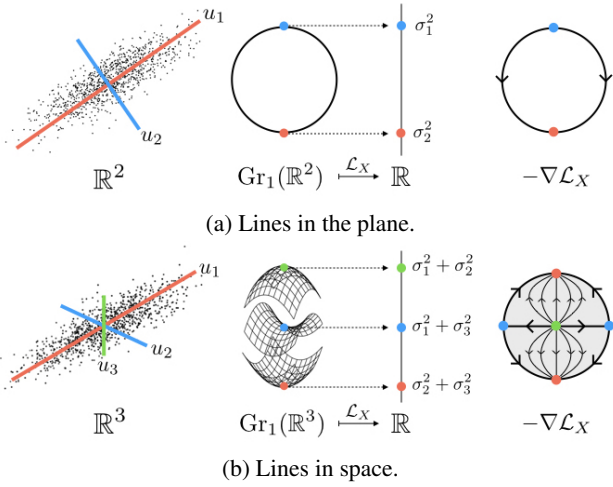(a) Lines in the plane.

(b) Lines in space.

Figure 2. Left: Principal directions of a point cloud X. *Middle:* $\mathcal{L}_X$ as height function on the manifold of lines through the origin. *Right:* Negative gradient flow of $\mathcal{L}_X$.

whose value on a k-plane is the sum of square distances from the points to the plane. Figure 2 depicts $\mathcal{L}_X$ as a height function for $\mathrm{Gr}_1(\mathbb{R}^2)$ and $\mathrm{Gr}_1(\mathbb{R}^3)$. Note that the min and max in (a) are the principal directions u_1 and u_2, while the min, saddle, and max in (b) are the principal directions u_1, u_2, and u_3. At right, we depict the negative gradient flow $-\nabla\mathcal{L}_X$ on each space, with (b) represented as a disk with glued boundary. In (a), u_1 may descend to u_2 by rotating clockwise or counterclockwise[11]. In (b), u_1 may descend to u_2 or u_3 by rotating in either of two directions in the plane they span, and u_2 may similarly descend to u_3.

The following theorem requires our assumption that the singular values of X are distinct.

Theorem 4.1 (Grassmannian Theorem). *$\mathcal{L}_X$ is a smooth function with $\binom{m}{k}$ critical points given by all rank-k principal subspaces. In local coordinates near the critical point with principal directions $i_1 < \ldots < i_k$, $\mathcal{L}_X$ takes the form of a standard non-degenerate saddle with*

$$d_{\mathcal{I}} = \sum_{j=1}^{k}(i_j - j). \tag{4}$$

descending directions.

The latter formula counts the total number of pairs $i < j$ with $i \in \mathcal{I}$ and $j \notin \mathcal{I}$. These correspond to directions to flow along $-\nabla\mathcal{L}_X$ by rotating one principal direction u_i to another u_j of higher eigenvalue, fixing the rest.

In Appendix E, we prove a stronger form of the Grassman-

[11]Formally, we mean there are two gradient trajectories (modulo time translation) that converge to u_1 and u_2 in each time direction asymptotically, namely the left and right halves of the circle.

nian Theorem by combining Theorem 4.2, a commutative diagram (10) relating $\mathcal{L}_X$ and $\mathcal{L}$, and techniques from algebraic topology.

4.2. Critical manifolds

Translating the Grassmannian Theorem back to the coordinate representation $\mathbb{R}^{k \times m} \times \mathbb{R}^{m \times k}$ introduces two additional phenomena we saw in the scalar case in Section 2.1.

- Each critical point on $\mathrm{Gr}_k(\mathbb{R}^m)$ corresponds to a manifold $\mathrm{GL}_k(\mathbb{R})$ or $\mathrm{O}_k(\mathbb{R})$ of rank-k critical points.
- Critical manifolds appear with rank less than k. In particular, $(0, 0)$ is a critical point for all three losses.

Now let's now combine our topological and scalar intuition to understand the the loss landscapes of LAEs in all dimensions and for all three losses.

Theorem 4.2 requires our assumption[12] that X has distinct singular values $\sigma_1 > \cdots > \sigma_m > 0$. Let $u_1, \ldots, u_m$ denote the corresponding left singular vectors (or principal directions) of X. For an index set $\mathcal{I} \subset \{1, \ldots, m\}$ we define:

- $\ell = |\mathcal{I}|$ and increasing indices $i_1 < \cdots < i_\ell$,
- $\Sigma_{\mathcal{I}} = \mathrm{diag}(\sigma_{i_1}, \ldots, \sigma_{i_\ell}) \in \mathbb{R}^{\ell \times \ell}$,
- $U_{\mathcal{I}} \in \mathbb{R}^{m \times \ell}$ consisting of columns $i_1, \ldots, i_\ell$ of U,
- $F_{\mathcal{I}}$, the open submanifold of $\mathbb{R}^{k \times \ell}$ whose points are matrices G with independent columns[13],
- $V_{\mathcal{I}}$, the closed submanifold of $\mathbb{R}^{k \times \ell}$ whose points are matrices O with orthonormal columns[14].

Theorem 4.2 (Landscape Theorem). *For each loss, the critical points form a smooth submanifold of $\mathbb{R}^{k \times m} \times \mathbb{R}^{m \times k}$.*

For $\mathcal{L}$ and $\mathcal{L}_\pi$, this submanifold is diffeomorphic to the disjoint union of $F_{\mathcal{I}}$ over all $\mathcal{I} \subset \{1, \ldots, m\}$ of size at most k.

For $\mathcal{L}_\sigma$, this submanifold is diffeomorphic to the disjoint union of $V_{\mathcal{I}}$ over all $\mathcal{I} \subset \{1, \ldots, m_0\}$ of size at most k, where m_0 is the largest index such that $\sigma_{m_0}^2 > \lambda$.

These diffeomorphisms map $G \in F_{\mathcal{I}}$ or $O \in V_{\mathcal{I}}$ to a critical point (W_1, W_2) as follows:

	W_2	W_1
$\mathcal{L}$	$U_{\mathcal{I}}G^+$	$GU_{\mathcal{I}}^{\mathsf{T}}$
$\mathcal{L}_\pi$	$U_{\mathcal{I}}(I_\ell + \lambda\Sigma_{\mathcal{I}}^{-2})^{-\frac{1}{2}}G^+$	$G(I_\ell + \lambda\Sigma_{\mathcal{I}}^{-2})^{-\frac{1}{2}}U_{\mathcal{I}}^{\mathsf{T}}$
$\mathcal{L}_\sigma$	$U_{\mathcal{I}}(I_\ell - \lambda\Sigma_{\mathcal{I}}^{-2})^{\frac{1}{2}}O^{\mathsf{T}}$	$O(I_\ell - \lambda\Sigma_{\mathcal{I}}^{-2})^{\frac{1}{2}}U_{\mathcal{I}}^{\mathsf{T}}$

The proof of the Landscape Theorem 4.2 follows quickly

[12]For the sum loss, we also assume λ is distinct from all σ_i^2.
[13]Also known as the manifold of l-frames in $\mathbb{R}^k$.
[14]Also known as the Stiefel manifold $V_l(\mathbb{R}^k)$.

from the Transpose Theorem 2.1 and Proposition 4.3.

Proposition 4.3. *Let $D, S \in \mathbb{R}^{m \times m}$ be diagonal matrices such that S is invertible and the diagonal of $D^2 S^{-2} D^2$ has distinct non-zero elements. Then the critical points of*

$$\mathcal{L}_*(Q_1, Q_2) = \text{tr}(Q_2 Q_1 S^2 Q_1^\mathsf{T} Q_2^\mathsf{T} - 2Q_2 Q_1 D^2)$$

are smoothly parameterized as the disjoint union of $F_\mathcal{I}$ over all $\mathcal{I} \subset \{1, \ldots, m\}$ of size at most k. The diffeomorphism maps $G \in F_\mathcal{I}$ to a critical point (Q_1, Q_2) as follows[15]:

	Q_2	Q_1
$\mathcal{L}_*$	$D_\mathcal{I} S_\mathcal{I}^{-1} I_\mathcal{I} G^+$	$G I_\mathcal{I}^\mathsf{T} D_\mathcal{I} S_\mathcal{I}^{-1}$

Proof of Theorem 4.2. Given the singular value decomposition $X = U\Sigma V^\mathsf{T}$, let $Q_1 = W_1 U$ and $Q_2 = U^\mathsf{T} W_2$. By invariance of the Frobenius norm under the smooth action of the orthogonal group, we may instead parameterize the critical points of the following loss functions and then pull the result back to $W_1 = Q_1 U^\mathsf{T}$ and $W_2 = U Q_2$:

$$\mathcal{L}(Q_1, Q_2) = ||\Sigma - Q_2 Q_1 \Sigma||_F^2$$
$$\mathcal{L}_\pi(Q_1, Q_2) = \mathcal{L}(Q_1, Q_2) + \lambda ||Q_2 Q_1||_F^2$$
$$\mathcal{L}_\sigma(Q_1, Q_2) = \mathcal{L}(Q_1, Q_2) + \lambda(||Q_1||_F^2 + ||Q_2||_F^2)$$

$\mathcal{L}$ expands to

$$\text{tr}(Q_2 Q_1 \Sigma^2 Q_1^\mathsf{T} Q_2^\mathsf{T} - 2Q_2 Q_1 \Sigma^2 + \Sigma^2).$$

By Proposition 4.3 with $S = \Sigma$ and $D = \Sigma$, the critical points have the form

	Q_2	Q_1
$\mathcal{L}$	$I_\mathcal{I} G^+$	$G I_\mathcal{I}^\mathsf{T}$

$\mathcal{L}_\pi$ expands to

$$\text{tr}(Q_2 Q_1 (\Sigma^2 + \lambda I) Q_1^\mathsf{T} Q_2^\mathsf{T} - 2Q_2 Q_1 \Sigma^2 + \Sigma^2).$$

By Proposition 4.3 with $S = (\Sigma^2 + \lambda I)^{\frac{1}{2}}$ and $D = \Sigma$, the critical points have the form

	Q_2	Q_1
$\mathcal{L}_\pi$	$I_\mathcal{I}(I_\ell + \lambda \Sigma_\mathcal{I}^{-2})^{-\frac{1}{2}} G^+$	$G(I_\ell + \lambda \Sigma_\mathcal{I}^{-2})^{-\frac{1}{2}} I_\mathcal{I}^\mathsf{T}$

By Lemma A.1 with $A = Q_2$ and $B = Q_1$, $\mathcal{L}_\sigma$ expands to the sum of two functions:

$$\mathcal{L}_1(Q_1, Q_2) = \text{tr}(Q_2 Q_1 \Sigma^2 Q_1^\mathsf{T} Q_2^\mathsf{T} - 2Q_2 Q_1 (\Sigma^2 - \lambda I) + \Sigma^2)$$
$$\mathcal{L}_2(Q_1, Q_2) = \lambda ||Q_1 - Q_2^\mathsf{T}||_F^2.$$

---

[15]Here $D_\mathcal{I}$ and $S_\mathcal{I}$ are defined like $\Sigma_\mathcal{I}$. $I_\mathcal{I}$ is defined like $U_\mathcal{I}$.

So at a critical point, $\nabla \mathcal{L}_1(Q_1, Q_2) = -\nabla \mathcal{L}_2(Q_1, Q_2)$ and $Q_1 = Q_2^\mathsf{T}$ by the Transpose Theorem 2.1. The latter also implies $\nabla \mathcal{L}_2(Q_1, Q_2) = 0$. So the critical points of $\mathcal{L}_\sigma$ coincide with the critical points of $\mathcal{L}_1$ such that $Q_1 = Q_2^\mathsf{T}$.

By Proposition 4.3 with $S = \Sigma$ and $D = (\Sigma^2 - \lambda I)^{\frac{1}{2}}$, these critical points have the form

	Q_2	Q_1
$\mathcal{L}_\sigma$	$I_\mathcal{I}(I_\ell - \lambda \Sigma_\mathcal{I}^{-2})^{\frac{1}{2}} O^\mathsf{T}$	$O(I_\ell - \lambda \Sigma_\mathcal{I}^{-2})^{\frac{1}{2}} I_\mathcal{I}^\mathsf{T}$

In particular, real solutions do not exist for $\sigma_i^2 < \lambda$. $\square$

4.3. Local curvature

By Theorem 4.2, the critical landscape is a disjoint union of smooth manifolds, each at some height. We now prove that the Hessian is non-degenerate in the normal directions to the critical landscape. As discussed in Appendix E, such functions are called *Morse-Bott* and studied extensively in differential and algebraic topology.

Theorem 4.4 (Curvature Theorem). *In local coordinates near any point on the critical manifold indexed by $\mathcal{I}$, all three losses take the form of a standard degenerate saddle with $d_\mathcal{I} + (k - \ell)(m - \ell)$ descending directions.*

- *$\mathcal{L}$ and $\mathcal{L}_\pi$ have $k\ell$ flat directions.*
- *$\mathcal{L}_\sigma$ has $k\ell - \binom{\ell+1}{2}$ flat directions.*

The remaining directions are ascending.

Proof. In addition to the descending directions of $\mathcal{L}_X$, there are $(k - \ell)(m - \ell)$ more that correspond to scaling one of $k - \ell$ remaining slots in G or O toward one of $m - \ell$ available principal directions[16]. For $\mathcal{L}$ and $\mathcal{L}_\pi$, the ascending directions are the $k(m - k) - d_\mathcal{I}$ ascending directions of $\mathcal{L}_X$; for $\mathcal{L}_\sigma$, an additional $\binom{\ell+1}{2}$ ascending directions preserve the reconstruction term while increasing the regularization term by decreasing orthogonality. The remaining (flat) directions are tangent to the critical manifold, itself diffeomorphic to the manifold of (orthonormal) l-frames in $\mathbb{R}^k$. $\square$

5. Empirical Illustration

In this section, we illustrate the Landscape Theorem 4.2 by training an LAE on synthetic data and visualizing properties of the learned weight matrices. See Appendix D.1 for experiments on real data. Corollary 4.4 implies that gradient descent and its extensions will reach a global minimum, regardless of the initialization, if trained for a sufficient number of epochs with a small enough learning rate (Zhu

---

[16]In Figure 1c, these two gradient trajectories descend from the yellow saddle at 0 to the two red minima at $\pm u_1$.
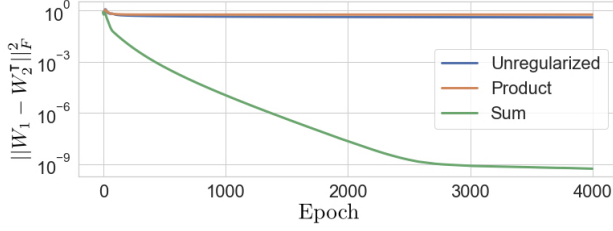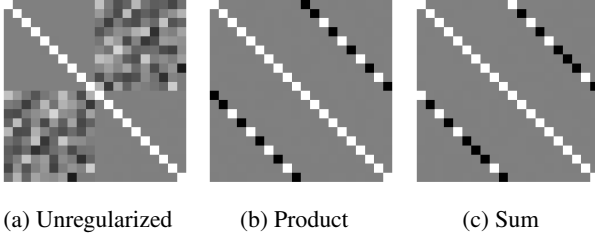
Figure 3. Distance between W_1 and $W_2^\intercal$ during training.

(a) Unregularized   (b) Product   (c) Sum

Figure 4. Heat map of the matrix $\begin{bmatrix} U & V_* \end{bmatrix}^\intercal \begin{bmatrix} U & U_* \end{bmatrix}$. Black and white correspond to -1 and 1, respectively.

et al., 2018).

5.1. Synthetic data

In the following experiments, we set $k = \lambda = 10$ and fix a data set $X \in \mathbb{R}^{20 \times 20}$ with singular values $\sigma_i = i$ and random left and right singular vectors under Haar measure. We train the LAE for each loss using the Adam optimizer for 4000 epochs with random normal initialization, full batch, and learning rate 0.05.

Figure 3 tracks the squared distance between W_1 and $W_2^\intercal$ during training. Indeed, only the the sum loss pulls the encoder and (transposed) decoder together as claimed in the Transpose Theorem 2.1.

Let W_* be the product $W_2 W_1$ after training and fix the singular value decompositions

$$X = U \Sigma V^\intercal, \quad W_* = U_* \Sigma_* V_*^\intercal.$$

For each loss, the heat map of

$$\begin{bmatrix} U^\intercal U & U^\intercal U_* \\ V_*^\intercal U & V_*^\intercal U_* \end{bmatrix}$$

in Figure 4 is consistent with W_* approximating a global minimum defined by the Landscape Theorem. Namely, the lower right quadrant shows $U_* \approx V_*$ for each loss and the upper right and lower left quadrants show $U \approx U_*$ and $U \approx V_*$ up to column sign for the product and sum losses, *but not for the unregularized loss*. That is, for the product and sum losses, the left singular vectors of X are obtained as the right and left singular vectors of W_*.

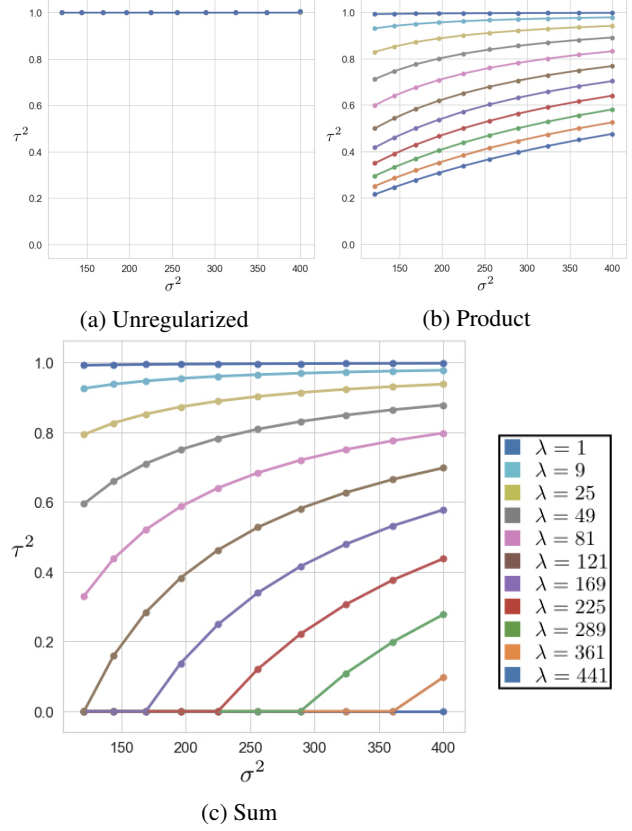The Landscape Theorem also gives explicit formulae for the eigenvalues of W_* at convergence. Letting σ_i^2 and τ_i^2 be

(c) Sum

Figure 5. Illustration of the relationship between the eigenvalues of the weight matrix (τ^2) and data matrix (σ^2) for various values of λ. Points are the empirical and lines are theoretical.

the ith largest eigenvalues of $XX^\intercal$ and W_*, respectively, in Figure 5 we plot the points $\left(\sigma_i^2, \tau_i^2 \right)$ for many values of λ. We superimpose a curve for each value of λ defined by the theoretical relationship between σ_i^2 and τ_i^2 in the Landscape Theorem. The (literal) alignment of theory and practice is visually perfect.

6. Implications

6.1. PCA algorithms

The Landscape Theorem for $\mathcal{L}_\sigma$ implies those principal directions of X with eigenvalues greater than λ coincide with the top left singular vectors of the trained decoder. Hence we can perform PCA by training a regularized LAE and computing SVD of the decoder. For example, full-batch gradient descent corresponds to the following algorithm, which is guaranteed to converge to a (global) minimum for sufficiently small learning rate by the Curvature Theorem.

Algorithm 1 LAE–PCA

> **input** $X \in \mathbb{R}^{m \times n}$; $k \le m$; $\lambda, \alpha > 0$
> **initialize** $W_1, W_2^\intercal \in \mathbb{R}^{k \times m}$
> **while not converged**
> $\quad W_1 \mathrel{-}= \alpha \left(W_2^\intercal (W_2 W_1 - I) X X^\intercal + \lambda W_1 \right)$
> $\quad W_2 \mathrel{-}= \alpha \left((W_2 W_1 - I) X X^\intercal W_1^\intercal + \lambda W_2 \right)$
> $U, \Sigma, \_ = \text{SVD}(W_2)$
> **return** $U, \ \lambda (I - \Sigma^2)^{-1}$

Note the SVD step is trivial because the decoder is only $m \times k$ dimensional with $k \ll n$. Hence optimizing this formulation of PCA reduces to optimizing the training of a regularized LAE. We have so far explored several simple ideas for optimizing performance[17]:

- *First order.* The Transpose Theorem suggests "tying" $W_1 = W_2^\intercal$ *a priori*. In fact, for $\lambda = 0$ and $k = 1$, the gradient update for $\mathcal{L}_\sigma$ with tied weights is equivalent to Oja's rule for how neurons in the brain adapt synaptic strength (Oja, 1982). See Appendix A for a derivation.
- *Second order.* The loss is convex in each parameter fixing the other, so W_1 and W_2 may be updated exactly per iteration by solving a system of m or k linear equations respectively.

While many other methods have been proposed for recovering principal components with neural networks, they all require specialized algorithms for iteratively updating weights, similar to classical numerical SVD approaches Warmuth & Kuzmin (2007); Feng et al. (2013b;a). By contrast, reducing PCA to a small SVD problem by training a regularized LAE more closely parallels randomized SVD (Halko et al., 2011). We hope others will join us in investigating how far one can push the performance of LAE-PCA.

6.2. Neural alignment

In gradient descent via backpropagation, the weight matrix before a layer is updated by an error signal that is propagated by the transposed weight matrix after the layer. In the brain, there is no known physical mechanism to reuse feedforward connections for feedback or to enforce weight symmetry between distinct forward and backward connections. The latter issue is known as the *weight transport problem* (Grossberg, 1987) and regarded as central to the implausibility of backpropagation in the brain. Recently Lillicrap et al. (2016) showed that forward weights can sufficiently align to fixed, random feedback weights to support learning in shallow networks, but Bartunov et al. (2018) demonstrated that this *feedback alignment* and other biologically-plausible archi-

tectures break down for deep networks.

We are now investigating two approaches to weight transport inspired by the results herein. First, the Transpose Theorem 2.1 suggests that weight symmetry emerges dynamically when feedback weights are updated to optimize forward-backward reconstruction between consecutive layers with weight decay. This *information alignment* algorithm balances task prediction, information transmission, and energy efficiency. Second, Lemma A.1 expresses weight symmetry as balancing weight decay and self-amplification, lending greater biological plausibility to *symmetric alignment*. We have verified that both information and (not surprisingly) symmetric alignment indeed align weights and are competitive with backprop when trained on MNIST and CIFAR10. In parallel, Akrout et al. (2019) has independently shown that similar forms of local dynamic alignment scale to ImageNet.

6.3. Morse homology

In Appendix E we expand from the algebraic topology of learning PCA to that of learning by gradient descent in general. For example, we explain how Morse homology provides a principled foundation for the empirical observation of low-lying valley passages between minima in Garipov et al. (2018). We are hopeful that this perspective will yield insights for efficient training and improved robustness and interpretation of consensus representations and ensemble predictions, for non-convex models arising in matrix factorization and deep learning.

7. Conclusion

In 1989, Baldi & Hornik (1989) characterized the loss landscape of an LAE. In 2018, Zhou & Liang (2018) characterized the loss landscape of an autoencoder with ReLU activations on a single hidden layer. This paper fills out and ties together the rich space of research on linear networks over the last forty years by deriving from first principles a unified characterization of the loss landscapes of LAEs with and without regularization, while introducing a rigorous topological lens. By considering a simple but fundamental model with respect to regularization, orthogonality, and Morse homology, this work also suggests new principles and algorithms for learning.

---

[17]NumPy implementations and benchmarks of these ideas are available on GitHub.

Acknowledgements

References

Akrout, M., Wilson, C., Humphreys, P. C., Lillicrap, T., and Tweed, D. Deep Learning without Weight Transport. 2019.

Alain, G., Bengio, Y., and Rifai, S. Regularized Auto-Encoders Estimate Local Statistics. *CoRR*, abs/1211.4246, 2012.

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

Banyaga, A. and Hurtubise, D. *Lectures on Morse Homology*. Springer Netherlands, 2004.

Bartunov, S., Santoro, A., Richards, B. A., Marris, L., Hinton, G. E., and Lillicrap, T. P. Assessing the Scalability of Biologically-Motivated Deep Learning Algorithms and Architectures. In *NIPS*, pp. 9390–9400, 2018.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Bloom, J. M. The Local Structure of Smooth Maps of Manifolds, 4 2004. URL http://www.math.harvard.edu/theses/phd/bloom/ThesisXFinal.pdf.

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

Choromanska, A., Henaff, M., Mathieu, M., Ben Arous, G., and LeCun, Y. The loss surfaces of multilayer networks. *Journal of Machine Learning Research*, 38:192–204, 2015.

Duan, H. Morse functions and cohomology of homogeneous spaces. 2004.

Eckart, C. and Young, G. The Approximation of One Matrix by Another of Lower Rank. *Psychometrika*, 1(3):211–218, 1936.

Feng, J., Xu, H., Mannor, S., and Yan, S. Online PCA for Contaminated Data. In *NIPS*, 2013a.

Feng, J., Xu, H., and Yan, S. Online Robust PCA via Stochastic Optimization. In *NIPS*, 2013b.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., and Wilson, A. G. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. In *NIPS*, 2018.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Grossberg, S. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11(1):23 – 63, 1987.

Halko, N., Martinsson, P. G., and Tropp, J. A. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Rev.*, 53(2):217–288, May 2011.

Hansen, M. C. Morse Theory on Complex Grassmannians. Master's thesis, University of Copenhagen, 6 2012.

Hatcher, A. *Algebraic Topology*. Cambridge University Press, 2002.

Izenman, A. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5:248–264, 06 1975.

Josse, J. and Wager, S. Bootstrap-Based Regularization for Low-Rank Matrix Estimation. *Journal of Machine Learning Research*, 17(124):1–29, 2016.

LeCun, Y. and Cortes, C. The MNIST Database of handwritten digits. http://yann.lecun.com/exdb/mnist/.

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7, 2016.

Mehta, D., Chen, T., Tang, T., and Hauenstein, J. D. The loss surface of deep linear networks viewed through the algebraic geometry lens. *CoRR*, abs/1810.07716, 2018.

Mianjy, P., Arora, R., and Vidal, R. On the Implicit Bias of Dropout. In *ICML*, 2018.

Milnor, J. *Morse Theory*. Princeton University Press, 1963.

Mukherjee, A. and Zhu, J. Reduced Rank Ridge Regression and Its Kernel Extensions. *Statistical Analysis and Data Mining*, 4(6):612–622, 2011.

Oja, E. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982.

Plaut, E. From Principal Subspaces to Principal Components with Linear Autoencoders. *ArXiv e-prints*, 2018.

Pretorius, A., Kroon, S., and Kamper, H. Learning Dynamics of Linear Denoising Autoencoders. In *ICML*, 2018.

Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. Contractive Auto-Encoders: Explicit Invariance During Feature Extraction. In *ICML*, 2011.

Solgun, M. Perfect Morse Function on $SO(n)$. *European Journal of Pure and Applied Mathematics*, 9(3):314–321, 2016.

Tipping, M. E. and Bishop, C. M. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society*, 61(3):611–622, 1999.

van den Bos, A. *Parameter Estimation for Scientists and Engineers*, chapter Appendix C. John Wiley Sons, 2007.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

Warmuth, M. K. and Kuzmin, D. Randomized Online PCA Algorithms with Regret Bounds that are Logarithmic in the Dimension. 2007.

Zhou, Y. and Liang, Y. Critical Points of Linear Neural Networks: Analytical Forms and Landscape Properties. In *ICLR*, 2018.

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. Global Optimality in Low-Rank Matrix Optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.