Regularized Linear Autoencoders and Topology of Loss Landscapes

Stefan Petrevski
Christ Church, Oxford

Abstract

Guided by recent deep learning literature, we explore classical and topological results on the loss landscapes of regularized linear autoencoders (LAEs). We illustrate the main theorems in original settings using the Fashion-MNIST and CIFAR-10 datasets. Moreover, we generate new artificial data to investigate the limitations of over-regularization in context of the theory presented, and propose an intermediate synthetic loss with surprising properties to resolve these issues.

1 Introduction and Motivation

Autoencoders are neural networks designed to copy the networks' inputs to their outputs. This property has brought them to the forefronts of feature learning and generative modeling (Goodfellow et al. [2016]). To illustrate their structure, we consider a data matrix $X \in \mathbb{R}^{m \times n}$, consisting of points $x_1, ..., x_n$ in $\mathbb{R}^m$. First, an **encoder** $f : \mathbb{R}^m \to \mathbb{R}^k$, sends x_i to a latent representation $h_i = f(x_i)$, whereas the **decoder** $g : \mathbb{R}^k \to \mathbb{R}^m$ reconstructs $h_i \mapsto g(f(x_i))$. We are interested in undercomplete autoencoders with $k < \min(m, n)$, which learn the most important features of the data due to restricted dimensionality. In this report, we study linear autoencoders (**LAEs**) with identity activations, so that $f(x_i) = W_1 x_i$ and $g(h_i) = W_2 h_i$, for $W_1 \in \mathbb{R}^{k \times m}$ and $W_2 \in \mathbb{R}^{m \times k}$.

A common and natural loss function for training is the **unregularized** squared Frobenius norm:

$$\mathcal{L}(W_1, W_2) = \|X - W_2 W_1 X\|_F^2 := \text{tr}\left((X - W_2 W_1 X)^T (X - W_2 W_1 X)\right). \tag{1}$$

For a fixed k, it is well known that the optimal $W_2 W_1$ is the rank-k projector of X onto the subspace spanned by its leading k principal directions, and as such is equivalent to PCA (Eckart and Young [1936]). However, for any M in the group of invertible $k \times k$ matrices $\text{GL}_k(\mathbb{R})$, it is clear that $\mathcal{L}(W_1, W_2) = \mathcal{L}(MW_1, W_2 M^{-1})$. In fact, this symmetry may deform W_1 and W_2 during traning, thereby preventing LAEs from recovering the principal directions themselves. Motivated by this issue, we explore how appropriate regularization of $\mathcal{L}$ can not only help recover the principal directions, but will also lead to the full classification of critical points and loss landscapes.

2 Regularization and Preliminary Results

Following Bourlard and Kamp [1988], we may assume that X in equation (1) is mean centered, by appropriately absorbing any bias parameters. This implies that given a singular value decomposition $X = U \Sigma V^T$ with the diagonal entries of Σ descending, the columns of U are the **principal directions** of X (i.e. the eigenvectors of the covariance of X, here also being left **singular vectors**).

Our first regularizer is inspired by the **denoising** autoencoder, which aims to reconstruct X from corrupted input $\tilde{X}$. Pretorius et al. [2018] show that if $\tilde{X} = X + \epsilon$, where $\epsilon \in \mathbb{R}^{m \times n}$ has i.i.d.

$\mathcal{N}(0, \lambda/n)$ entries, and we want to minimize the reconstruction $\mathcal{L}_{\text{noise}} = \left\| X - W_2 W_1 \tilde{X} \right\|_F^2$, then:

$$\mathbb{E}\left[\mathcal{L}_{\text{noise}}\right] = \frac{1}{2n} \sum_{i=1}^{n} \|x_i - W_2 W_1 x_i\|^2 + \frac{\lambda}{2n} \text{tr}\left(W_2 W_1 W_1^T W_2^T\right). \tag{2}$$

This motivates the **product** loss: $\mathcal{L}_\pi(W_1, W_2) = \|X - W_2 W_1 X\|_F^2 + \lambda \|W_2 W_1\|_F^2$.

Alternatively, one may regularize an LAE by penalizing derivatives, so the model learns a function which changes only slightly when the input is perturbed. An example is a **contractive** autoencoder as considered by Rifai et al. [2011], who restrict the Jacobian using the loss $\mathcal{L}(W_1, W_2) + \frac{\lambda}{2} \|J_f(x)\|_F^2$. When the encoder and decoder are tied together by $W_1 = W_2^T$, they prove that this is equivalent to minimizing the **sum** loss:

$$\mathcal{L}_\sigma(W_1, W_2) = \|X - W_2 W_1 X\|_F^2 + \lambda \left(\|W_1\|_F^2 + \|W_2\|_F^2\right). \tag{3}$$

Restricting to 1D with $m = n = 1$, we may find the critical points (w_1, w_2) of $\mathcal{L}_\sigma$ by differentiating $(x - w_2 w_1 x)^2 + \lambda(w_1^2 + w_2^2)$ with respect to w_1 and w_2, then setting these expressions to zero. The origin is always a critical point, whereas the other two points are at the intersection of the hyperbola $w_1 w_2 = 1 - \lambda x^{-2}$ and the line $w_1 = w_2$. We can therefore infer three important observations:

1. There is symmetry in w_1 and w_2 for all critical points of $\mathcal{L}_\sigma$.
2. $L_\sigma(W_1, W_2)$ is no longer invariant under the action of $\text{GL}_k(\mathbb{R})$ on (W_1, W_2).
3. There is loss of information associated with "over-regularizing" - if λ is too large (in this case $\lambda > x^2$), the origin is the only remaining critical point.

The first observation was extended to any dimension by Kunin et al. [2019], using only linear algebra:

Theorem 2.1 (Transpose Theorem)**.** *Any critical point of $\mathcal{L}_\sigma$ satisfies $W_1 = W_2^T$.*

Sketch Proof. Let $A = W_1 - W_2^T$, $B = (I - W_2 W_1) X X^T$. As gradients vanish at critical points:

$$0 = \frac{\partial \mathcal{L}_\sigma}{\partial W_1} - \left(\frac{\partial \mathcal{L}_\sigma}{\partial W_2}\right)^T = 2A(B + \lambda I). \tag{4}$$

B can be shown to be a positive semi-definite matrix (see Kunin et al. [2019] for details), so as $\lambda > 0$, $B + \lambda I$ must be positive definite. However, then $A(B + \lambda I)A^T = 0$, which implies $A = 0$. □

The second observation hints that the invariance might have been reduced to orthogonal matrices $O_k(\mathbb{R})$ (and not $\text{GL}_k(\mathbb{R})$), whereas the third observation flags a tradeoff between regularizing and the loss landscape of $\mathcal{L}_\sigma$. These notions can be formulated rigorously by turning to algebraic topology.

3 Topology of Landscapes

Instead of approaching the problem via linear algebra, one can think of minimizing the losses over the space of k-dimensional planes in $\mathbb{R}^m$ passing through the origin. In literature, this space is known as the **Grassmannian** $\text{Gr}_k(\mathbb{R}^m)$, and is a $k(m-k)$ dimensional smooth, compact manifold[1]. If we imagine our dataset X as a **point cloud** in $\mathbb{R}^m$, it naturally induces a function $\mathcal{L}_X : \text{Gr}_k(\mathbb{R}^m) \to \mathbb{R}$, mapping a k-dimensional plane to the Euclidean distance between this plane and the point cloud X.

This approach lies at the heart of *Morse theory*, an area of topology which studies smooth, compact manifolds M through smooth functions $f : M \to \mathbb{R}$ (Milnor et al. [1969]). Using commutative diagrams, one can show that $\mathcal{L}_X$ is a smooth function on $\text{Gr}_k(\mathbb{R}^m)$ with all critical points non-degenerate, if and only if the singular values of X are positive and distinct (say $\sigma_1 > ... > \sigma_m > 0$). *Under this assumption, for any fixed $k \leq m$, there are exactly $\binom{m}{k}$ critical points of $\mathcal{L}_X$, which translated back to Euclidean space become rank-k principal subspaces* (Kunin et al. [2019]).

The above topological intuition led Kunin et al. [2019] to a result for each of $\mathcal{L}, \mathcal{L}_\pi$, and $\mathcal{L}_\sigma$. To accomplish this, they preserved the above assumption on the ordering of the singular values of X. Before we state their crucial theorem, we must first introduce and motivate appropriate notation.

[1] Roughly, this means that $\text{Gr}_k(\mathbb{R}^m)$ is bounded and locally well-behaved enough for one to do calculus.

Table 1: Relationship between losses and critical points under the diffeomorphisms in Theorem 3.1.

	W_2	W_1
$\mathcal{L}$ (unregularized loss)	$U_\mathcal{I} G^+$	$G U_\mathcal{I}^T$
$\mathcal{L}_\pi$ (product loss)	$U_\mathcal{I}(I_l + \lambda \Sigma_\mathcal{I}^{-2})^{-\frac{1}{2}} G^+$	$G(I_l + \lambda \Sigma_\mathcal{I}^{-2})^{-\frac{1}{2}} U_\mathcal{I}^T$
$\mathcal{L}_\sigma$ (sum loss)	$U_\mathcal{I}(I_l - \lambda \Sigma_\mathcal{I}^{-2})^{\frac{1}{2}} O^T$	$O(I_l - \lambda \Sigma_\mathcal{I}^{-2})^{\frac{1}{2}} U_\mathcal{I}^T$

- Let $\mathcal{I} \subset \{1, ..., m\}$ be an indexing set, with $|\mathcal{I}| = l \leq k$ and increasing indices $i_1 < ... < i_l$. This will capture all combinations of at most k singular values and their principal directions.
- We write u_j for the principal direction associated with σ_j, and $U_\mathcal{I}$ for the $k \times l$ matrix of columns $u_{i_1}, ..., u_{i_l}$ in that order. The singular value matrix is then $\Sigma_\mathcal{I} = \text{diag}(\sigma_{i_1}, ..., \sigma_{i_l})$.
- Let $F_\mathcal{I}$ and $V_\mathcal{I}$ be the submanifolds of $\mathbb{R}^{k \times l}$ whose points are $\mathbb{R}^{k \times l}$ matrices with **independent** and **orthonormal** columns, respectively. These manifolds will capture the symmetries that define the subspaces of critical points. Then, the following result is true:

Theorem 3.1 (Landscape Theorem). *The critical points (W_1, W_2) of each of $\mathcal{L}, \mathcal{L}_\sigma, \mathcal{L}_\pi$ form a smooth submanifold of $\mathbb{R}^{k \times m} \times \mathbb{R}^{m \times k}$. Moreover, we have the following diffeomorphisms*[2]*:*

1. *For $\mathcal{L}$ and $\mathcal{L}_\pi$, this submanifold is diffeomorphic to the disjoint union $\bigsqcup_{\mathcal{I}:|\mathcal{I}| \leq k} F_\mathcal{I}$.*
2. *For $\mathcal{L}_\sigma$, the submanifold is diffeomorphic to $\bigsqcup_{\mathcal{I}:|\mathcal{I}| \leq k} V_\mathcal{I}$, with the added restriction that $\mathcal{I} \subset \{1, ..., m_0\}$, where $m_0 \leq m$ is the largest index for which $\sigma_{m_0}^2 > \lambda$.*

The diffeomorphisms map matrices $G \in F_\mathcal{I}$ or $O \in V_\mathcal{I}$ to critical points (W_1, W_2) as in Table 1. ***Locally, each of these points represents a saddle, with*** *$d_\mathcal{I} + (k - l)(m - l)$* ***descending directions****, where $d_\mathcal{I} = \sum_{j=1}^{k} (i_j - j)$ counts the number of pairs $i < j$ with $i \in \mathcal{I}$ and $j \notin \mathcal{I}$.*

This abstract result explicitly gives all critical points, only defined up to symmetries of matrices with independent/orthonormal columns which form manifolds. The loss landscape is then fully characterised as a disjoint union of these manifolds, each at some height. While the authors do not discuss this, the Landscape Theorem implies that for a fixed $k < \min(m, n)$, there is a **unique global minimum**. Namely, for $l < k$, each point in Table 1 has multiple descending directions given by the counting formula, and is a true saddle. Meanwhile, when $l = k$, we have $d_\mathcal{I} = 0$ if and only if $i_j = j$, for all $j \in \{1, ..., k\}$. This is the point with no descending directions and all top k singular vectors represented in the matrix $U_\mathcal{I}$, and suggests a direct relationship with PCA explored in Section 3.1.

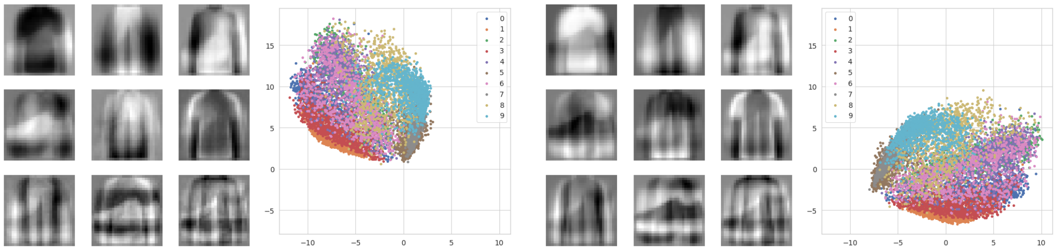
The Landscape Theorem also quantifies the "loss of information" for $\mathcal{L}_\sigma$ we observed in Section 2: as λ increases past the singular values, the submanifold of the critical points admits fewer elements, since the critical index m_0 decreases. If λ is too large, the minus sign in the expressions for $\mathcal{L}_\sigma$ in Table 1 dominates, so there may be no real solutions to $(I_l - \lambda \Sigma_\mathcal{I}^{-2})^{\frac{1}{2}}$ for certain index sets $\mathcal{I}$ and critical points will be lost. We explore this new insight computationally in full details in Section 4.2.

3.1 Relation to PCA

One further observation from Theorem 3.1 and Table 1 is that for the sum loss, $W_1^T = W_2$ is the matrix of principal directions $(U_\mathcal{I})$, compressed by $(I_l - \lambda \Sigma_\mathcal{I}^{-2})^{\frac{1}{2}}$, and rotated/reflected by $O \in F_\mathcal{I}$. This is reminiscent of classical PCA, with only the compression appearing to be novel. However, if we remove regularization ($\lambda = 0$) to eliminate compression, we lose orthogonality and end up with $W_2 = U_\mathcal{I} G^+$, where $G \in F_\mathcal{I}$ as indicated by Table 1. Therefore, regularization is essential to the link with principal components, and the true connection turns out to be with **probabilistic PCA**.

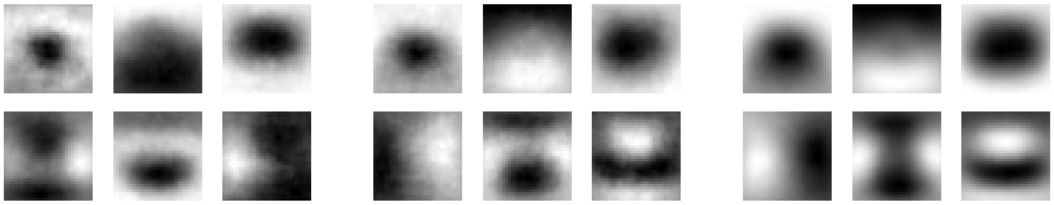
Probabilistic PCA (pPCA) can be viewed as a Gaussian framework, in which the latent vectors $z_i \sim \mathcal{N}_k(0, 1)$ reconstruct the original data points $x_i \in \mathbb{R}^m$ via the matrix W_{PCA} and some noise $\epsilon_i \sim \mathcal{N}_m(0, \tau^2)$. This is written simply as $x_i = W_{PCA} z_i + \epsilon_i$. By considering the log likelihood of this model, Tipping and Bishop [1999] show that the MLE is given by $W_{PCA} =$

[2] A diffeomorphism $F : M \to N$ of manifolds is a differentiable bijection whose inverse is differentiable.



a) Sum LAE ($k = 9$) b) Sum LAE ($k = 2$) c) PCA ($k = 9$) d) PCA ($k = 2$)

Figure 1: The sum-regularized LAE shows excellent agreement with PCA for Fashion-MNIST.



a) Sum LAE ($\lambda = 5$) b) Sum LAE ($\lambda = 15$) c) PCA

Figure 2: Higher levels of λ in $\mathcal{L}_\sigma$ better approximate PCA for the CIFAR-10 dataset.

$U_{\mathcal{I}}(\Sigma_{\mathcal{I}}^2 - \tau^2 I_l)^{\frac{1}{2}} O^T = U_{\mathcal{I}}\Sigma_{\mathcal{I}}(I_l - \tau^2 \Sigma_{\mathcal{I}}^{-2})^{\frac{1}{2}} O^T$ for $O \in V_{\mathcal{I}}$. Therefore, for $\tau = \sqrt{\lambda}$, *pPCA is equivalent to the sum-regularized LAE*, up to compression of the principal vectors by $\Sigma_{\mathcal{I}}$.

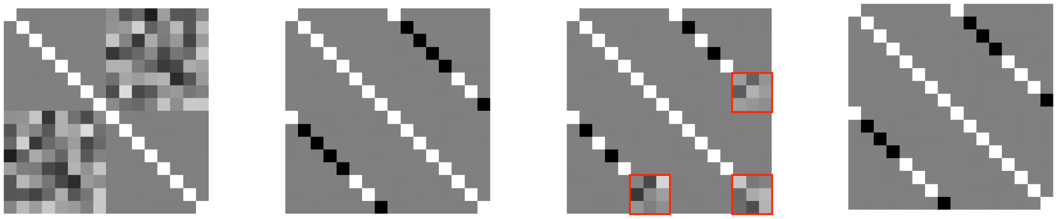
4 Numerical Simulations

We now illustrate the main results, building on code from the GitHub repository of Kunin et al. [2019], who provide functions to regularize and train LAEs using Keras and Tensorflow. The experiments were done on two new datasets from literature, as well as on original generated data. All simulations use the Adam optimizer (Kingma and Ba [2017]), learning rate 0.05 and random normal initialization.

4.1 Relationship with PCA: Fashion MNIST and CIFAR-10 datasets

We first consider the Fashion-MNIST testing dataset (Xiao et al. [2017]) of $10,000$ 28×28 greyscale images, each representing one of ten clothing items. We trained a sum-regularized LAE, with $\lambda = 8$ and $k = 9$, for 200 epochs with a batch size of 32. We first converted the images into 784×1 vectors in order to pass them as inputs. In Figure 1, we plot the left $k = 9$ singular vectors of the trained decoder W_2, converted back into 28×28 images. As guaranteed by the Landscape Theorem and the discussion in Section 3.1, the agreement between the LAE with sum loss and PCA is almost perfect: in a) and c) of Figure 1, we see that the principal directions match up to sign (inversion of black and white). Moreover, if we visualize the embeddings for $k = 2$ in Figure 1 b) and d), we witness virtually identical classification patterns for the two methods (up to rotation and reflection).

We may expand on the above discussion by considering the more complex CIFAR-10 testing dataset (Krizhevsky [2009]), which consists of $10,000$ 32×32 color images. For simplicity, using the CCIR 601 Luma weighted average (Fischer [2010]), we can convert these images into greyscale to apply the above methodology. Increasing the number of epochs to 400 when training, we plot the $k = 6$ singular vectors for $\lambda = 5$ and $\lambda = 15$ in Figure 2. In our simulations, a low level of regularization failed to learn the later singular vectors when compared to PCA, seen by the disagreement of the lower rows of Figure 2 a) versus c). Meanwhile, a loss with large λ converges more clearly to the PCA optimum. Adaptive optimizers such as Adam have been shown to escape saddles effectively (Staib et al. [2019]), so it is unlikely the problem is due to trouble with convergence. Instead, a more plausible explanation comes from the probabilistic PCA model as discussed in Section 3.1, where higher λ better captures the intrinsic noise and variation of the more sophisticated CIFAR-10 dataset.



a) Unregularized b) Product ($\mathcal{L}_\pi$) c) Sum ($\mathcal{L}_\sigma$) d) Synthetic ($\tilde{\mathcal{L}}$)

Figure 3: Heat map of $\begin{bmatrix} U & V_\star \end{bmatrix}^T \begin{bmatrix} U & U_\star \end{bmatrix}$. Black entries represent -1, white entries are 1.

.

4.2 Landscape Theorem: Artificial data and synthetic loss

While the above discussion illustrates the merits of regularization, we show that care must be taken, as we engineer an original dataset for which high λ fails to give the minimum of the Landscape Theorem. To this end, we take $m = 20, n = 30$, and we generate $X = U\Sigma V^T$ via an SVD. Here, we choose U to be an orthogonal matrix obtained via the QR factorisation of a matrix with independent $\mathcal{N}(0,1)$ entries, and obtain V similarly. Meanwhile, we set the singular values to be $\Sigma_{ii} = (21 - i)/5$. For $\lambda = 10$ and $k = 8$, this means that λ is greater than the squares of all but the largest 5 singular values, so the critical index for $\mathcal{L}_\sigma$ from the Landscape Theorem is $m_0 = 5$. Therefore, the manifold for the sum loss cannot admit a rank-k minimum, as any critical point has at least $(k - m_0)(m - m_0) > 0$ descending directions. To illustrate this, we train the LAEs with full batch for 4000 epochs.

To visualize the results, we write $W_\star = W_2W_1 = U_\star\Sigma_\star V_\star^T$ for the trained autoencoders, and compare them to $X = U\Sigma V^T$ in Figure 3 via the following matrix: $\begin{bmatrix} U^TU & U^TU_\star \\ V_\star^TU & V_\star^TU_\star \end{bmatrix}$. This is plotted in a heat map for $\mathcal{L}$, $\mathcal{L}_\pi$, and $\mathcal{L}_\sigma$ in Figure 3, a visualization method similarly used by Plaut [2018] and Kunin et al. [2019]. Clearly, the top left quadrant is the identity for all losses since U is orthogonal. Moreover, as the Landscape Theorem guarantees $W_\star$ approximates a minimum for the product and unregularized losses regardless of λ, we see that $U_\star = V_\star$ in the lower right quadrant as expected from the expressions in Table 1. The product loss also obtains the left singular vectors of X as the left and right singular vectors of $W_\star$. Meanwhile, the sum loss fails to achieve this: as $k - m_0 = 3$, 3×3 red blocks in the heat map correspond to noise, consistent with the lack of a rank-k minimum.

We may resolve this issue by defining a **synthetic** loss $\tilde{\mathcal{L}} = w_\sigma\mathcal{L}_\sigma + w_\pi\mathcal{L}_\pi$, for non-negative weights such that $w_\sigma + w_\pi = 1$. While a reduction of λ for $\mathcal{L}_\sigma$ itself would restore the landscape lost, the idea is that this new loss gives autoencoders both denoising and contractive properties for an appropriate balance of weights. We see in Figure 3 d) that for $w_\sigma = w_\pi = 0.5$, the synthetic loss admits the properties of the product loss in terms of convergence to a minimum. Additionally, computational evidence on this data in Figure 4 shows that the Transpose Theorem still holds for this synthetic loss (even when $w_\sigma = 0.05$, although it fails for $w_\sigma = 0.005$), inheriting this symmetry from the sum loss. This is somewhat surprising, as the proof presented in Section 2 for $\mathcal{L}_\sigma$ does not extend directly. Regardless, provided there is sufficient sum regularization, we expect symmetry within matrices with orthonormal (and not just independent) columns. Then, heuristically the matrices $G \in F_\mathcal{I}$ in Table 1 become $O \in V_\mathcal{I}$, so a transpose relationship should hold. We conjecture that under some conditions on w_σ and λ, one might establish a precise transpose result. This could have major implications, as the Transpose Theorem is a key ingredient in Kunin's proof of the Landscape Theorem, and so the critical landscape of this synthetic loss might also be exactly classified topologically.

5 Conclusion

In this report, we presented a recent topological breakthrough on the loss landsapes of regularized autoencoders. We explored the effect of sum regularization for two novel datasets, and constructed an original scenario which illustrates the boundaries of the scope of the theorems. The promising properties of the synthetic loss may open further research avenues, which nonetheless remain ample with questions on the extensions of the results to non-linear autoencoders and general loss functions.

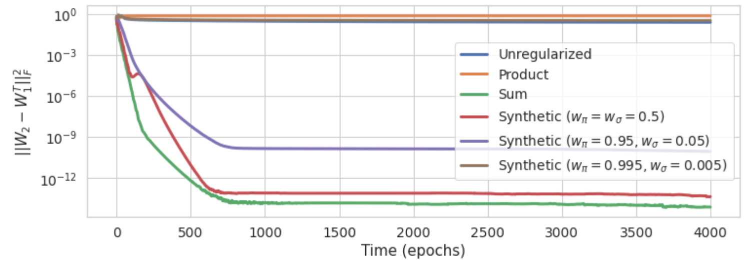


Figure 4: Tracking the squared Frobenius norm between W_1 and W_2^T for $\mathcal{L}_\sigma$ and $\tilde{\mathcal{L}}$.

References

C. Bishop. *Pattern Recognition in Machine Learning*. Springer, 2006.

H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59:291–294, 1988.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

W. Fischer. *Digital Video and Audio Broadcasting Technology*. Springer, 2010.

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

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

D. Kunin, J. Bloom, A. Goeva, and C. Seed. Loss landscapes of regularized linear autoencoders. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 3560–3569. PMLR, 2019. URL http://proceedings.mlr.press/v97/kunin19a.html.

J. Milnor, M. Spivak, and R. Wells. *Morse Theory. (AM-51), Volume 51*. Princeton University Press, 1969. ISBN 9780691080086. URL http://www.jstor.org/stable/j.ctv3f8rb6.

E. Plaut. From principal subspaces to principal components with linear autoencoders, 2018. URL https://arxiv.org/abs/1804.10253.

A. Pretorius, S. Kroon, and H. Kamper. Learning dynamics of linear denoising autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4141–4150. PMLR, 2018.

S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 833–840. ACM, 2011.

M. Staib, S. J. Reddi, S. Kale, S. Kumar, and S. Sra. Escaping saddle points with adaptive gradient methods. *CoRR*, abs/1901.09149, 2019. URL http://arxiv.org/abs/1901.09149.

M. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, 61(3):611–622, 1999.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.