## A. Deferred Proofs

***Proof of Proposition 4.3.*** Setting the gradient to zero:

$$\frac{\partial \mathcal{L}_*}{\partial Q_1} = 0 \implies Q_2^\intercal Q_2 Q_1 S^2 = Q_2^\intercal D^2 \tag{5}$$

$$\frac{\partial \mathcal{L}_*}{\partial Q_2} = 0 \implies Q_2 Q_1 S^2 Q_1^\intercal = D^2 Q_1^\intercal \tag{6}$$

Let $Q = Q_2 Q_1$. Multiplying (5) on the right by $D^{-2}$ and on the left by $D^{-2} S^2 Q_1^\intercal$ gives

$$D^{-2} S^2 Q^\intercal Q S^2 D^{-2} = D^{-2} S^2 Q^\intercal,$$

which implies $D^{-2} S^2 Q^\intercal$ is symmetric and idempotent. Multiplying (6) on the right by $Q_2^\intercal$ gives

$$Q S^2 Q^\intercal = D^2 Q^\intercal,$$

which can be rewritten as

$$Q S^2 Q^\intercal = \left(D^2 S^{-2} D^2\right)\left(D^{-2} S^2 Q^\intercal\right).$$

Since the left-hand side is symmetric, $D^{-2} S^2 Q^\intercal$ is diagonal and idempotent by Lemma A.2 with $A = D^{-2} S^2 Q^\intercal$ and $B = D^2 S^{-2} D^2$. Lemma A.3 with the same $A$ implies there exists an index set $\mathcal{I}$ of size $\ell$ with $0 \le \ell \le k$ such that

$$D^{-2} S^2 Q^\intercal = I_\mathcal{I} I_\mathcal{I}^\intercal$$

and hence

$$Q = I_\mathcal{I} I_\mathcal{I}^\intercal D^2 S^{-2} = I_\mathcal{I} D_\mathcal{I}^2 S_\mathcal{I}^{-2} I_\mathcal{I}^\intercal. \tag{7}$$

Consider the smooth map $(Q_1, Q_2) \mapsto (\mathcal{I}, G)$ with

$$G = Q_1 S_\mathcal{I} D_\mathcal{I}^{-1} I_\mathcal{I}$$

from the critical submanifold of $\mathbb{R}^{k\times m} \times \mathbb{R}^{m\times k}$ to the manifold of pairs $(\mathcal{I}, G)$ with $G$ full-rank. Note

$$G^+ = I_\mathcal{I}^\intercal S_\mathcal{I} D_\mathcal{I}^{-1} Q_2$$

by (7). Commuting diagonal matrices to rearrange terms in (5) and (6), we obtain a smooth inverse map from pairs $(\mathcal{I}, G)$ to critical points:

$$Q_1 = Q_1 S^2 Q^\intercal D^{-2} = Q_1 I_\mathcal{I} I_\mathcal{I}^\intercal = G I_\mathcal{I}^\intercal D_\mathcal{I} S_\mathcal{I}^{-1},$$
$$Q_2 = D^{-2} S^2 Q^\intercal Q_2 = I_\mathcal{I} I_\mathcal{I}^\intercal Q_2 = D_\mathcal{I} S_\mathcal{I}^{-1} I_\mathcal{I} G^+.$$

□

***Equating bias parameters and mean centering.*** Consider the loss function

$$\mathcal{L}_\beta(W_1, W_2, b_1, b_2) = ||X - W_2(W_1 X + b_1 e_n^\intercal) + b_2 e_n^\intercal||_F^2,$$

where $b_1 \in \mathbb{R}^k$ and $b_2 \in \mathbb{R}^m$ are bias vectors and $e_n \in \mathbb{R}^n$ is the vector of ones. With $b = W_2 b_1 + b_2$, $\mathcal{L}_b$ becomes

$$||X - W_2 W_1 X - b e_n^\intercal||_F^2. \tag{8}$$

At a critical point,

$$\frac{\partial \mathcal{L}_\beta}{\partial b} = 2(X - W_2 W_1 X - b e_n^\intercal) e_n = 0,$$

which implies

$$b = \frac{1}{n} X e_n - W_2 W_1 \frac{1}{n} X e_n.$$

Substituting into (8), $\mathcal{L}_b$ reduces to

$$||\bar{X} - W_2 W_1 \bar{X}||_F^2$$

with $\bar{X} = X - \frac{1}{n} X e_n e_n^\intercal$. Thus, at the optimal bias parameters, $\mathcal{L}_\beta$ with $X$ is equivalent to $\mathcal{L}$ with $X$ mean-centered. □

**Lemma A.1.** *Let $A \in \mathbb{R}^{m\times k}$ and $B \in \mathbb{R}^{k\times m}$, then*

$$||A||_F^2 + ||B||_F^2 = ||A - B^\intercal||_F^2 + 2tr(AB)$$

*Proof.*

$$\begin{aligned} ||A - B^\intercal||_F^2 &= \mathrm{tr}((A - B^\intercal)^\intercal (A - B^\intercal)) \\ &= \mathrm{tr}(A^\intercal A - A^\intercal B^\intercal - BA + B^\intercal B) \\ &= ||A||_F^2 + ||B||_F^2 - 2tr(AB) \end{aligned}$$

□

**Lemma A.2.** *Let $A, B \in \mathbb{R}^{m\times m}$ with $B$ diagonal with distinct diagonal elements. If $AB = BA$ then $A$ is diagonal.*

*Proof.* Expand the difference of $(i, j)$ elements:

$$\begin{aligned} (AB)_{ij} - (BA)_{ij} &= a_{ij} b_{jj} - b_{ii} a_{ij} \\ &= a_{ij}(b_{jj} - b_{ii}) = 0. \end{aligned}$$

So for $i \neq j$, $b_{ii} \neq b_{jj}$ implies $a_{ij} = 0$. □

**Lemma A.3.** *If $A \in \mathbb{R}^{m\times m}$ is diagonal and idempotent then $a_{ii} \in \{0, 1\}$.*

*Proof.* $0 = (AA - A)_{ii} = a_{ii}^2 - a_{ii} = a_{ii}(a_{ii} - 1)$. □

***Relationship between Oja's rule and LAE-PCA.*** The update step used in Oja's rule is

$$\nabla w = \alpha(xy - wy^2),$$

where $\alpha$ is a fixed learning rate, $x, w \in \mathbb{R}^m$ and $y = x^\intercal w$. Substituting $y$ into this update and factoring out $xx^\intercal w$ on the right gives,

$$\nabla w = \alpha(1 - ww^\intercal) xx^\intercal w,$$

which is the (negative) gradient for an unregularized LAE with tied weights in the $k = 1$ case. □

## B. Positive (semi-)definite matrices

We review positive definite and semi-definite matrices as needed to prove the Transpose Theorem (2.1).

**Definition B.1.** *A real, symmetric matrix $A$ is positive semi-definite, denoted $A \succeq 0$, if $x^\intercal Ax \geq 0$ for all vectors $x$. $A$ is positive definite, denoted $A \succ 0$, if the inequality is strict.*

*The Loewner partial ordering of positive semi-definite matrices defines $A \succeq B$ if $A - B \succeq 0$.*

**Lemma B.1.** *The following properties hold.*

1. *If $\lambda > 0$ then $\lambda I \succeq 0$.*
2. *If $A \succeq 0$ then $BAB^\intercal \succeq 0$ for all $B$.*
3. *If $A \succeq 0$ and $B \succeq 0$ then $A + B \succeq 0$.*
4. *If $A \succ 0$ and $B \succeq 0$ then $A + B \succ 0$.*
5. *If $A \succeq 0$ and $AB^\intercal \succeq BAB^\intercal$ then $A \succeq BA$.*
6. *If $B \succ 0$ and $A^\intercal BA = 0$ then $A = 0$.*

*Proof.* Property 5 follows from Properties 1 and 2 and

$$A - BA = (B - I)A(B - I)^\intercal + (AB^\intercal - BAB^\intercal).$$

The other properties are standard exercises; see Appendix C of (van den Bos, 2007) for a full treatment. □

## C. Denoising and contractive autoencoders

Here we connect regularized LAEs to the linear case of denoising (DAE) and contrastive (CAE) autoencoders.

A linear DAE receives a corrupted data matrix $\tilde{X}$ and is trained to reconstruct $X$ by minimizing

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

As shown in Pretorius et al. (2018), if $\tilde{X} = X + \epsilon$ is the corrupting process, where $\epsilon \in \mathbb{R}^{m \times n}$ is a noise matrix with elements sampled iid from a distribution with mean zero and variance $s^2$, then

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

With $\lambda = ns^2$, we have

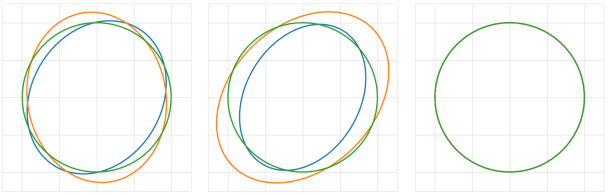
$$\mathbb{E}\left[\mathcal{L}_{\text{DAE}}\right] = \frac{1}{2n} \mathcal{L}_\pi.$$

The loss function of a linear CAE includes a penalty on the derivative of the encoder:

$$\mathcal{L}_{\text{CAE}}(W_1, W_2) = \mathcal{L}(W_1, W_2) + \gamma ||J_f(x)||_F^2.$$

As shown in Rifai et al. (2011), if the encoder and decoder are tied by requiring $W_1 = W_2^\intercal$, then $\mathcal{L}_{\text{CAE}}$ equals $\mathcal{L}_\sigma$ with $\lambda = \frac{\gamma}{2}$:

$$\mathcal{L}_{\text{CAE}}(W_1) = \mathcal{L}_\sigma(W_1, W_1^\intercal).$$



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

*Figure 6.* Image of the unit circle (green) under $A$ (blue), $B$ (orange), and $AB$ (green) from (9). Non-orthogonal transformations deform the circle to an ellipse; orthogonal transformations preserve the circle.

## D. Further empirical exploration

The Landscape Theorem also gives explicit forms for the trained encoder $W_{1*}$ and decoder $W_{2*}$ such that the matrices

$$A = \Sigma_*^{-\frac{1}{2}} U^\intercal W_{2*} \quad \text{and} \quad B = W_{1*} U \Sigma_*^{-\frac{1}{2}} \tag{9}$$

satisfy $AB = I_k$ for all losses and are each orthogonal for the sum loss. In Figure 6, we illustrate these properties by applying the linear transformations $A$, $B$, and $AB$ to the unit circle $\mathbb{S}^1 \subset \mathbb{R}^2$. Non-orthogonal transformations deform the circle to an ellipse, whereas orthogonal transformations (including the identity) preserve the unit circle. This experiment used the same setup described in 5.1 with $k = 2$.

### D.1. MNIST

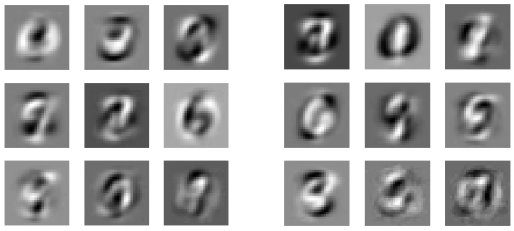
In the following experiment, the data set $X \in \mathbb{R}^{784 \times 10000}$ is the test set of the MNIST handwritten digit database (LeCun & Cortes). We train an LAE with $k = 9$ and $\lambda = 10$ for each loss, again using the Adam optimizer for 100 epochs with random normal initialization, batch size of 32, and learning rate 0.05.

Figure 7 further illustrates the Landscape Theorem 4.2 by reshaping the left singular vectors of the trained decoder $W_{2*}$ and the top $k$ principal direction of $X$ into $28 \times 28$ greyscale images. Indeed, only the decoder from the LAE trained on the sum loss has left singular vectors that match the principal directions up to sign.

As described in Section 3, for an LAE trained on the sum loss, the latent representation is, up to orthogonal transformation, the principal component embedding compressed along each principal direction. We illustrate this in Figure 8 by comparing the $k = 2$ representation to that of PCA.

## E. Morse homology of the real Grassmannian

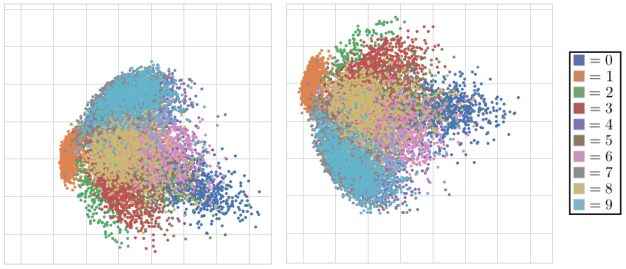
This section embraces the language and techniques of differential and algebraic topology to dive into the topology underlying LAEs. To complement Wikipedia, the following



(a) Unregularized (b) Product

(c) Sum (d) PCA

*Figure 7.* Left singular vectors of the decoder from an LAE trained on unregularized, product, and sum losses and the principal directions of MNIST reshaped into images.




(a) Sum (b) PCA

*Figure 8.* Latent representations of MNIST learned by an LAE with sum loss and by PCA. Colors represent class label.

resources cover the italicized terminology in depth: (Milnor, 1963; Hatcher, 2002; Banyaga & Hurtubise, 2004).

Let $M$ be a smooth, compact manifold. In this section, we prove the Grassmannian Theorem through the lens of *Morse theory*, a subfield of differential and algebraic topology that relates the topology of $M$ to smooth functions $f\colon M \to \mathbb{R}$.

A critical point of $f$ is *non-degenerate* if the eigenvalues of the Hessian are non-zero. A *Morse function* is a smooth function all of whose critical points are non-degenerate. Morse functions are generic and stable; see Section 4.2 of Bloom (2004) for precise statements.

The *Morse index* $d$ of a critical point is the number of negative eigenvalues of the Hessian. At each index-$d$ non-degenerate critical point, one can choose a local coordinate system under which the function takes the form

$$-x_1^2 - \ldots - x_d^2 + x_{d+1}^2 + \ldots + x_m^2.$$

Hence $d = 0$ and $d = m$ correspond to parabolic minima and maxima, respectively, which all other values of $d$ correspond to saddles with, in local coordinates, $d$ orthogonal descending directions and $m-d$ orthogonal ascending directions. For example, the red, blue, and green critical points in Figure 2 have Morse indices 0, 1, and 2, respectively.

The *Morse inequalities* state that any Morse function on $M$ must have at least as many index-$d$ critical points as the *Betti number* $b_d$, i.e. the rank of the *singular homology group* $H_d(M;\mathbb{Z})$. This follows from a realization, called *Morse homology*, of singular homology as the homology of a chain complex generated in dimension $d$ by the index-$d$ critical points. The boundary map $\partial$ counts negative gradient trajectories between critical points of adjacent index. A Morse function is perfect if this signed count is always zero, in which case $\partial$ vanishes. A Morse function is $\mathbb{F}_2$*-perfect* if this count is always even, in which case $\partial$ vanishes over the field of two elements.

Not all smooth manifolds admit perfect Morse functions. For example, the projective plane $\mathbb{RP}^2 \cong \mathrm{Gr}_1(\mathbb{R}^3)$ cannot since $H_1(\mathbb{RP}^2;\mathbb{Z}) \cong \mathbb{F}_2$ implies that $\partial$ is non-zero. The *Poincaré homology sphere* is a famous example of a manifold without a perfect Morse over $\mathbb{Z}$ or any field[18].

The Grassmannian $\mathrm{Gr}_k(\mathbb{R}^m)$ provides a coordinate-free representation of the space of rank-$k$ orthogonal projections, a submanifold of $\mathbb{R}^{m\times m}$. Through the identification of a projection with its image, $\mathrm{Gr}_k(\mathbb{R}^m)$ is endowed with the structure of a smooth, compact Riemannian manifold of dimension $k(m-k)$.

**Theorem E.1.** *$\mathcal{L}_X$ is an $\mathbb{F}_2$-perfect Morse function. Its critical points are the rank-$k$ principal subspaces.*

*Proof.* Consider the commutative diagram

$$\begin{array}{ccc} \mathrm{V}_k(\mathbb{R}^m) & \xrightarrow{\pi:O\mapsto \mathrm{Im}(OO^\intercal)} & \mathrm{Gr}_k(\mathbb{R}^m) \\ {\scriptstyle \iota:O\mapsto(O^\intercal,O)}\downarrow & & \downarrow{\scriptstyle \mathcal{L}_X} \\ \mathbb{R}^{k\times m}\times\mathbb{R}^{m\times k} & \xrightarrow[\mathcal{L}]{} & \mathbb{R} \end{array} \tag{10}$$

where $\mathrm{V}_k(\mathbb{R}^m)$ is the *Stiefel manifold* of $m \times k$ matrices with orthonormal columns. Since $\iota$ is an immersion, by Theorem 4.2 the critical points of $\mathcal{L}\circ\iota = \mathcal{L}_X \circ \pi$ are all $k$-frames spanning principal subspaces of $X$. Since $\pi$ is a submersion, the critical points of $\mathcal{L}_X$ are the image of this subset under $\pi$ as claimed.

Each critical point (that is, rank-$k$ principle subspace) is non-degenerate because each of the included $k$ principal

[18] This is because if a homology 3-sphere admits a perfect Morse function, then it consists of a 3-cell attached to a 0-cell, and is therefore the 3-sphere. Similarly, the *smooth 4-dimensional Poincaré conjecture* holds if and only if every smooth 4-sphere admits a perfect Morse function. This conjecture, whose resolution continues to drive the field, states that there is only one smooth structure on the topological 4-sphere.

directions may be rotated toward any of the excluded $m-k$ principal directions in the plane they span, fixing all other principal directions; this accounts for all $k(m-k)$ dimensions. Flowing from higher to lower eigenvalues, these rotations are precisely the $-\nabla\mathcal{L}_X$ trajectories between adjacent index critical points. Since there are exactly two directions in which to rotate, we conclude that $\mathcal{L}_X$ is $\mathbb{F}_2$-perfect. □

While this paper may be the first to directly construct an $\mathbb{F}_2$-perfect Morse function on the real Grassmannian, the existence of some $\mathbb{F}_2$-perfect Morse function is straightforward to deduce from the extensive literature on perfect Morse functions on complex Grassmanians (Hansen, 2012; Duan, 2004). Our simple and intuitive function is akin to that recently established for the special orthogonal group (Solgun, 2016).

Note that $\mathcal{L}_X$ is invariant to replacing $X = U\Sigma V^\intercal$ with $U\Sigma$ and therefore doubles by replacing $X$ with the $2m$ points bounding the axes of the principal ellipsoid of the covariance of $X$. Rotating by $U^\intercal$, one need only consider the data set

$$\{(\pm\sigma_1^2, 0, \dots, 0), (0, \pm\sigma_2^2, \dots, 0), \dots, (0, 0, \dots, \pm\sigma_m^2)\}$$

to appreciate the symmetries, dynamics, and critical values of the gradient flow in general.

We encourage the reader to check the Morse index formula (4) in the case of $\mathrm{Gr}_2(\mathbb{R}^4)$ in the table below. The symmetries of the table reflect the duality between a plane and its orthogonal complement in $\mathbb{R}^4$.

| $d$ | $u_1$ | $u_2$ | $u_3$ | $u_4$ |
|---|---|---|---|---|
| 4 | | | • | • |
| 3 | | • | | • |
| 2 | | • | • | |
| 2 | • | | | • |
| 1 | • | | • | |
| 0 | • | • | | |

Theorem 2 implies that $\mathcal{L}_X$ endows $\mathrm{Gr}_k(\mathbb{R}^m)$ with the structure of a *CW complex*; each index-$d$ critical point $P$ is the maximum of a $d$-dimensional cell consisting of all points that asymptotically flow up to $P$. This decomposition coincides with the classical, minimal CW construction of Grassmannians in terms of *Schubert cells*. Over $\mathbb{Z}$, pairs of rotations have the same sign when flowing from even to odd dimension, and opposite signs when flowing from odd to even dimension, due to the oddness and evenness of the antipodal map on the boundary sphere, respectively. In this way, Morse homology for $\mathcal{L}_X$ realizes the same chain complex as *CW homology* on the Schubert cell structure of the real Grassmannian.

### E.1. Morse homology and deep learning

We have seen how the rich topology of the real Grassmannian forces any generic smooth function to have at least $\binom{m}{k}$ critical points. More interestingly from the perspective of deep learning, Morse homology also explains why simple topology forces critical points of any generic smooth function to "geometrically cancel" through gradient trajectories. As an intuitive example, consider a generic smooth function $f : \mathbb{R} \to \mathbb{R}$ that is strictly decreasing for $x < a$ and strictly increasing for $x > b$ for some $a < b$. Then on $[a, b]$, $f$ wiggles up and down, alternating between local minima and maxima, with pairwise gradient cancellation leaving a single minimum.

More generally, for a generic smooth loss function over the connected parameter space $\mathbb{R}^p$, diverging strictly to infinity outside of a compact subset, each pair of minima is linked by a path of gradient trajectories between minima and index-1 saddles[19]. In fact, since $\mathbb{R}^p$ is contractible, we can flow upward along gradient trajectories from one minimum to all minima through index-1 saddles, from those index-1 saddles to other index-1 saddles through index-2 saddles, and so on until the resulting chain complex is contractible. Note there may exist additional critical points forming null-homotopic chain complexes.

The contractible complex containing the minima is especially interesting in light of Choromanska et al. (2015). For large non-linear networks under a simple generative model of data, the authors use random matrix theory[20] to prove that critical points are layered according to index: local minima occur at a similar height as the global minimum, index-1 saddles in a layer just above the layer of minima, and so on. Hence Morse homology provides a principled foundation for the empirical observation of low-lying valley passages between minima used in Fast Geometric Ensembling (FGE) (Garipov et al., 2018).

In FGE, after descending to one minimum, the learning rate is cycled to traverse such passages and find more minima. While the resulting ensemble prediction achieves state-of-the-art performance, these nearby minima may correspond to models with correlated error. With this in mind, we are exploring whether ensemble prediction is improved using less correlated minima, and whether many such minima may be found with logarithmic effort by recursively bifurcating gradient descent near saddles to descend alongside the Morse complex described above.

[19] For example, in Figure 1(c), the red minima are each connected to the yellow saddle by one gradient trajectory.

[20] From this perspective, the LAE is a toy model that more directly bridges loss landscapes and random matrix theory; the heights of critical points are sums of eigenvalues.