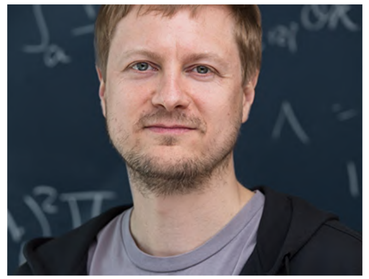
Pacific Ballroom #26

Loss Landscapes of Regularized Linear Autoencoders


Daniel Kunin


Jonathan M. Bloom


Aleksandrina Goeva


Cotton Seed

BROAD INSTITUTE

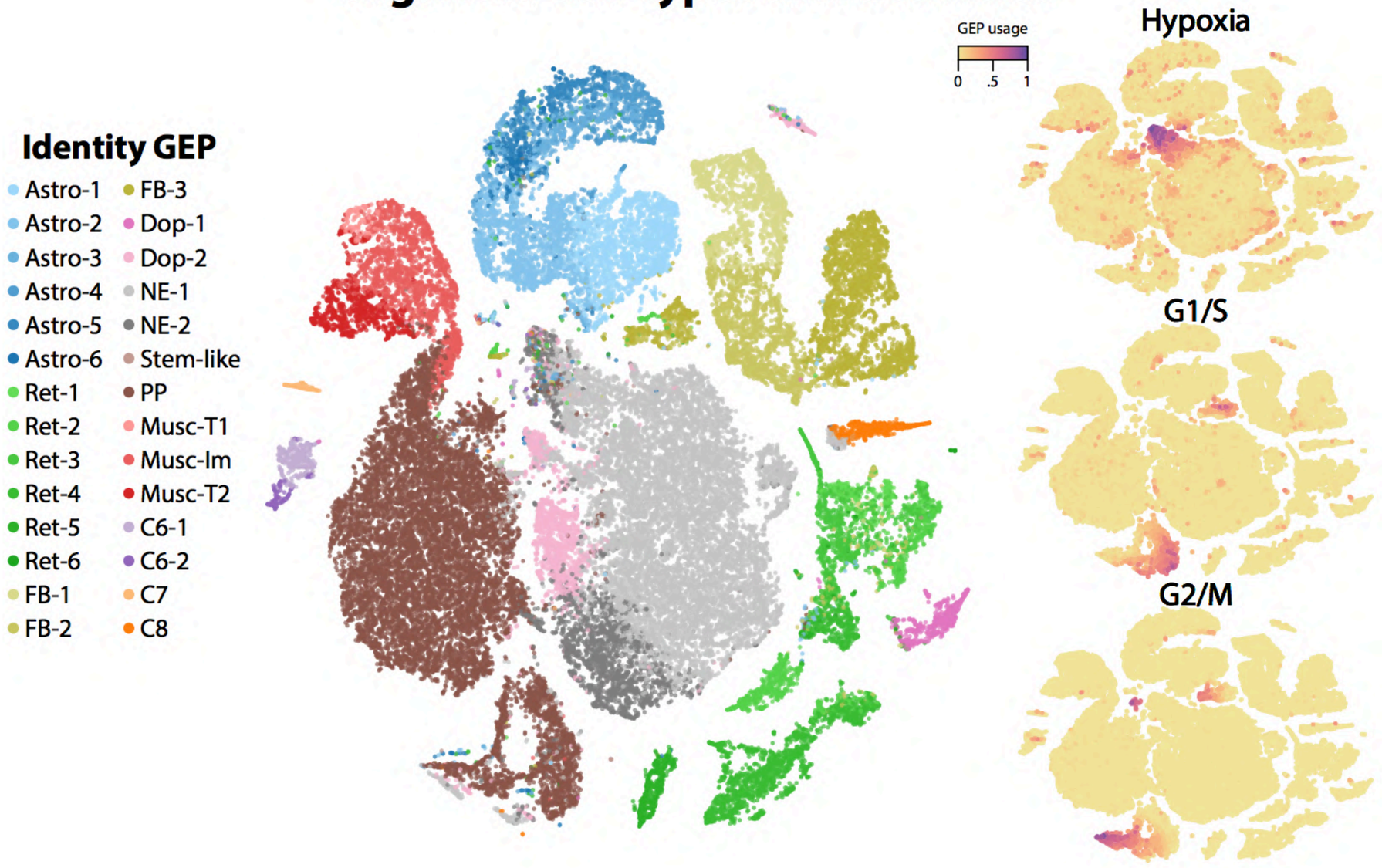
STANLEY CENTER FOR PSYCHIATRIC RESEARCH AT BROAD INSTITUTE

hail

Stanford University

Can we use **autoencoders** to learn meaningful representations of **neurons** from their gene expression?

Organoid cell-types and activities

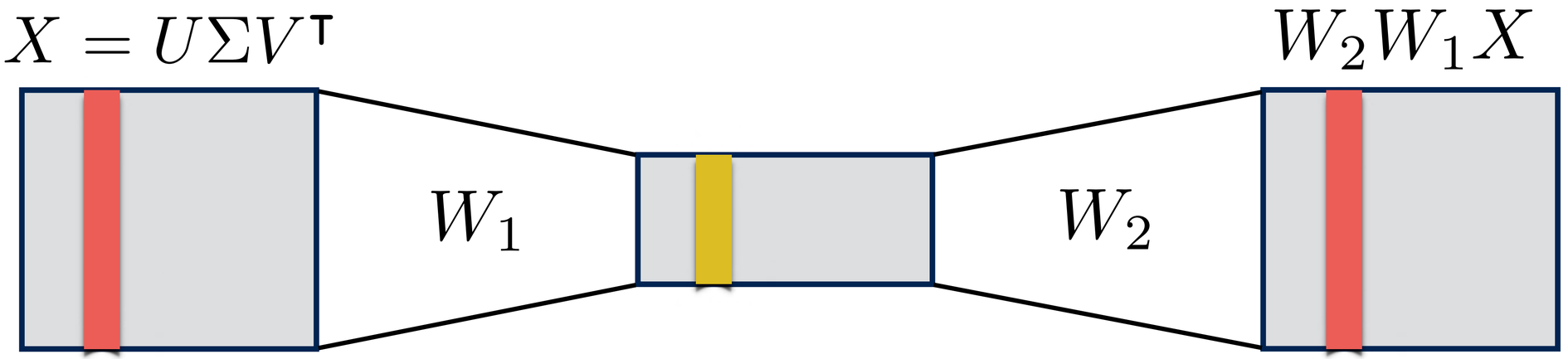


What **does** a linear autoencoder learn?

&

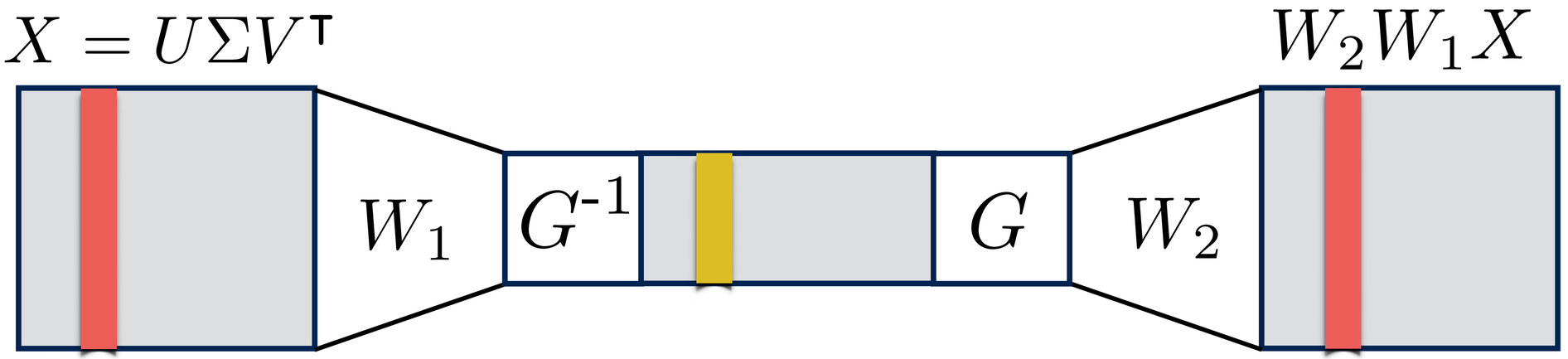
What **doesn’t** a linear autoencoder learn?

Does learn the **principal subspace**



$$\mathcal{L}(W_1, W_2) = ||X - \underbrace{W_2W_1}_{U_kU_k^\intercal}X||^2$$

Doesn't learn the **principal directions** or **eigenvalues**



$$\mathcal{L}(W_1, W_2) = ||X - \underbrace{W_2W_1}_{U_kGG^{-1}U_k^\intercal}X||^2$$

Principal Directions of X

Left Singular Vectors of W_2

Elad Plaut, 2018, From Principal Subspaces to Principal Components with Linear Autoencoders

Adding Regularization

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

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

Regularization and **Orthogonality**

1. Orthogonal matrices are the volume-preserving matrices of minimal Frobenius norm.

$$\min_A ||A||_F^2 \quad \text{s.t.} \;\; \det(A)^2 = 1$$

2. Orthogonal matrices are the inverse matrices of minimal Frobenius norm.

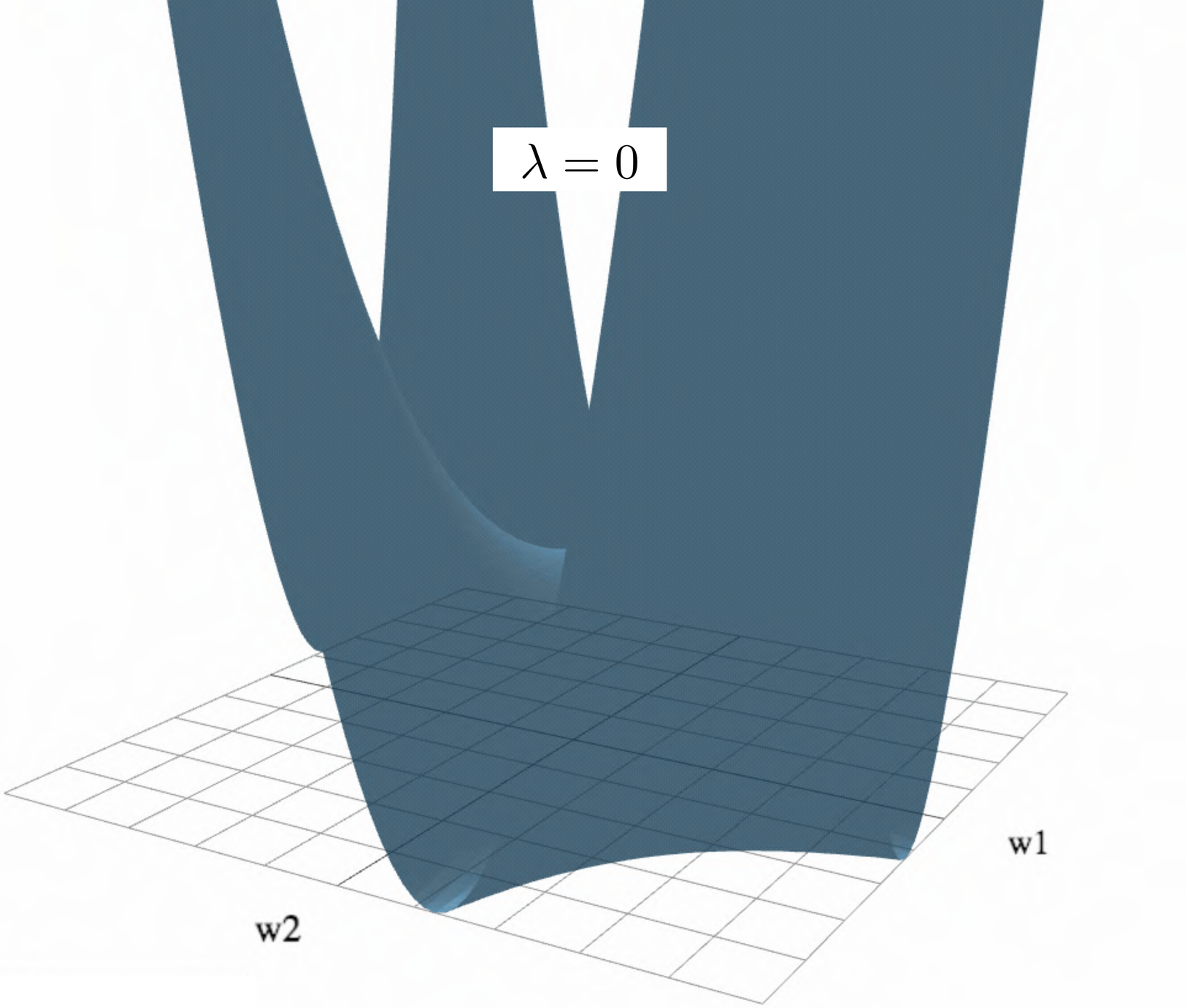
$$\min_{A,B} ||A||_F^2 + ||B||_F^2 \quad \text{s.t.} \;\; AB = I$$

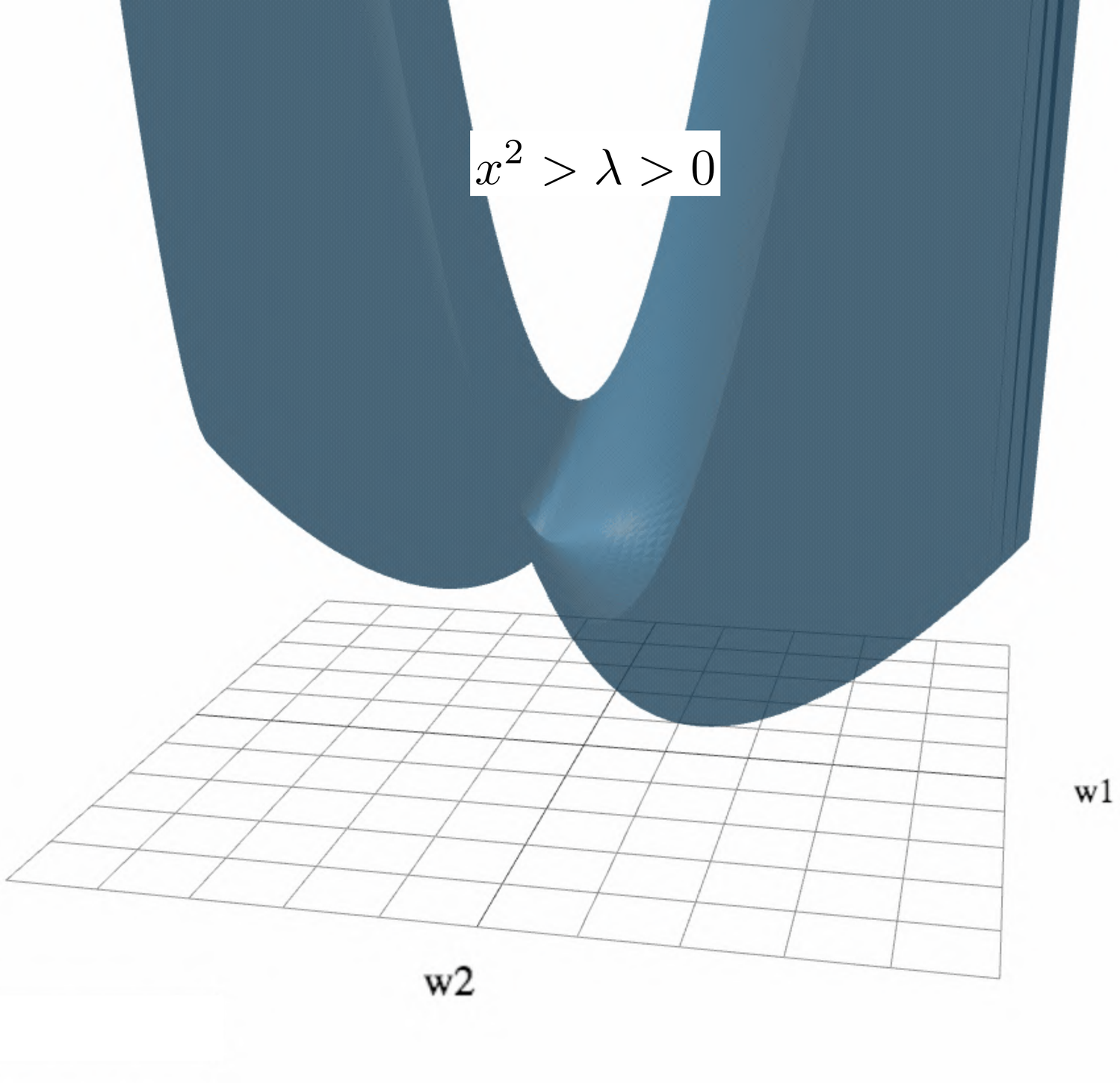
In particular $A = B^\intercal$ at all minima.

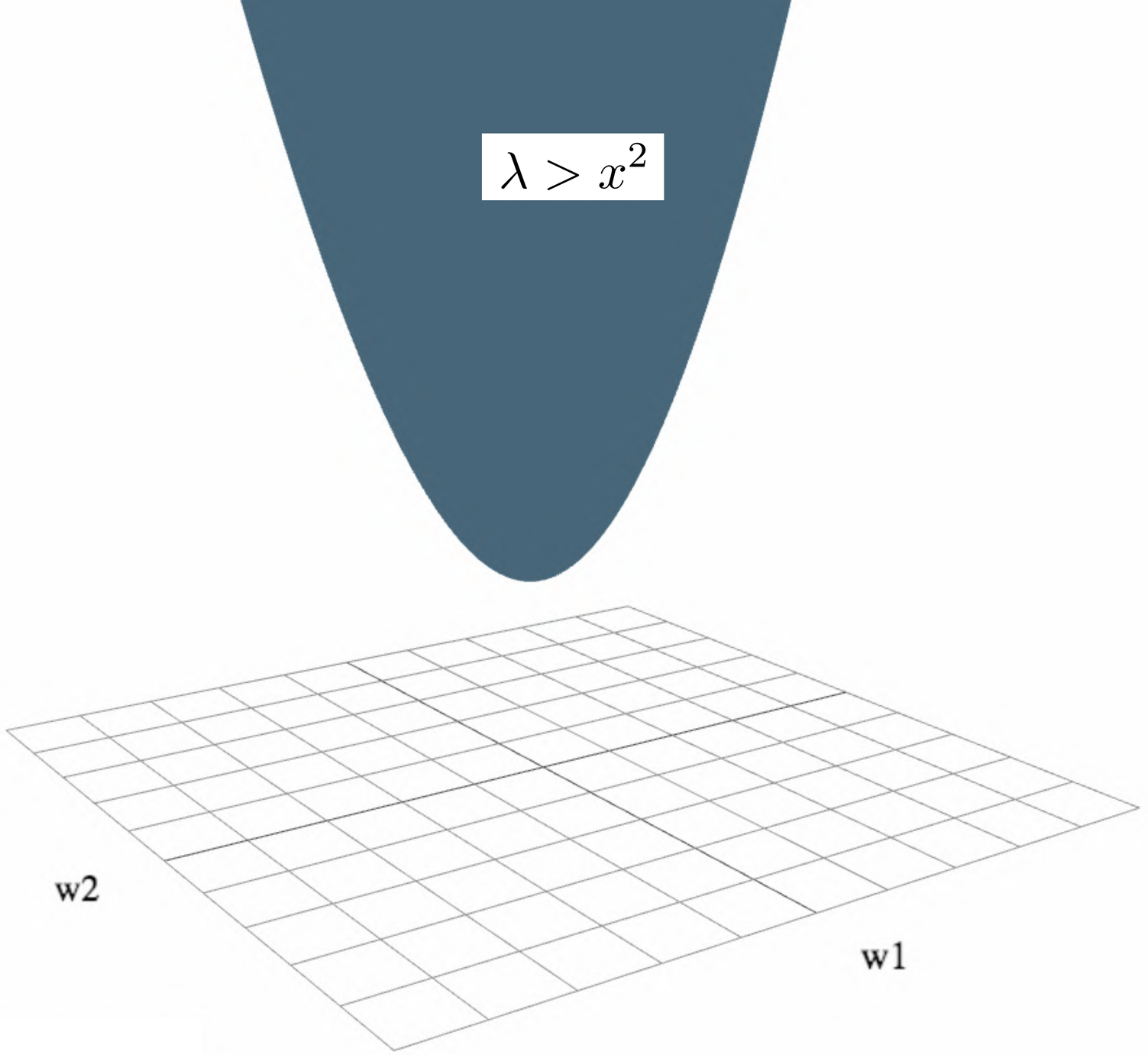
Scalar Case

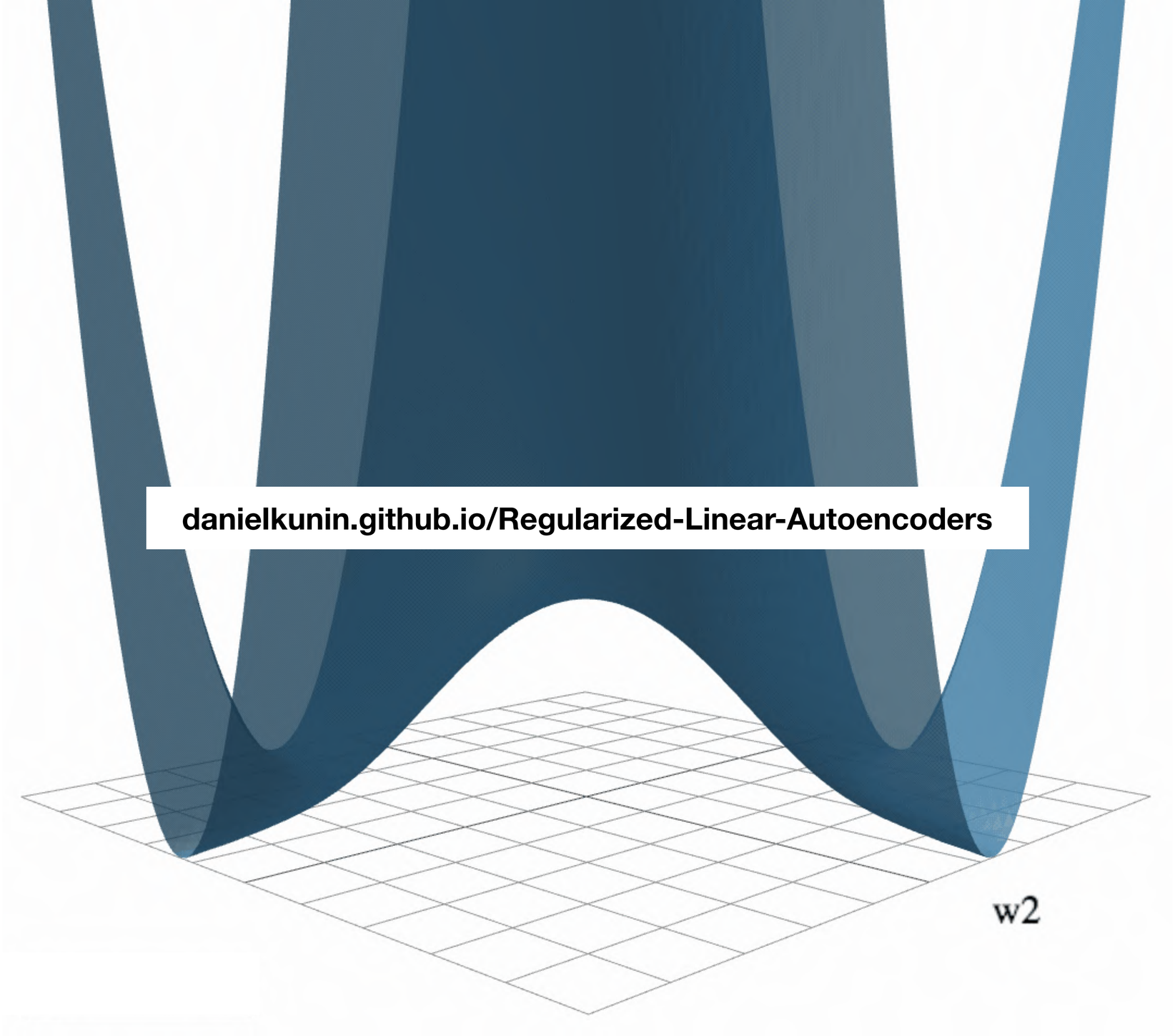
$$(x - w_2 w_1 x)^2 \xrightarrow{\text{Critical points}} \begin{cases} w_1 = w_2 = 0 \\ w_2 w_1 = 1 \end{cases}$$

$$(x - w_2 w_1 x)^2 + \lambda(w_1^2 + w_2^2) \xrightarrow{\text{Critical points}} \begin{cases} w_1 = w_2 = 0 \\ w_2 w_1 = (1 - \lambda x^{-2}) \\ w_1 = w_2 \end{cases}$$


$\lambda = 0$
w2
w1


$x^2 > \lambda > 0$
w1
w2


$\lambda > x^2$
w2
w1


danielkunin.github.io/Regularized-Linear-Autoencoders
w2

Loss Functions

Unregularized:

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

Product Regularized:

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

Sum Regularized:

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

Theorem 4.2 (Landscape Theorem).

The critical landscape is diffeomorphic to the space of pairs $(\mathcal{I}, G)$ or $(\mathcal{I}, O)$ with

- $\mathcal{I} \subset \{1, \ldots, m\}$ of size $0 \leq l \leq k$,
- $G \in \mathbb{R}^{k \times l}$ with independent columns,
- $O \in \mathbb{R}^{k \times l}$ with orthonormal columns.

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

Theorem 4.2 (Landscape Theorem).

The critical landscape is diffeomorphic to the space of pairs $(\mathcal{I}, G)$ or $(\mathcal{I}, O)$ with

- $\mathcal{I} \subset \{1, \ldots, m\}$ of size $0 \leq l \leq k$,
- $G \in \mathbb{R}^{k \times l}$ with independent columns,
- $O \in \mathbb{R}^{k \times l}$ with orthonormal columns.

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

Theorem 4.2 (Landscape Theorem).

The critical landscape is diffeomorphic to the space of pairs $(\mathcal{I}, G)$ or $(\mathcal{I}, O)$ with

- $\mathcal{I} \subset \{1, \ldots, m\}$ of size $0 \leq l \leq k$,
- $G \in \mathbb{R}^{k \times l}$ with independent columns,
- $O \in \mathbb{R}^{k \times l}$ with orthonormal columns.

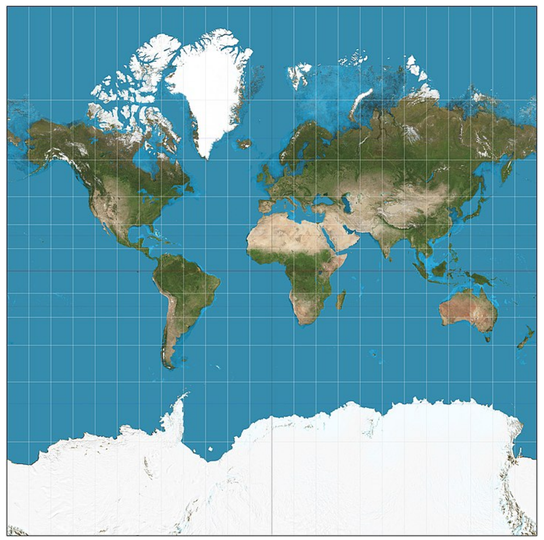
	W_2	W_1
$\mathcal{L}$	$U_{\mathcal{I}} G^{+}$	$G U_{\mathcal{I}}^{\intercal}$
$\mathcal{L}_{\pi}$	$U_{\mathcal{I}} (I_{\ell} + \lambda \Sigma_{\mathcal{I}}^{-2})^{-\frac{1}{2}} G^{+}$	$G (I_{\ell} + \lambda \Sigma_{\mathcal{I}}^{-2})^{-\frac{1}{2}} U_{\mathcal{I}}^{\intercal}$
$\mathcal{L}_{\sigma}$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

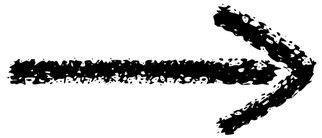
Algebraic Topology of PCA

Regularization reduces the symmetry from linear to orthogonal.

But LAEs are still over-parameterized, obscuring the gradient dynamics.

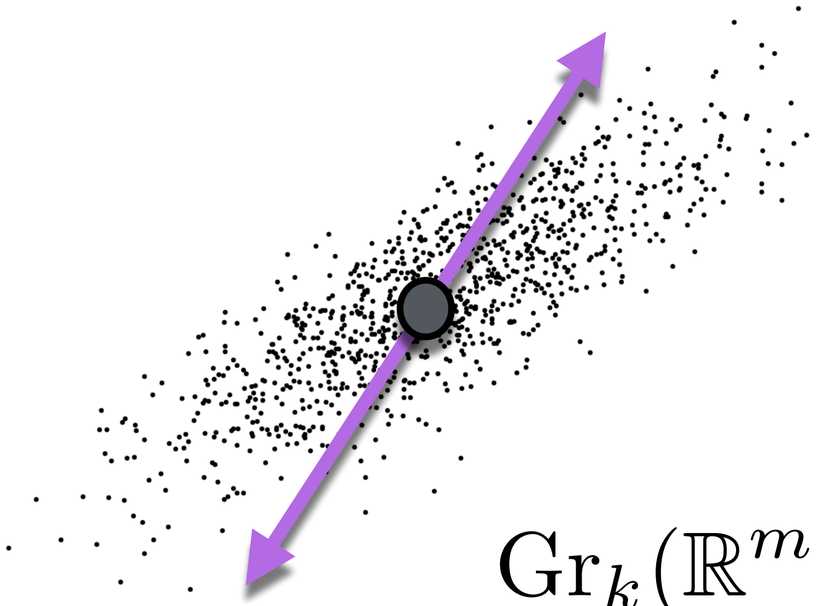
Issue: $\mathbb{R}^{k\times m} \times \mathbb{R}^{m\times k}$ is not the natural domain of PCA.





Algebraic Topology of PCA

Find the k-subspace nearest a point cloud in $\mathbb{R}^m$



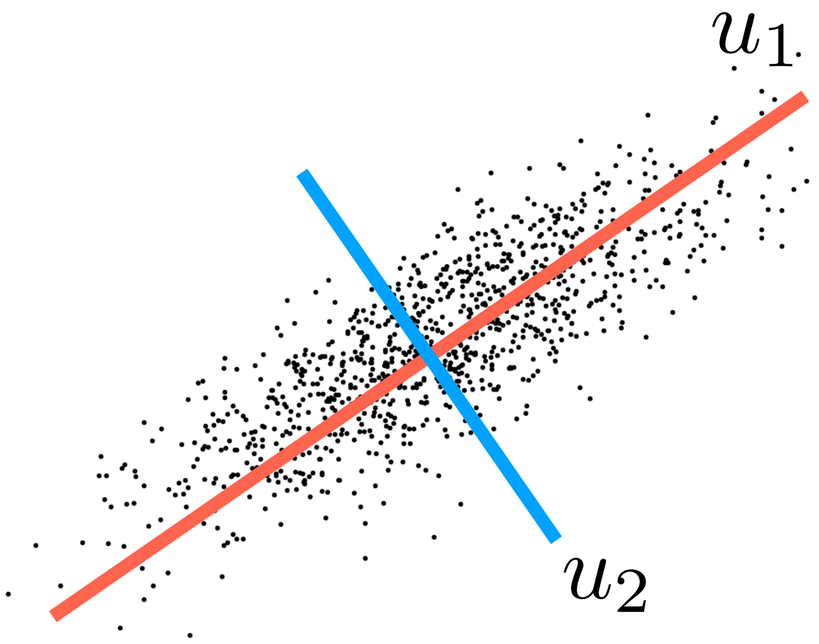
Natural domain is the *Grassmannian manifold.*

Points *are* subspaces. Distance is loss function.

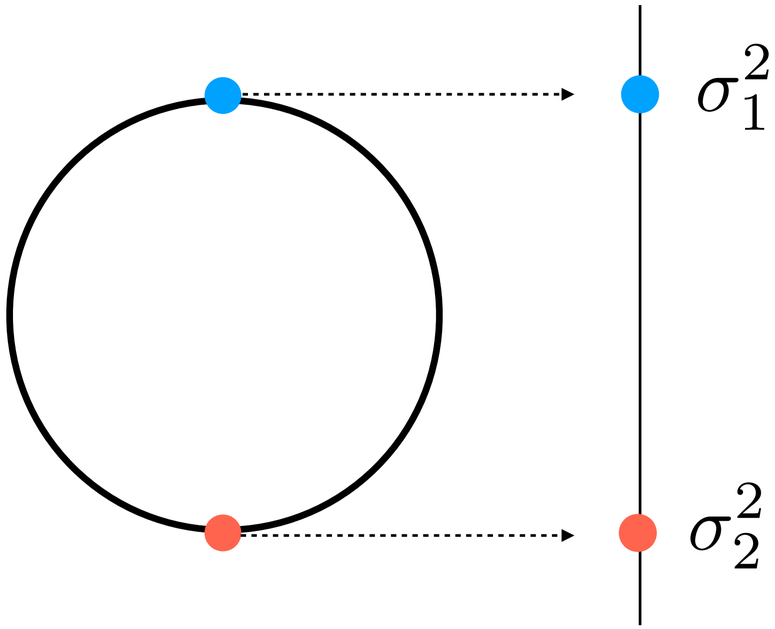
$$\mathrm{Gr}_k(\mathbb{R}^m) \cong \{P = P^2,\ P = P^\intercal,\ \mathrm{tr}\, P = k\} \subset \mathbb{R}^{m \times m}$$

$$\dim \mathrm{Gr}_k(\mathbb{R}^m) = k(m-k) \qquad \mathcal{L}_X(P) = ||X - PX||^2$$

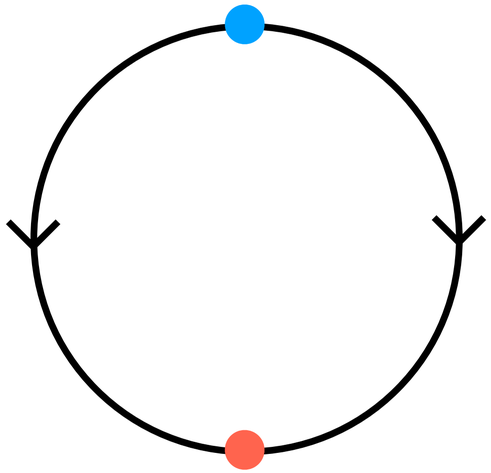
Algebraic Topology of PCA



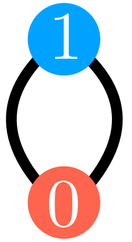
$X \subset \mathbb{R}^2$



$\mathrm{Gr}_1(\mathbb{R}^2) \xrightarrow{\mathcal{L}_X} \mathbb{R}$

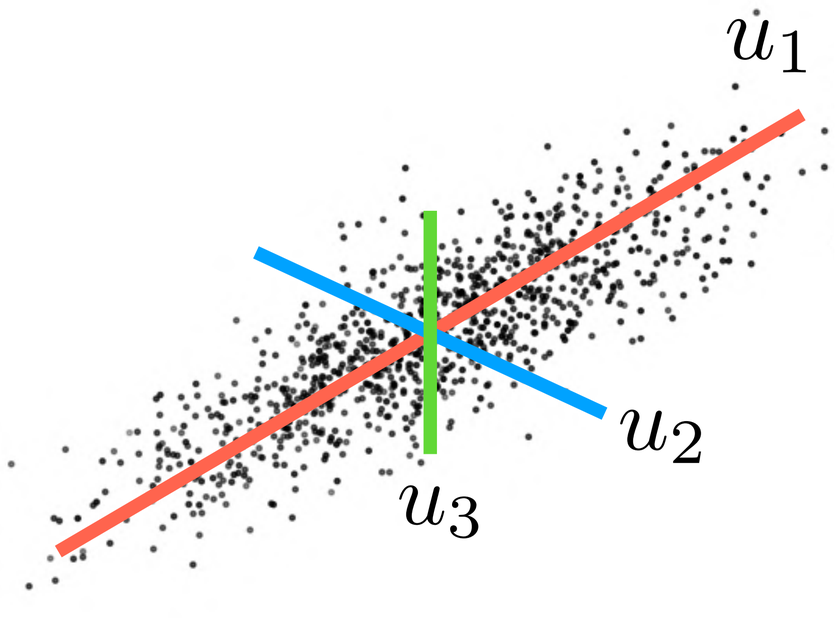


$-\nabla \mathcal{L}_X$

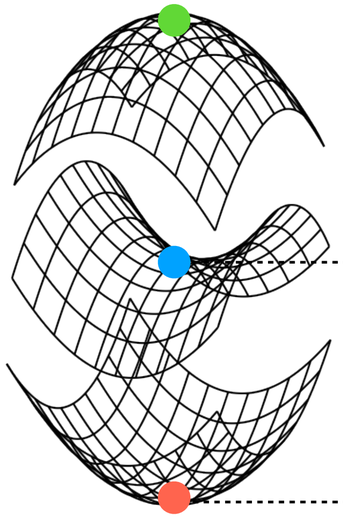


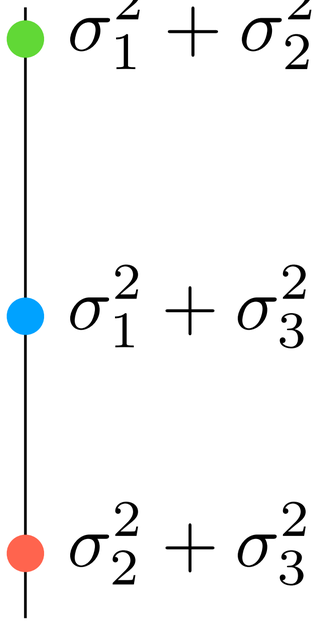
$H_*(\cdot)$

Algebraic Topology of PCA

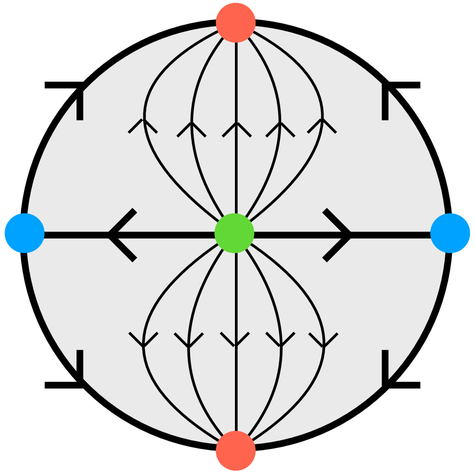


$$X \subset \mathbb{R}^3$$





$$\mathrm{Gr}_1(\mathbb{R}^3) \xrightarrow{\mathcal{L}_X} \mathbb{R}$$



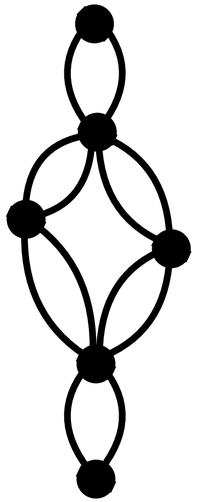
$$-\nabla \mathcal{L}_X$$

$$H_*(\cdot)$$

Algebraic Topology of PCA

$$\mathcal{L}_X : \mathrm{Gr}_2(\mathbb{R}^4) \to \mathbb{R}$$

d	u_1	u_2	u_3	u_4
4			•	•
3		•		•
2		•	•	
2	•			•
1	•		•	
0	•	•		



- $\binom{m}{k}$ critical points are principal subspaces
- Critical values are sums of eigenvalues
- Connecting gradient trajectories are rotations
- Loss is std saddle ($\mathbb{F}_2$-*perfect Morse* function)
- LAE loss is degenerate std saddle (*Morse-Bott*)
- Suggests principals, algorithms for deep learning

PCA Algorithms

PCA is a **two-step optimization**:

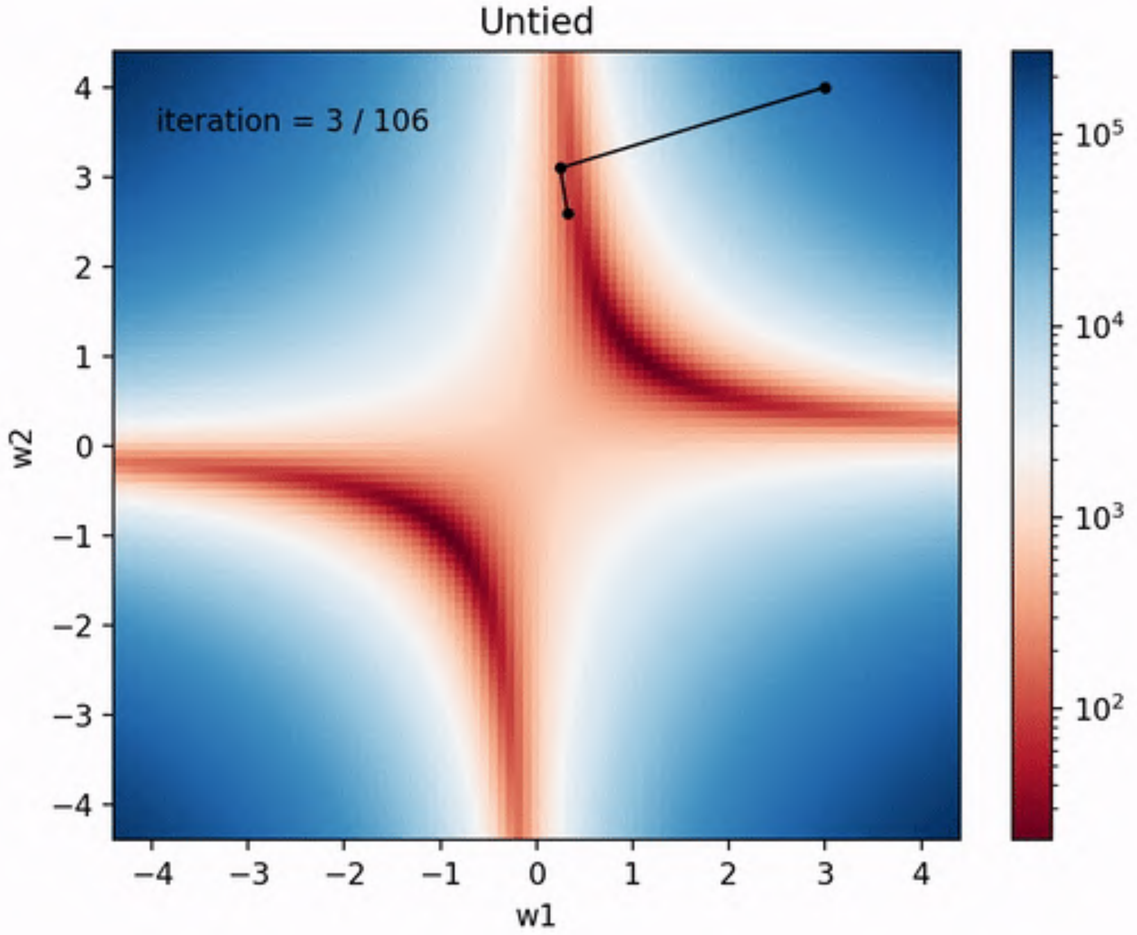
1. Train an L²-regularized LAE on $X \subset \mathbb{R}^{m \times n}$

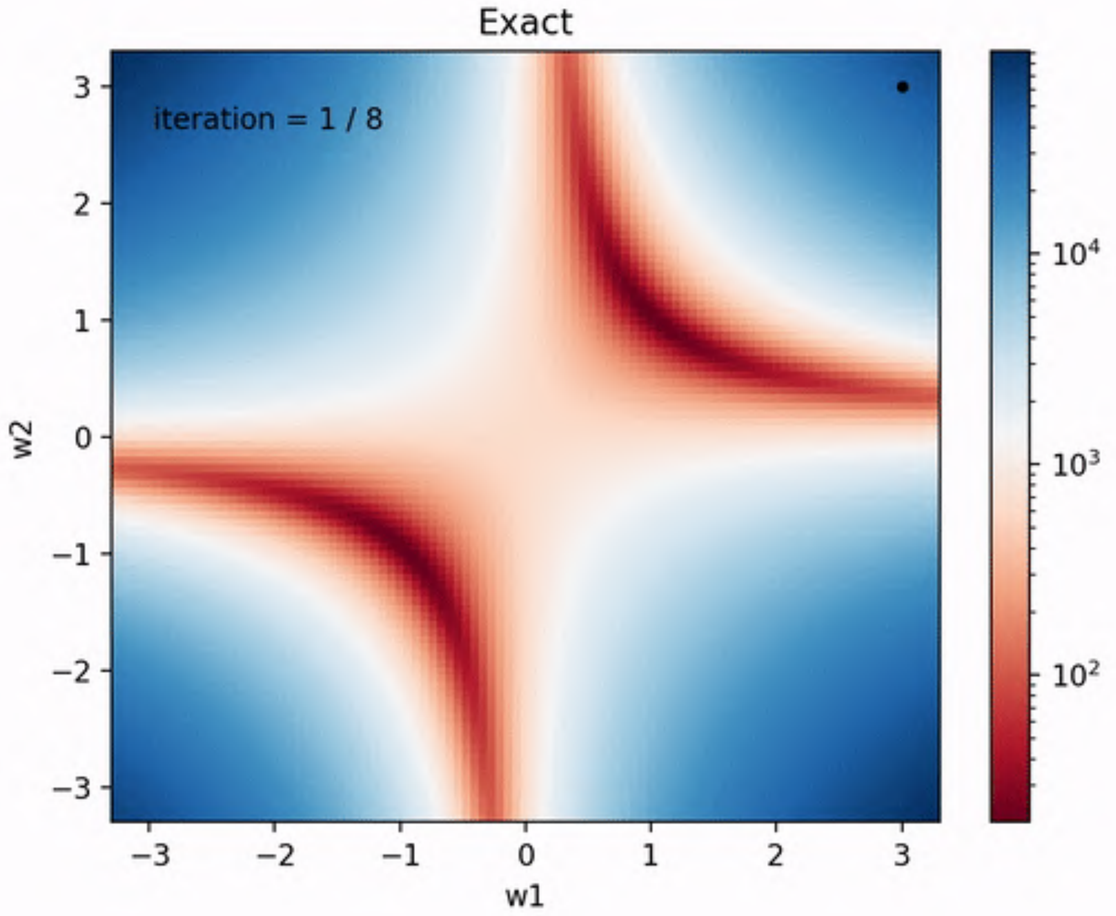
An Adventure!

2. Apply SVD to the decoder $W_2 \subset \mathbb{R}^{m \times k}$

Quick!

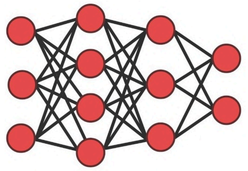
PCA Algorithms





Gradient descent

Solve for W_2, set $W_1 = W_2^\intercal$



Prediction in artificial neural networks is inspired by the brain.

Is *learning* in the brain inspired by artificial neural networks?

COGNITIVE SCIENCE 11, 23–63 (1987)

Competitive Learning: From Interactive Activation to Adaptive Resonance

STEPHEN GROSSBERG

Boston University

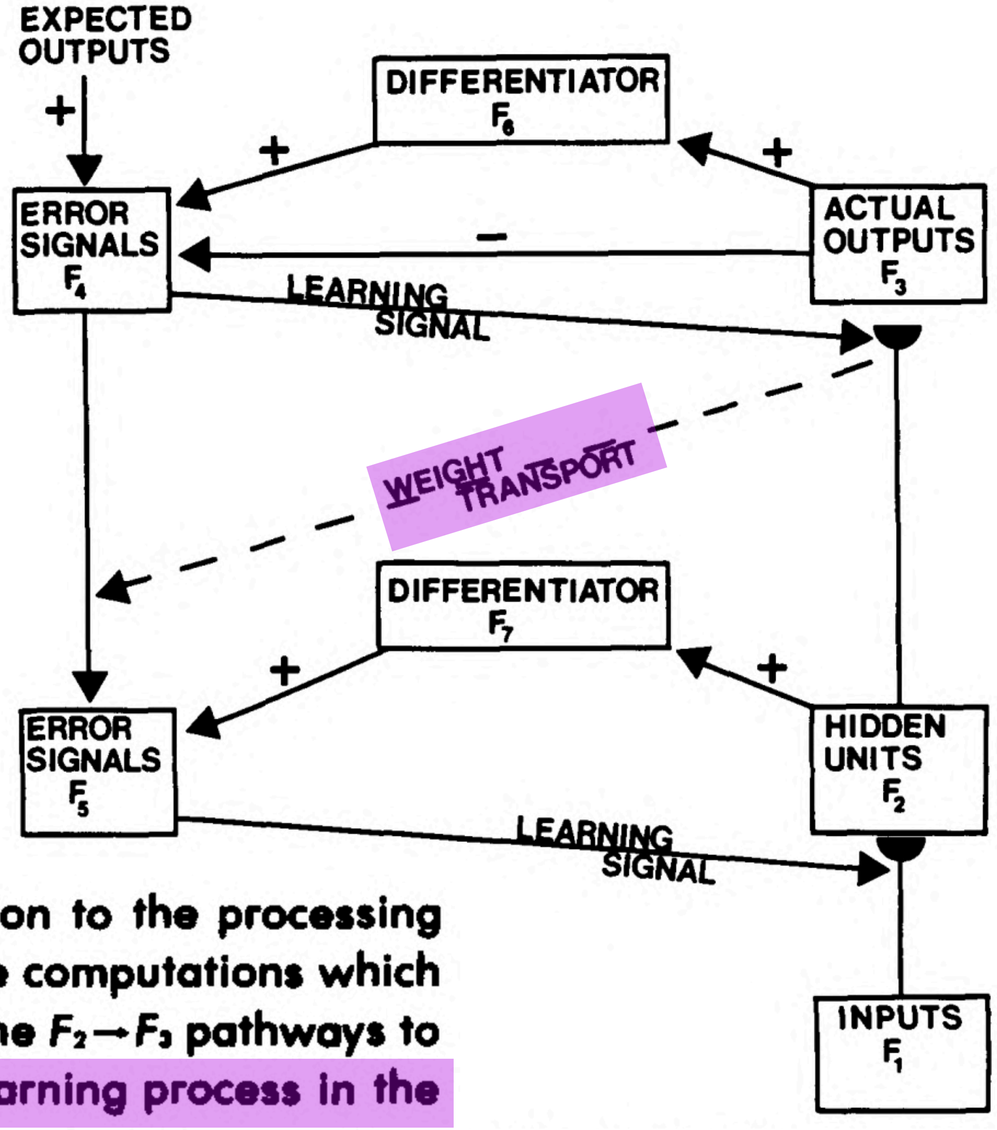
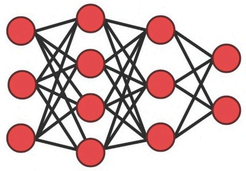
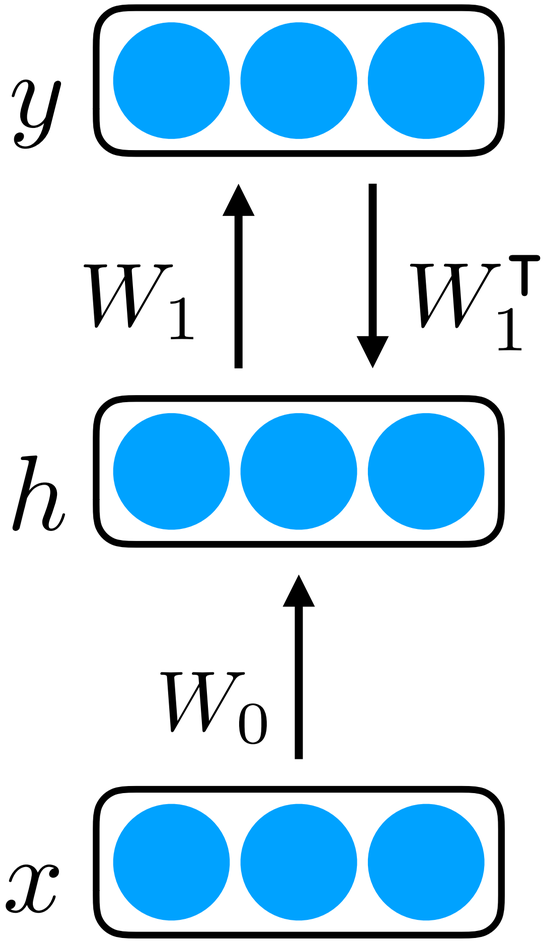


Figure 8. Circuit diagram of the back propagation model: In addition to the processing levels F_1, F_2, F_3, there are also levels F_4, F_5, F_6, and F_7 to carry out the computations which control the learning process. The transport of learned weights from the $F_2 \rightarrow F_3$ pathways to the $F_4 \rightarrow F_5$ pathways shows that this algorithm cannot represent a learning process in the brain.



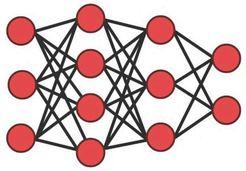
Prediction in artificial neural networks is inspired by the brain.

Is *learning* in the brain inspired by artificial neural networks?



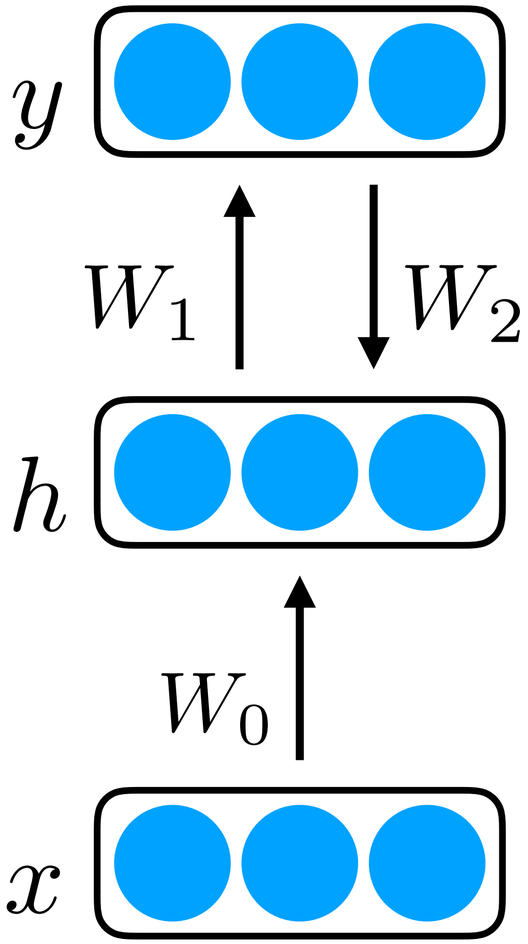
weight transport problem

$$\Delta W_0 \propto W_1^\intercal e x^\intercal$$



Prediction in artificial neural networks is inspired by the brain.

Is *learning* in the brain inspired by artificial neural networks?



Thm: L_2-regularized LAEs *symmetrize*.

W_2 *dynamically* aligns to $W_1^\intercal$ by:

- maximizing flow of information
- minimizing energy

$$\mathcal{L}_{\text{IA}} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{info}} + \mathcal{L}_{\text{reg}}$$

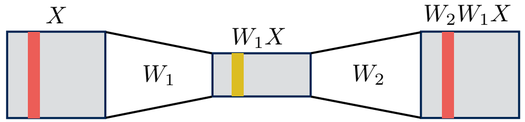
Loss Landscapes of Regularized Linear Autoencoders

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

[1] Broad Institute, Stanley Center for Psychiatric Research at Broad Institute
[2] Stanford University

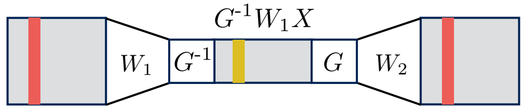
Background

A **linear autoencoder** maps $\mathbb{R}^m \to \mathbb{R}^k \to \mathbb{R}^m$.



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

LAEs learn the top principal *subspace* but **not** the principal *directions* or eigenvalues. The optimal latent representation is only defined up to a **linear map** $G \in \mathrm{GL}_k(\mathbb{R}^m)$.



LAEs are **pseudoinverses** at all critical pts.

Regularization

We prove that L²-*regularized* LAEs are **transposes** at all critical points and learn the principal directions as the left singular vectors of the decoder. Define $\mathcal{L}_\sigma$ by

$$||X - W_2 W_1 X||^2 + \lambda(||W_1||^2 + ||W_2||^2)$$

The minima of $\mathcal{L}_\sigma$ are defined up to an **orthogonal map** $O \in \mathrm{O}_k(\mathbb{R}^m)$ by

$$W_2 = U_k(I - \lambda\Sigma_k^2)^{\frac{1}{2}} O = W_1^\intercal$$

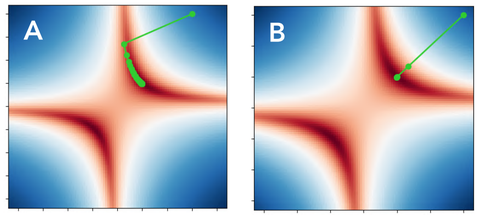
where $X = U\Sigma V^T$ and $\sigma_1^2 > \cdots > \sigma_k^2 > \lambda$.

$$W_2 W_1 = U_k(I - \lambda\Sigma_k^2)U_k^\intercal$$

PCA Algorithms

Hence PCA is a **two-step optimization**:
1. Train L²-regularized LAE on $X \subset \mathbb{R}^{m \times n}$.
2. Apply SVD to the decoder $W_2 \subset \mathbb{R}^{m \times k}$.

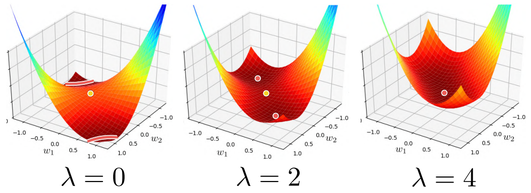
Step 2 is quick. Step 1 options include:
A. Gradient descent (below).
B. Solve for W_2, set $W_1 = W_2^\intercal$, iterate.



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

Posterior Collapse

Principal directions with eigenvalues below λ collapse as in **probabilistic PCA**.



Example of collapse for $X = [2]$.

Symmetry and Backprop

L²-reg LAEs are **symmetric** at all critical pts.

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

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

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

We first prove that the matrix

$$C = (I - W_2 W_1)XX^\intercal$$

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

$$XX^\intercal(W_2 W_1)^\intercal = (W_2 W_1)XX^\intercal(W_2 W_1)^\intercal + \lambda W_2 W_2^\intercal.$$

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

$$XX^\intercal(W_2 W_1)^\intercal \succeq (W_2 W_1)XX^\intercal(W_2 W_1)^\intercal.$$

Cancelling $(W_2 W_1)^\intercal$ via Lemma B.1 gives $C \succeq 0$.

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

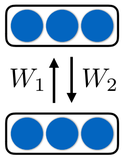
$$0 = \frac{\partial \mathcal{L}_\sigma}{\partial W_1} - \frac{\partial \mathcal{L}_\sigma}{\partial W_2}^\intercal = 2A(C + \lambda I).$$

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

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

that $A = 0$. □

Resolution to **weight transport problem**:



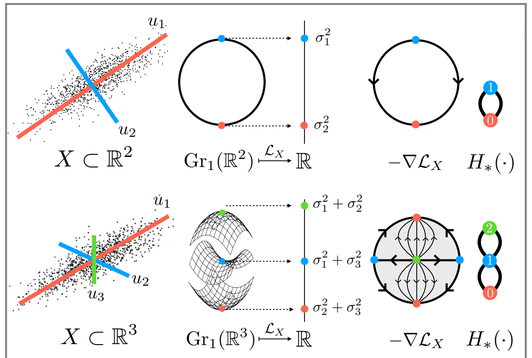
Backprop in  lacks $W_1^\intercal$ because neurons go one way.

Learn as W_2 by maximizing flow of info and minimizing energy.

Algebraic Topology

We **smoothly parameterize** the critical manifolds of LAEs with several forms of regularization via one elementary proof.

We factor the loss as a **Morse function** on the Grassmannian to reveal the dynamics near and between critical manifolds. **Morse homology** suggests principles and algorithms for deep learning.



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.

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

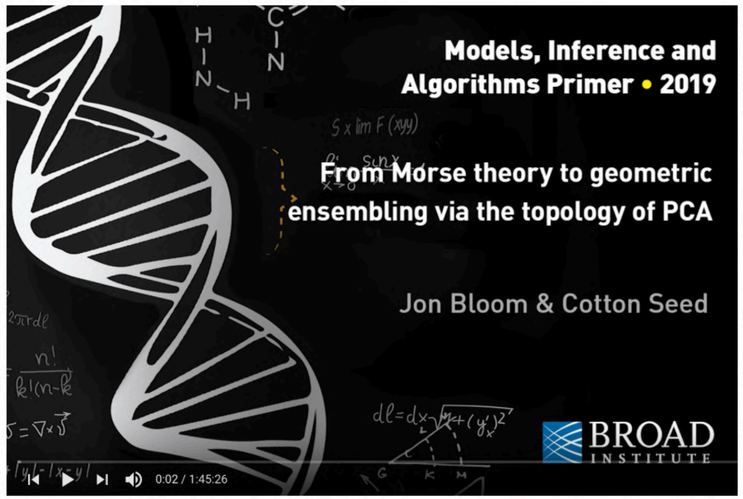
$$\begin{array}{ccc} \mathrm{V}_k(\mathbb{R}^m) & \xrightarrow{\pi : O \mapsto \mathrm{Im}(OO^\intercal)} & \mathrm{Gr}_k(\mathbb{R}^m) \\ {\scriptstyle \iota : O \mapsto (O^\intercal, O)}\big\downarrow & & \big\downarrow{\scriptstyle \mathcal{L}_X} \\ \mathbb{R}^{k \times m} \times \mathbb{R}^{m \times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

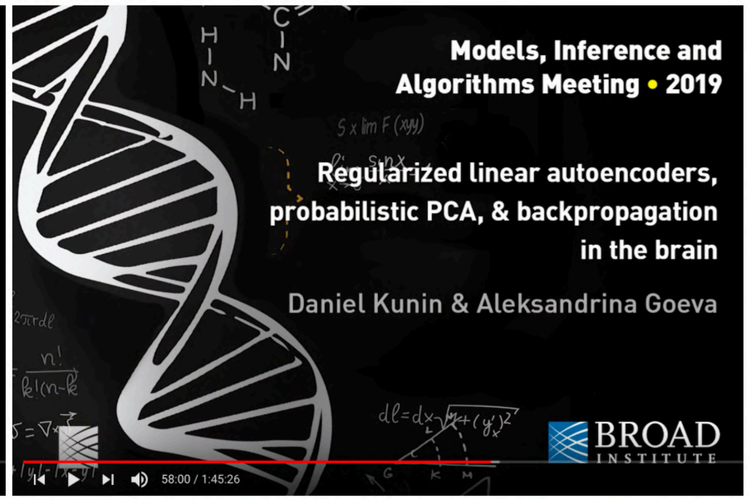
MIA talks at broadinstitute.org/mia

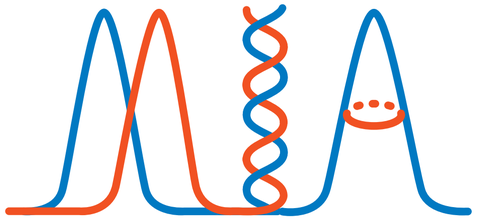
In the last decade, **biology** has been transformed by the ability to perturb and measure biological systems at massive scale.

However, new ideas in **ML** are needed to translate biomedical data into a *mechanistic understanding* of biology and disease.

Biology and **ML** are poised to powerfully advance one another.







broadinstitute.org/mia