Weight Matrix Evolution during MiniAlexNet training

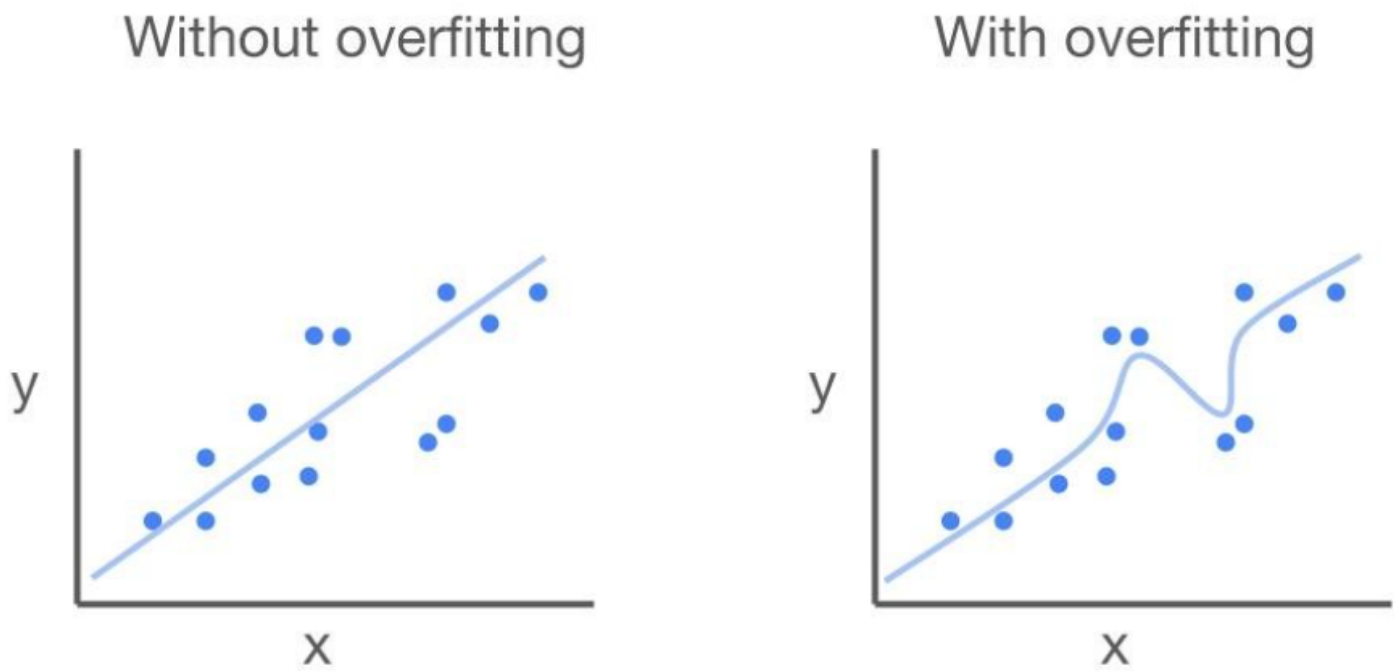
Presenter: Doyoon Kim, Enoch Yiu

UC Berkeley

2nd May, 2025

Background

According to Martin&Mahoney, Regularization ≈ Generalization
So that if in order for the deep learning model to perform well on unseen data(test data), it should be regularized well.
To prevent being overfitted to training data, the weight matrices of the model should be 'simple'.



Background

There are two kinds of regularization: explicit vs. implicit

- explicit regularization:
  - Dropout: randomly zeroes a fraction of activations during training
  - Weight norm constraints, e.g.
    - Lasso (L_1 penalty):

$$\min_w \frac{1}{N} \sum_{i=1}^{N} L\big(f(x_i; w), y_i\big) + \lambda \|w\|_1$$

- implicit regularization:
  Neural network's training dynamics drive weight matrices toward a "simple" state

How can we "check" that the matrix is getting simpler?

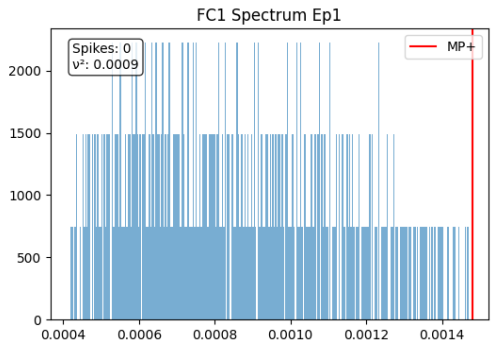
Background

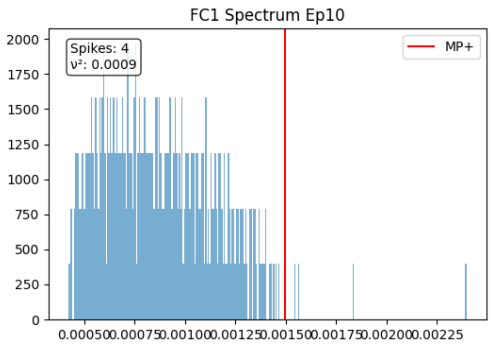
Marchenko–Pastur (MP) Law: For a random matrix $W \in^{p \times n}$ with iid entries of variance σ^2, the eigenvalue density of $\frac{1}{n} W W^T$ converges to

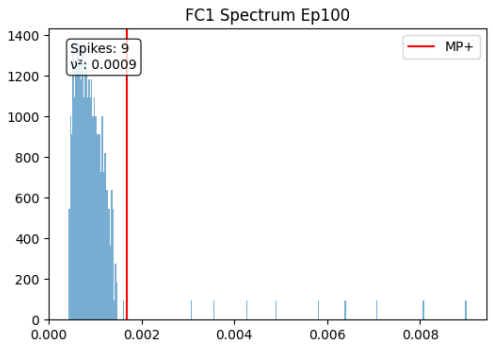
$$\rho_{\mathrm{MP}}(\lambda) = \frac{1}{2\pi\sigma^2 c\,\lambda}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}, \quad \lambda_\pm = \sigma^2(1 \pm \sqrt{c})^2, \quad c = \frac{p}{n}.$$

The interval $[\lambda_-, \lambda_+]$ is the *noise bulk*; eigenvalues outside are *spikes* (signal).

Background







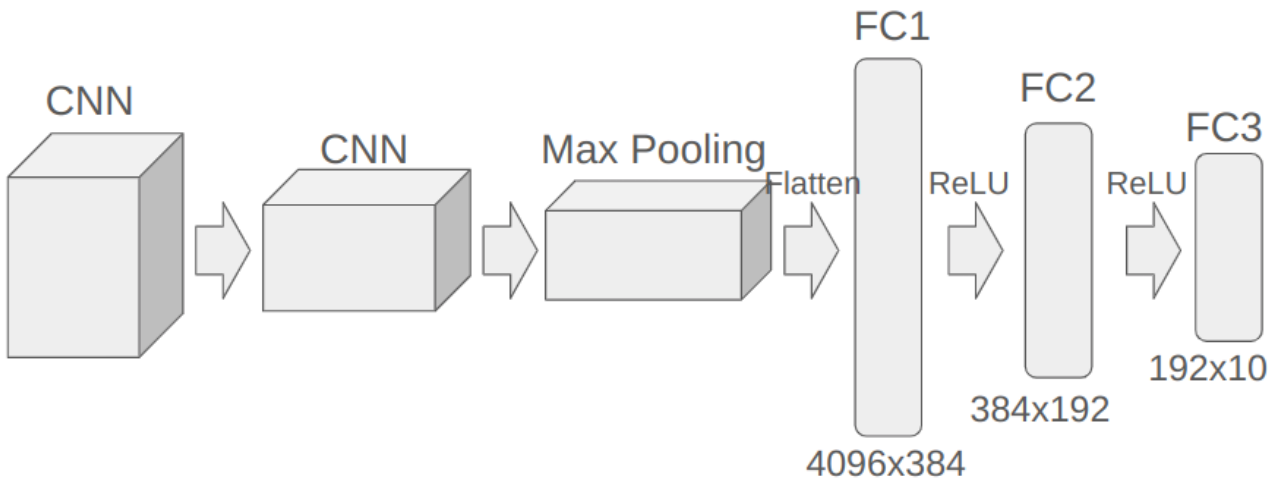
This growth in spike count signals that the weight matrix develops a simpler, more low-rank "signal" component.

Questions

- If the weight matrix follows MP law, can we reconstruct the matrix with a few spikes(factors)?
- Is it possible to shrink Eigenvectors of covariance matrices using JSE to improve regularization?
- If it's possible, where should we shrink it?

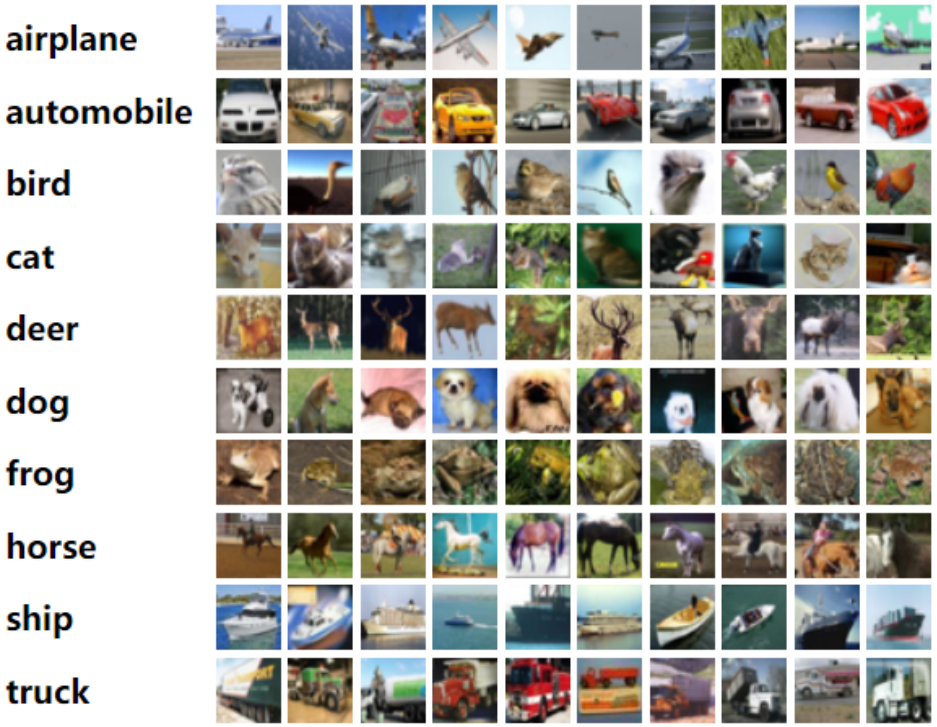
Model Architecture

Used the same architecture used in Martin&Marhoney, which is miniAlexnet, simplified version of alexnet.
It has 2 CNN layers, MaxPooling layer with 3 Fully Connected layers, which doesn't include explicit regularization.



Dataset

We used CIFAR10 dataset to train minialexnet.
The data consists of 60000 32x32 colour images in 10 classes, with 50000 training images and 10000 test images.



Reconstruction

Q. If the weight matrix follows MP law, can we reconstruct the matrix with a few spikes(factors)?
A. We can approximate it, but not perfect.
For FC1 (W_1), FC2 (W_2), and FC3 (W_3)

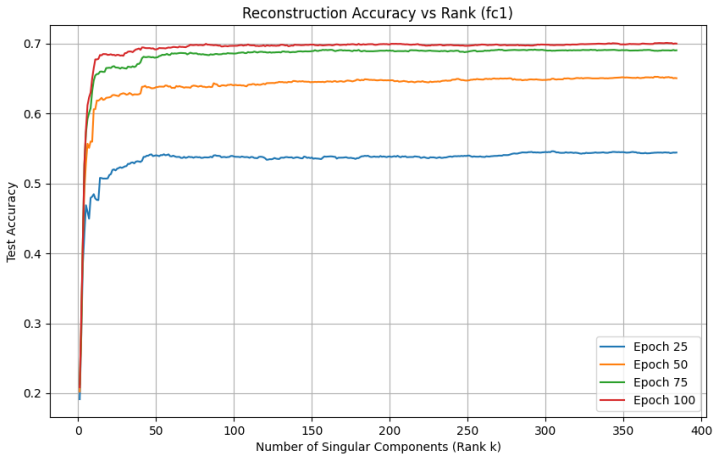
$$W_i = U\Sigma V^T = U_{p\times k}\,\Sigma_{k\times k}\,V^T_{n\times k} + Z,$$

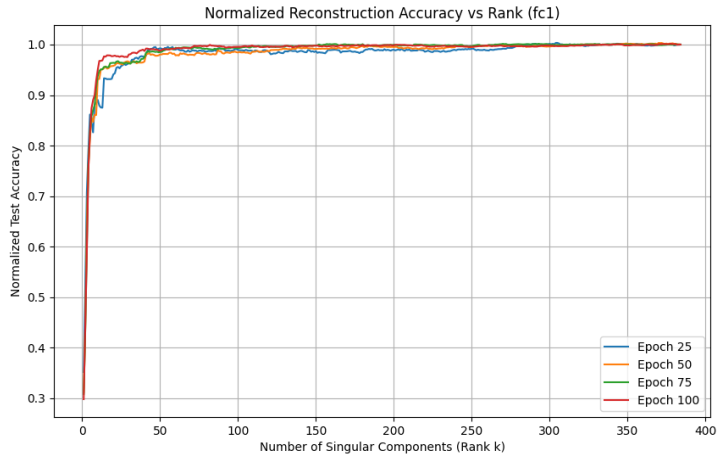
$$\hat{W}_i(k) = U_{p,k}\,\Sigma_{k,k}\,V^T_{n,k} = \sum_{i=1}^{k}\sigma_i\,u_i\,v_i^T,$$

$$\mathrm{Acc}(\hat{W}_i(k)) = \frac{\#\text{correct predictions}}{\#\text{test samples}}.$$

Reconstruction

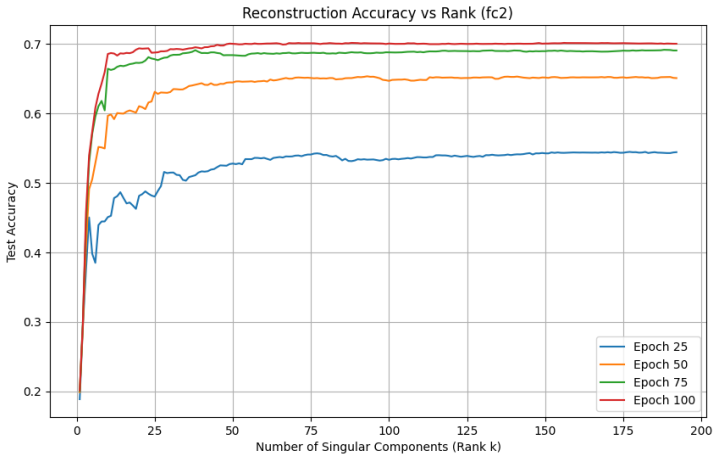
Fixed FC2, FC3 matrices and varied k in FC1 weight matrix(4096x384). As we train the model, the number of factors k needed to approximate full rank decreased.

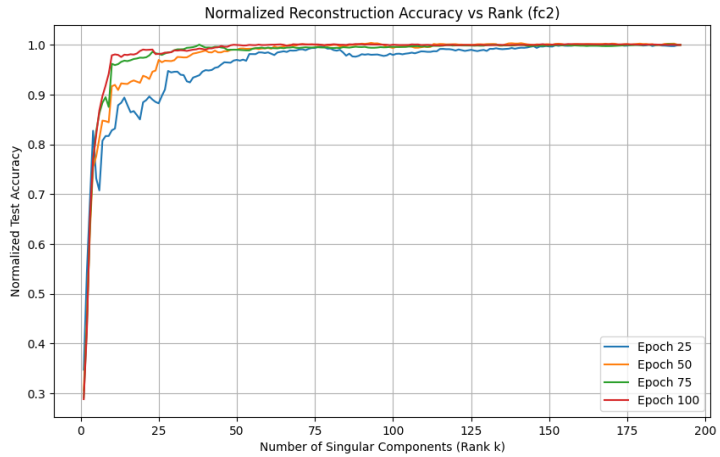




Reconstruction

Similarly for FC2 layer(384x192),
As we train the model, the number of factors k needed to approximate full rank decreased.





James–Stein Shrinkage for SVD Reconstruction

Q. Is it possible to shrink Eigenvectors of covariance matrices using JSE to improve regularization?
A. No. If we shrink it toward the mean.
Shrink each eigenvectors towards its mean

$$W = U\,\Sigma\,V^T \quad \longrightarrow \quad H \;=\; \frac{1}{\sqrt{n}}\,U_{:,1:k}\,\Sigma_{1:k,1:k}$$

$$M \;=\; \frac{1}{p}\,\mathbf{1}\mathbf{1}^T H, \quad R \;=\; H - M$$

$$\nu^2 \;=\; \frac{(RR^T)}{n_+ - k}, \quad J \;=\; R^T R, \quad C \;=\; I_k - \nu^2\,J^{-1}$$

$$\boxed{H_{\mathrm{JS}} \;=\; H\,C \;+\; M\,(I_k - C)} \quad \Longrightarrow \quad H_{\mathrm{JS}} = U_{\mathrm{JS}}\,\Sigma_{\mathrm{JS}}\,V_{\mathrm{JS}}^T$$

James–Stein Shrinkage for SVD Reconstruction

(a) Vector-only shrinkage:

$$\widehat{W}_{\text{vec}}(k) = U_{\text{JS}}\, \Sigma_{1:k,1:k}\, V^T_{:,1:k}$$

(b) Full shrinkage of singular values:

$$\widehat{W}_{\text{full}}(k) = U_{\text{JS}}\, \Sigma_{\text{JS}}\, V^T_{:,1:k}$$

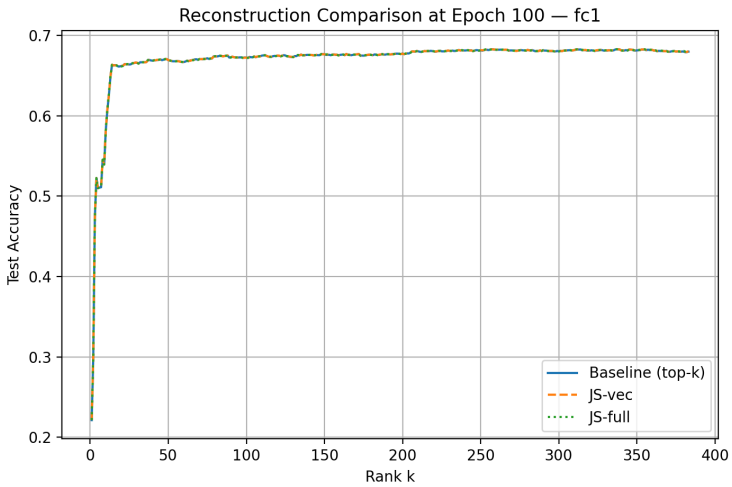
where

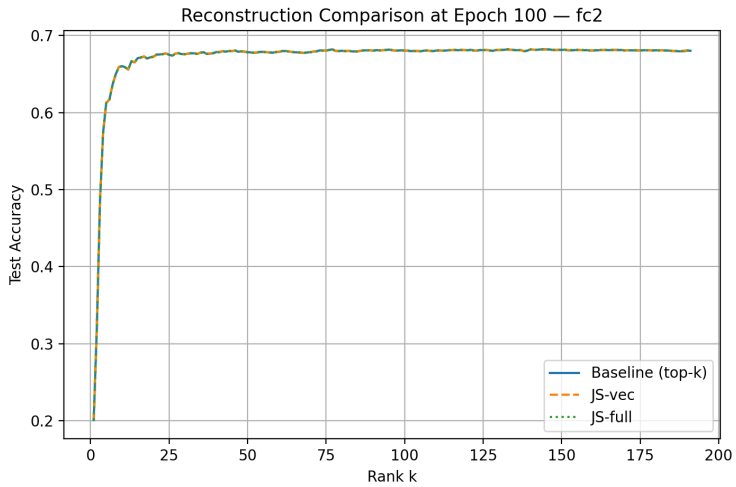
$$S^2 = \frac{1}{n}\Sigma^2, \quad \Psi^2 = I - \nu^2 S^{-2} = I - \nu^2 n \Sigma^{-2}, \quad \Phi = S^2 \Psi^2 = \frac{1}{n}\left(\Sigma^2 - n\nu^2 I\right),$$

$$\boxed{\Sigma_{\text{JS}} = (n\,\Phi)^{\frac{1}{2}} = \left(\Sigma^2 - n\,\nu^2 I\right)^{\frac{1}{2}} = \text{diag}\left(\sqrt{\sigma_i^2 - n\,\nu^2}\right)}$$

James–Stein Shrinkage for SVD Reconstruction

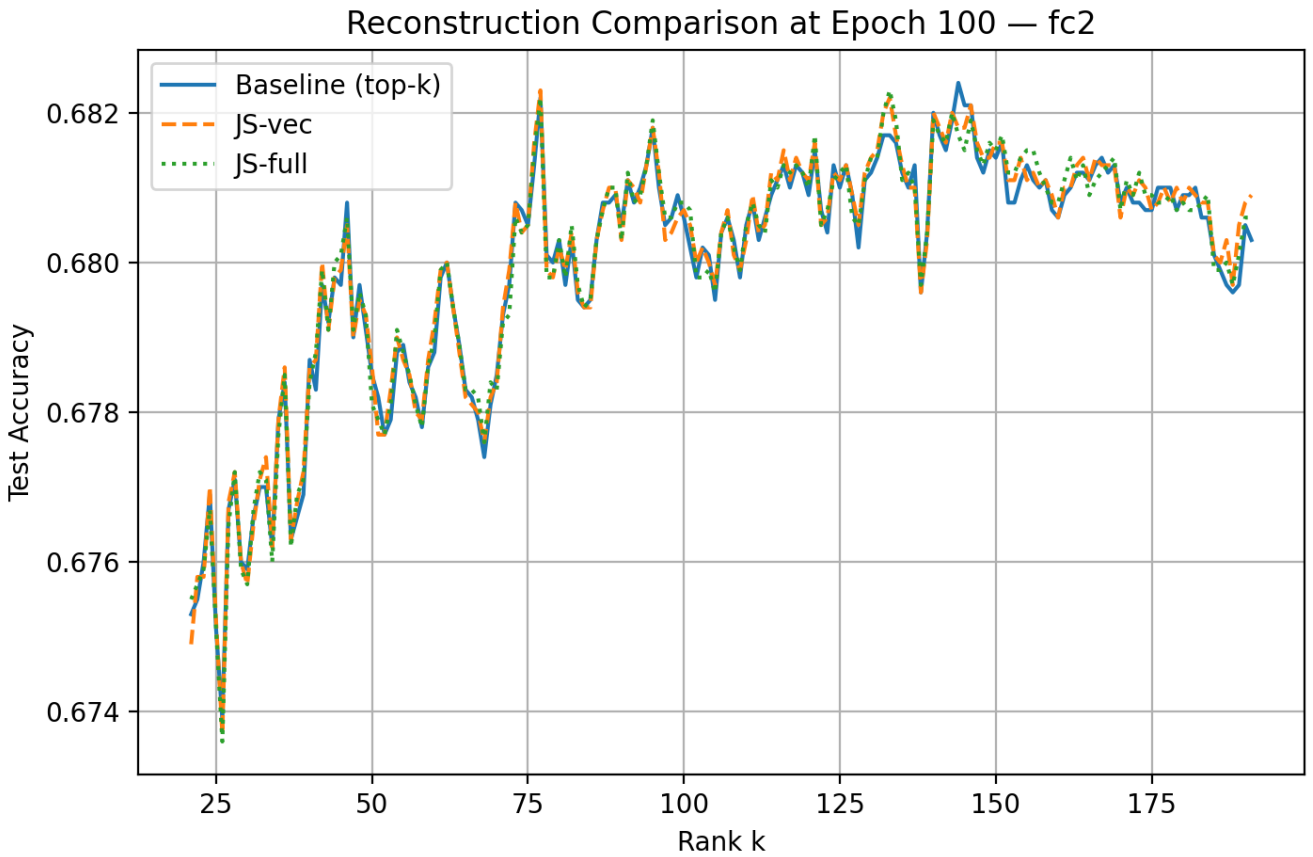
No dramatic increase in performance no matter what k is.





James–Stein Shrinkage for SVD Reconstruction

Zooming in, there are some factor numbers that JS shrinkage works better, but more like coincidence.



James–Stein Shrinkage for SVD Reconstruction

We need better shrinkage target than grand mean point.
For that, we checked the behavior of leading eigenvector upon epoch.

Geometric PCA Projections

Let $x_i \in^p$ be the leading eigenvector at epoch i, $\|x_i\| = 1$.

Define the projection basis matrix A by:

$$A = \begin{cases} \left[b \;\; \frac{t}{\|t\|} \right] \in^{p \times 2}, & \text{for 1D GPCA (tangent line);} \\ \left[b \;\; t_1 \;\; t_2 \right] \in^{p \times 3}, & \text{for 2D GPCA (tangent plane).} \end{cases}$$

Then in both cases:

$$P = A\,(A^T A)^{-1} A^T, \quad \hat{x}_i = \frac{P\,x_i}{\|P\,x_i\|} \in S^{p-1}.$$

RSS and Fitting Score

$$\text{Residual Sum of Squares (RSS)} \;=\; \sum_{i=1}^{L} d^2\big(x_i, \hat{x}_i\big) \quad \big(x_i, \hat{x}_i \in S^{p-1}\big),$$

where $d(\cdot,\cdot)$ is the geodesic distance on the sphere.

$$\text{Mixed Variance} \;=\; \underbrace{\sum_{i=1}^{L} d^2\big(x_i, \hat{x}_i\big)}_{\text{RSS}} \;+\; \underbrace{\sum_{i=1}^{L} d^2\big(\hat{x}_i, \mu\big)}_{\text{variance on } S^{p-1}},$$

where μ is the Fréchet mean of $\{x_i\}$, i.e. the point on S^{p-1} that

$$\mu \;=\; \arg\min_{p \in S^{p-1}} \sum_{i=1}^{L} d^2\big(x_i,\, p\big).$$

RSS and Fitting Score

$$\text{Fitting Score} \;=\; R^2 \;=\; 1 \;-\; \frac{\text{RSS}}{\text{Mixed Variance}} \qquad \in [0,1],$$

measuring the proportion of total dispersion explained by the projection.

2D GPCA fitting example

With proper learning rate(when weight doesn't converge within 100 epochs), leading eigen vector tend to follow geodesics

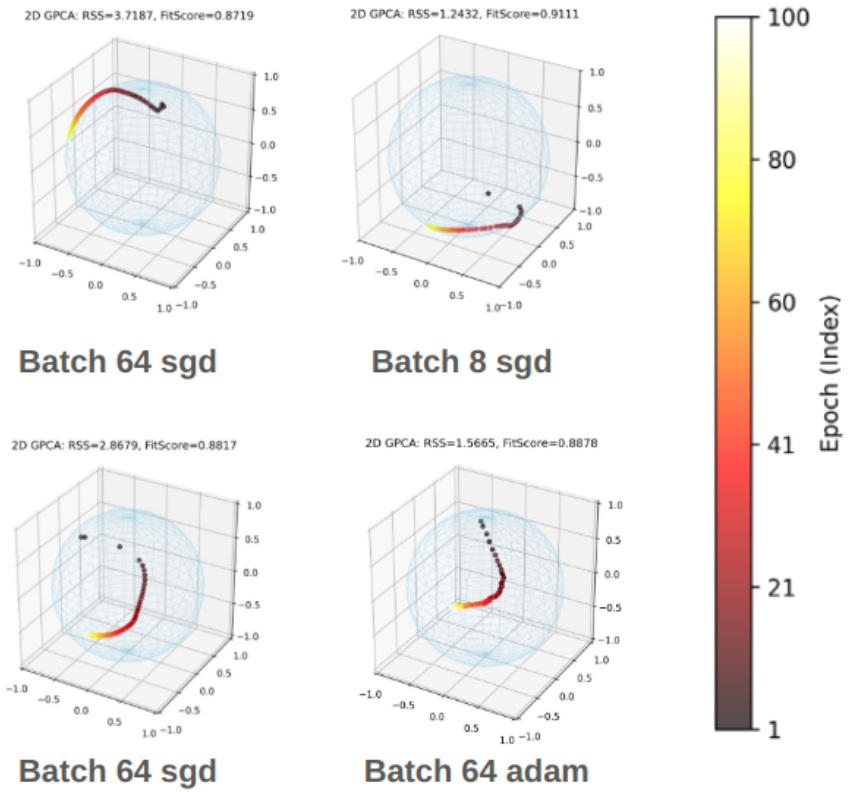


Figure 1: FC1 results

2D GPCA fitting example

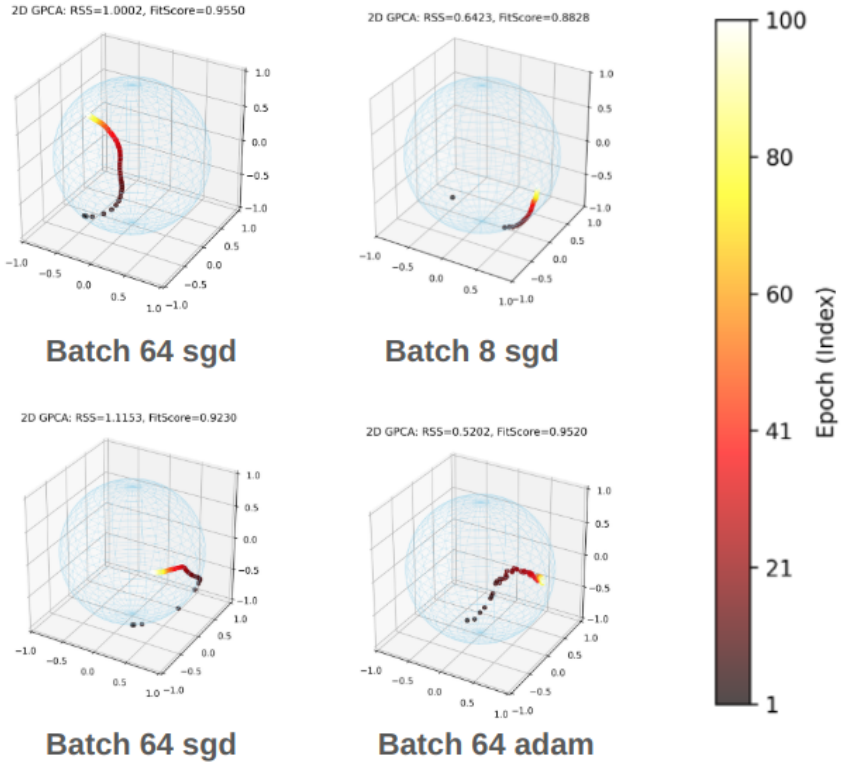
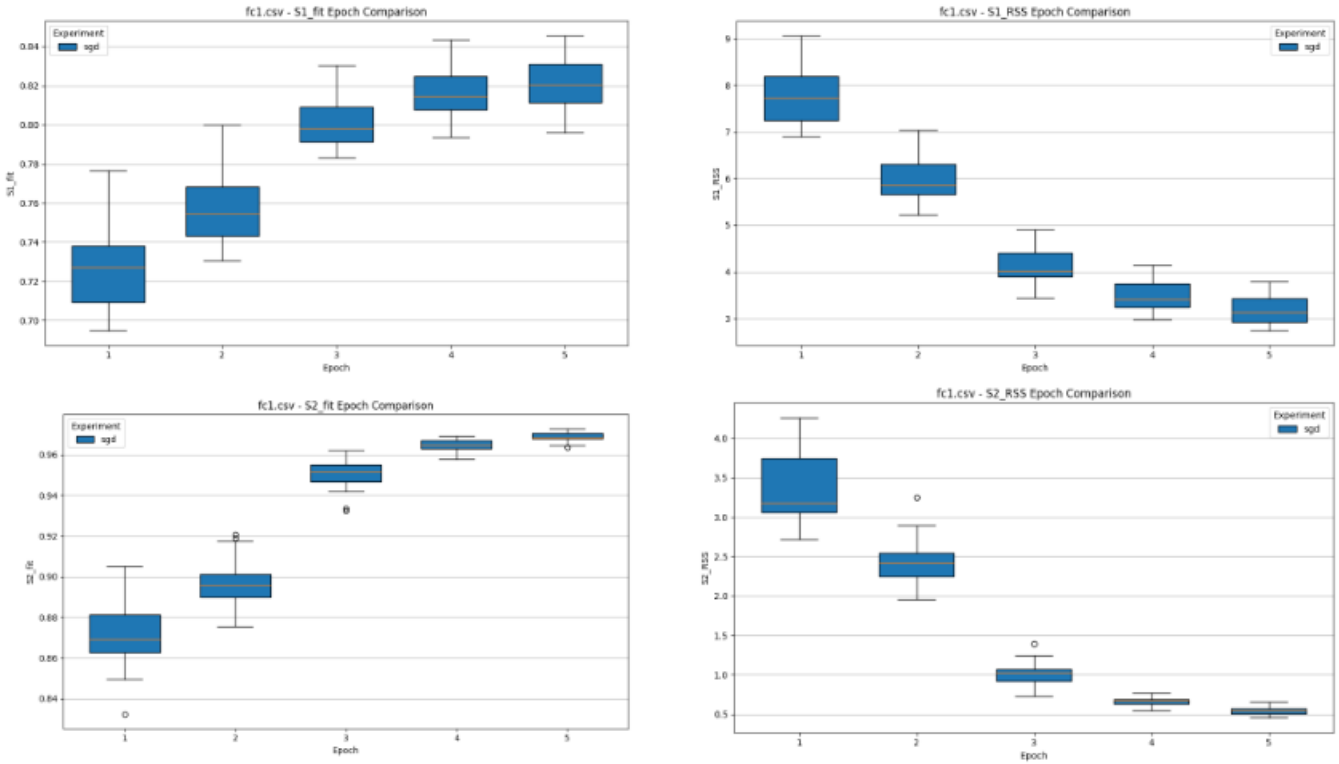


Figure 2: FC2 results

Choosing start of trajectory

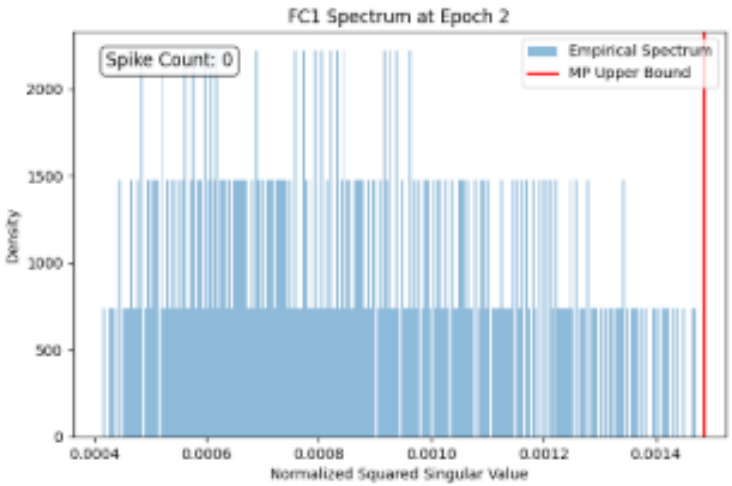
The trajectory of leading eigenvector stabilizes after few epochs.
The first few vectors have high residuals.
Truncating those leads to lower RSS and higher fitting score.

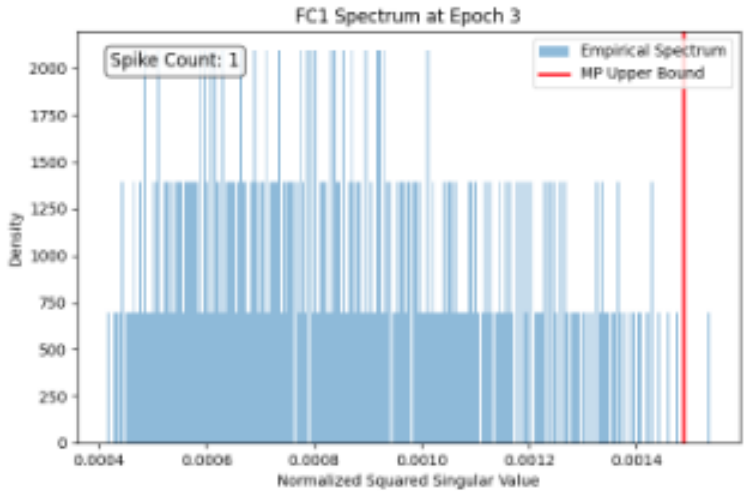


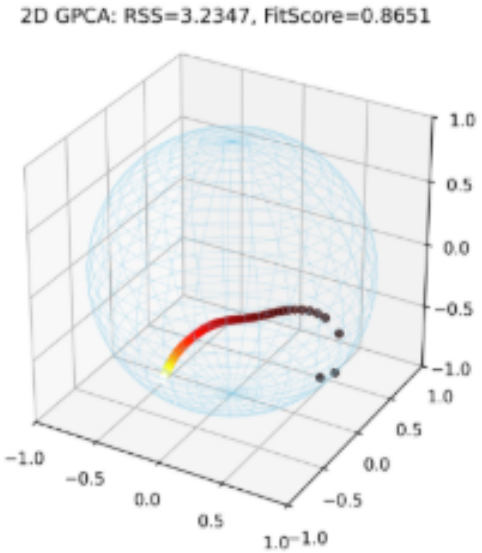
Choosing start of trajectory

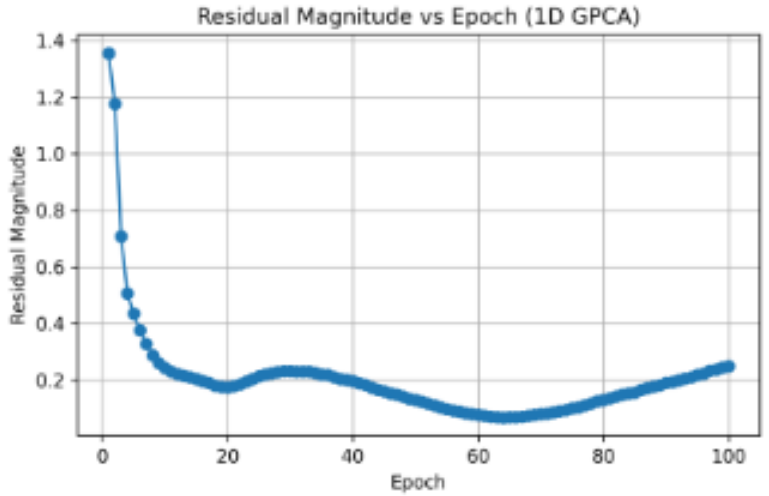
However, the optimal "start" of the trajectory depends on your hyperparameters.
To choose it automatically, monitor each epoch's weight-matrix spectrum using Marchenko–Pastur law.
The first epoch at which a clear outlier (spike) appears in the spectrum is then taken as the beginning of your trajectory.

Choosing start of trajectory



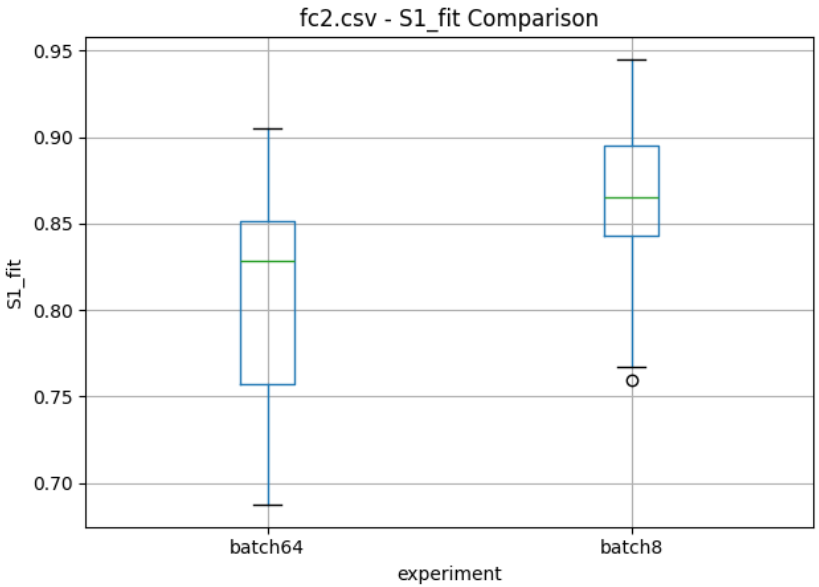


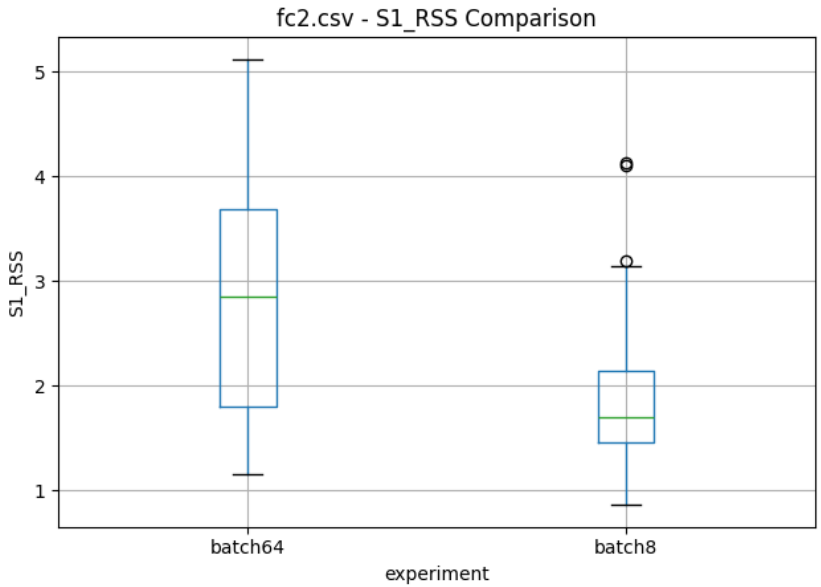




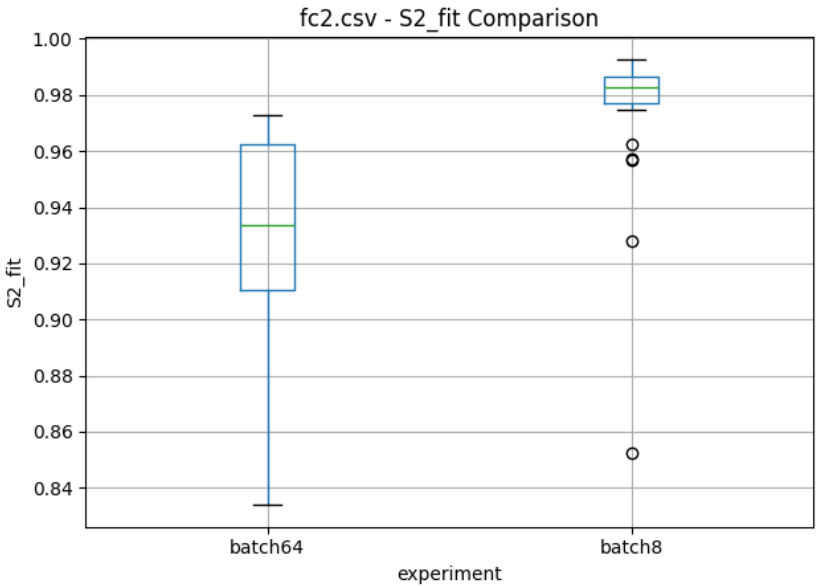
Multiple seed experiments

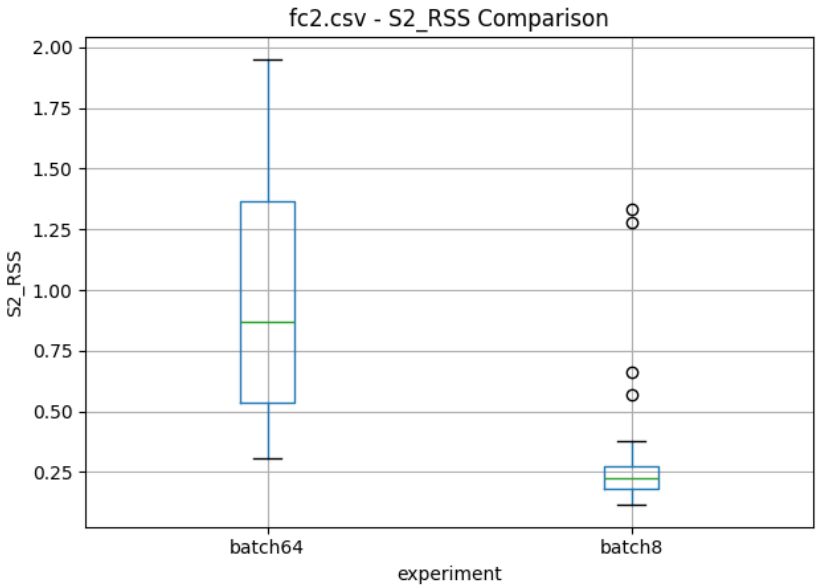
Trained model with 20 different seeds and projected 3-100 epoch points into S1, S2 sphere.
Recall: RSS of Random Walk was about 200, Fitting score was about 0.55



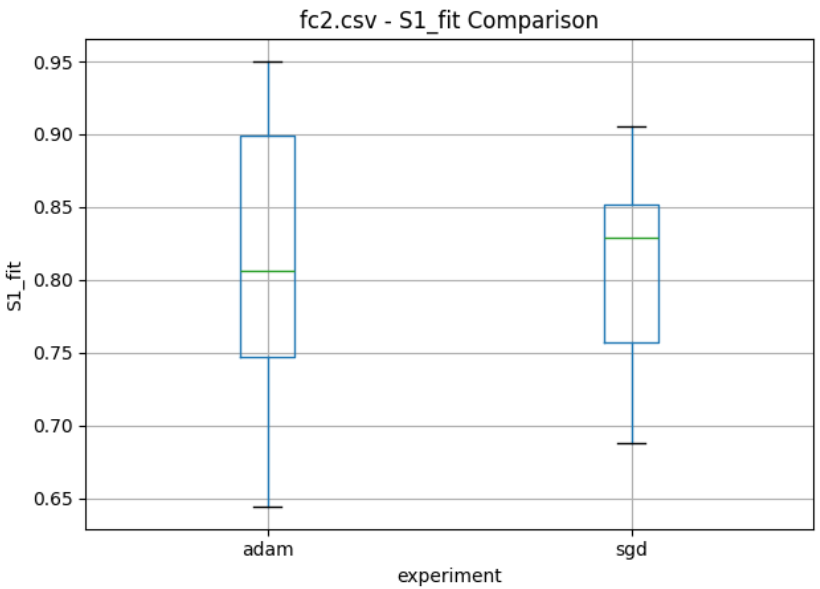


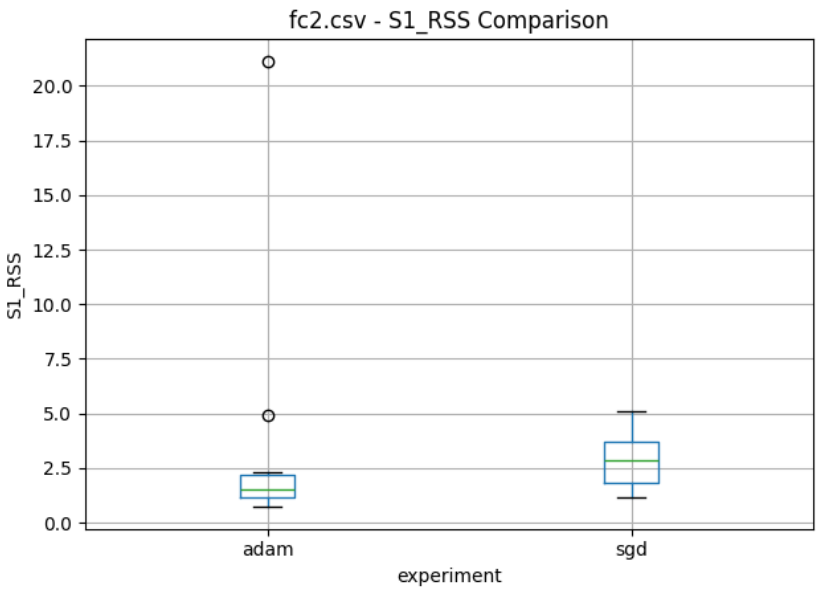
Multiple seed experiments



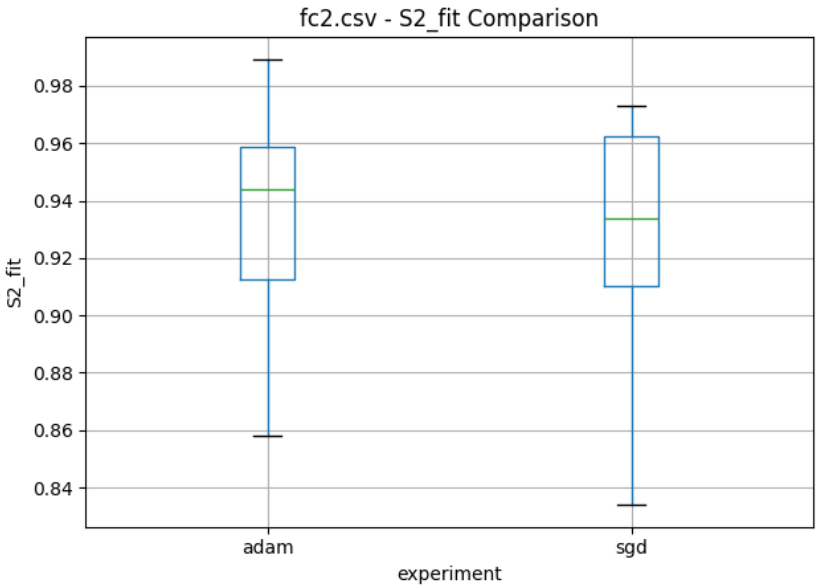


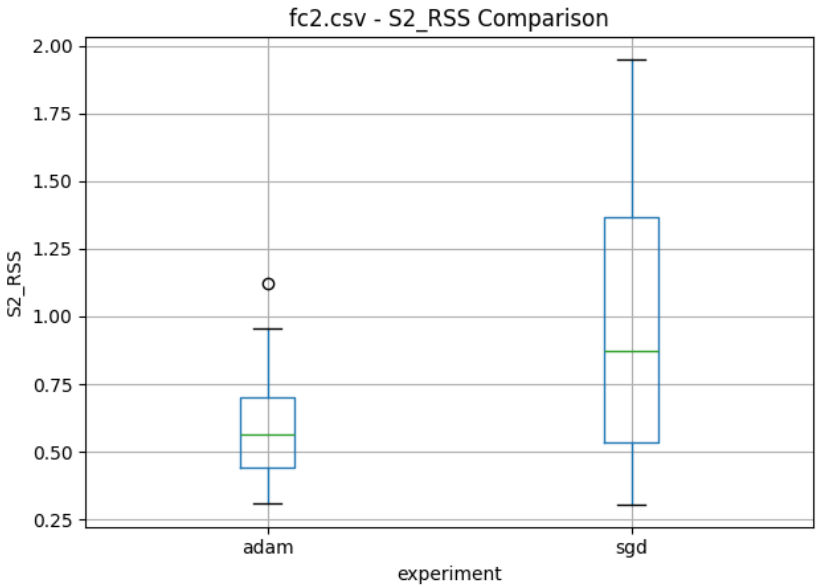
Multiple seed experiments





Multiple seed experiments

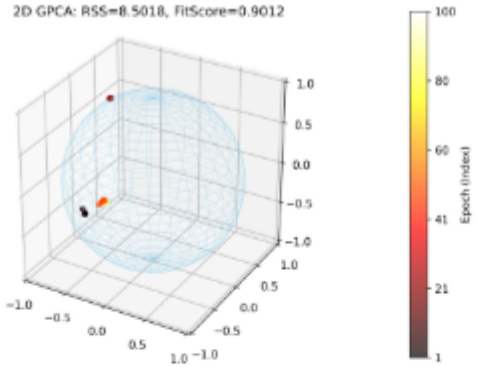


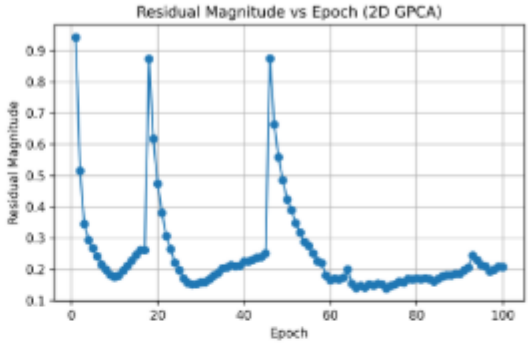


Multiple geodesic during training

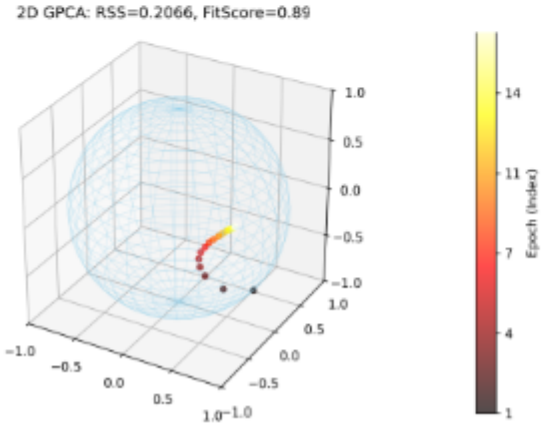
There were some cases that epochs follows three different geodesics.
After 18, 48 epochs, the projected eigenvector moves abruptly.
And has big resiual magnitude at that point.

Multiple geodesic during training

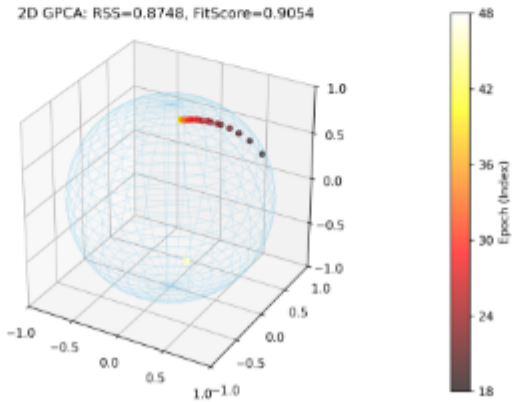




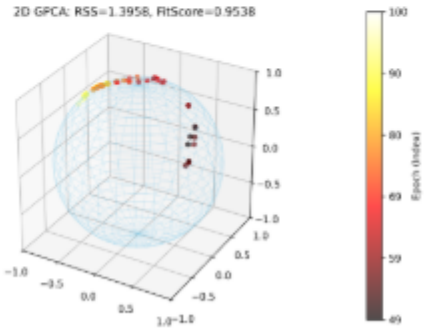
1-18



19-48

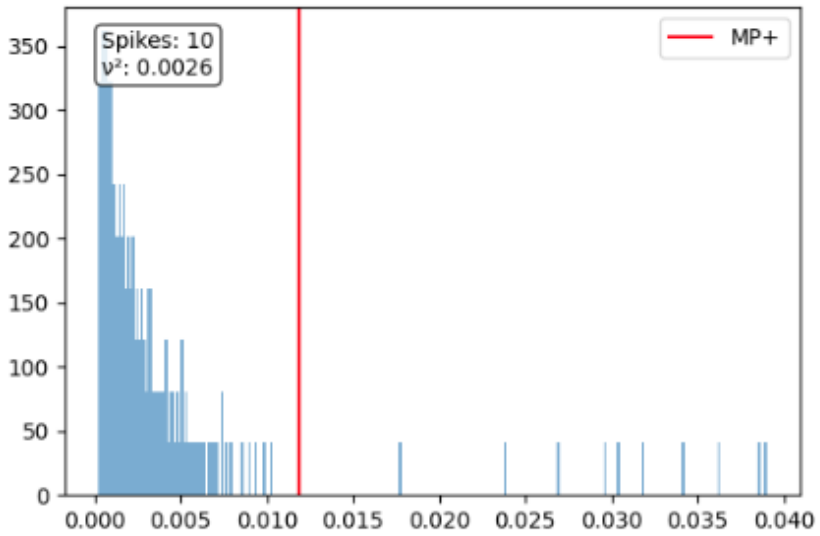


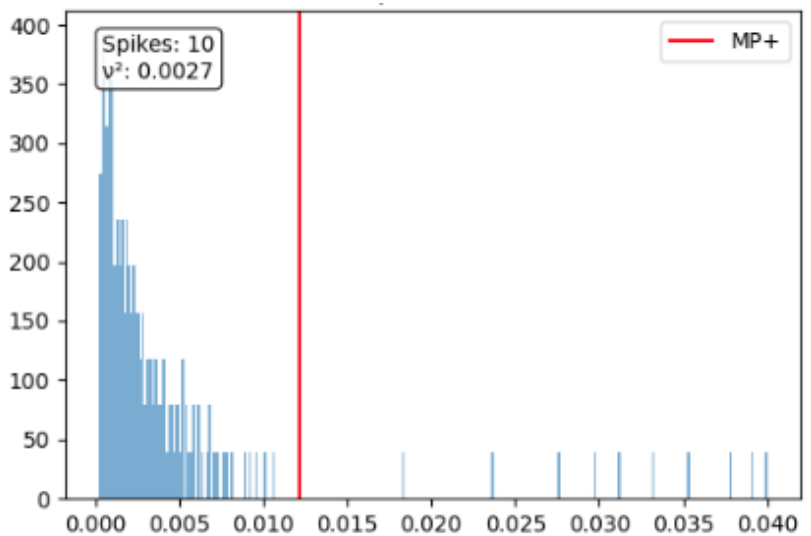
49-100



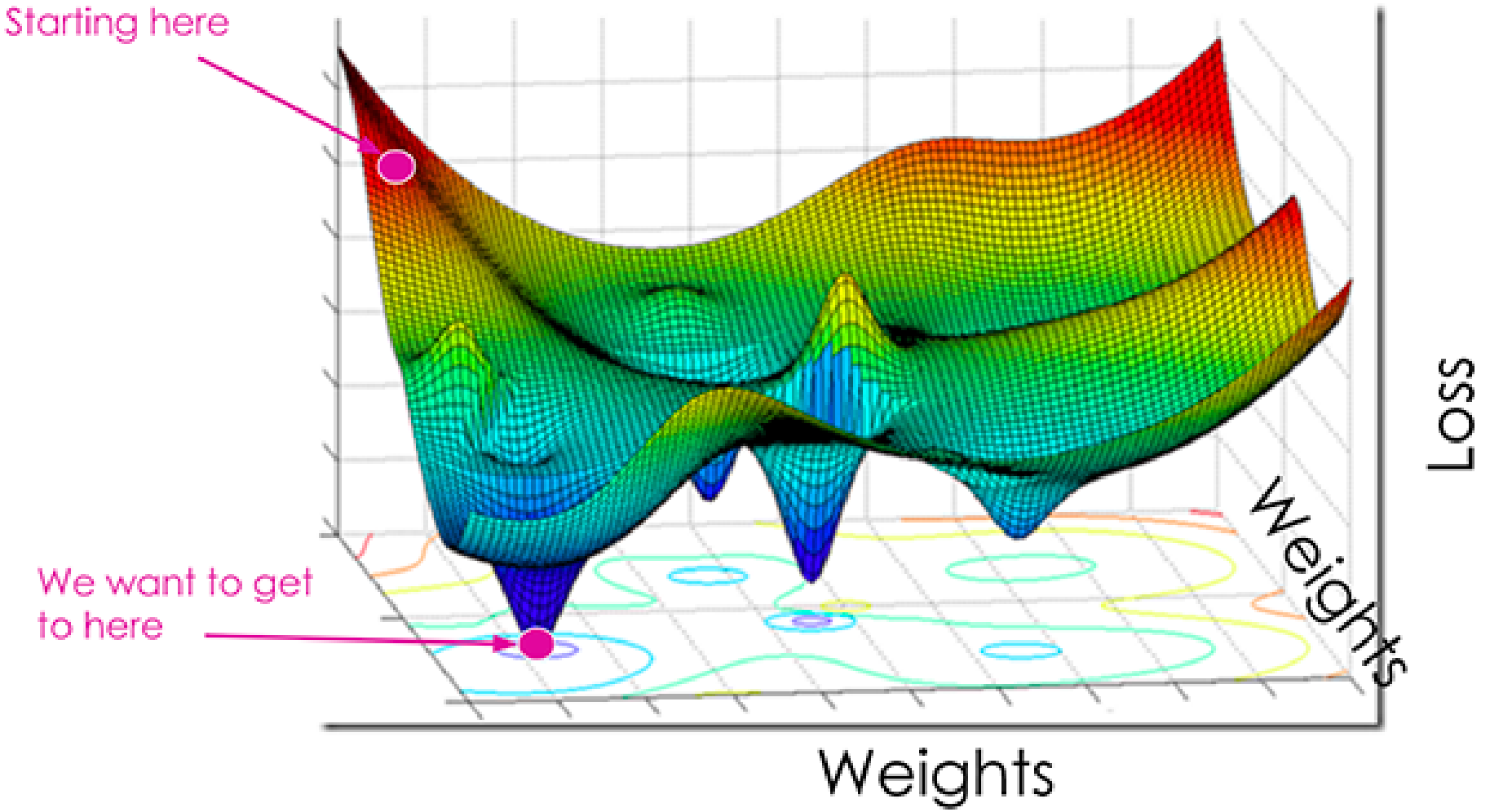
Multiple geodesic during training

This might be due to the leading eigenvalue crossing.
That the leading eigenvalue, vector changes.
This phenomena sometimes happened when the learning rate(step size) is big.





Multiple geodesic during training



Perturbation Studies

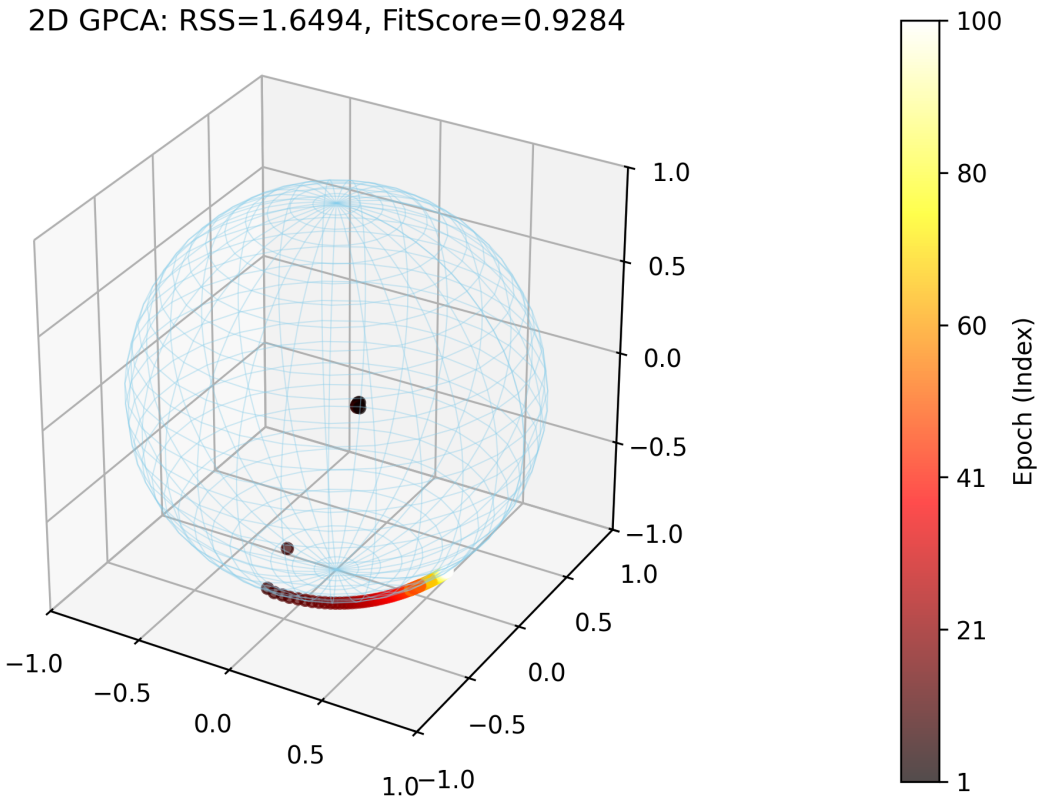
To understand how eigenvector trajectories react under perturbations, we compare:

- **Data shift at epoch 50:** Swap to a disjoint set of classes mid-training.

Varying training data after 50 epochs

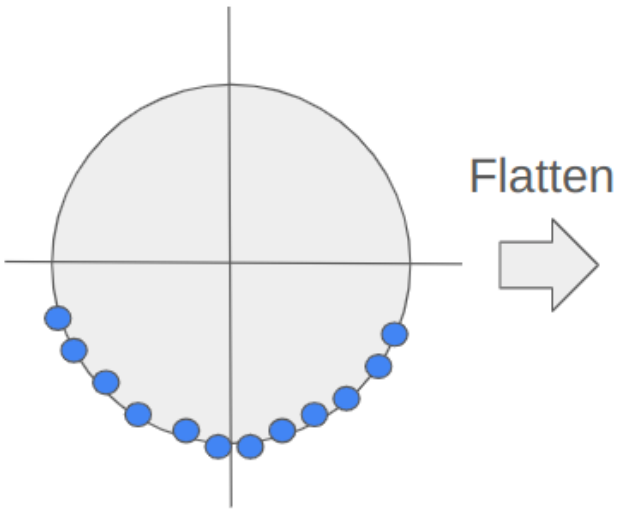
For first 50 epochs, train the model using the airplane, automobile, bird, cat and deer images.
After 50 epochs, train the model with dog, frog, horse, ship and truck images.
Will the leading eigenvector's trajectory follow the geodesic still?

Varying training data after 50 epochs

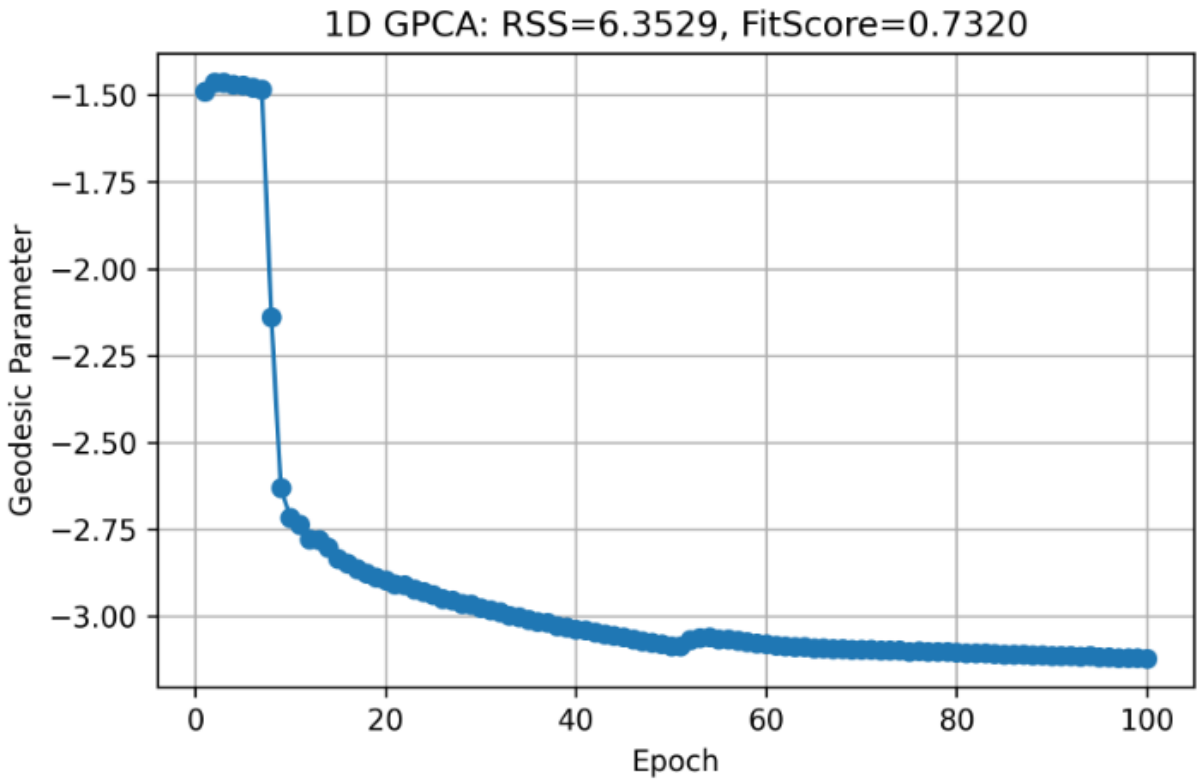


Varying training data after 50 epochs

Direction changes during 51 53 epochs, but after that, it goes to the similar direction as before.
Still, leading eigenvectors are on the geodesic line.

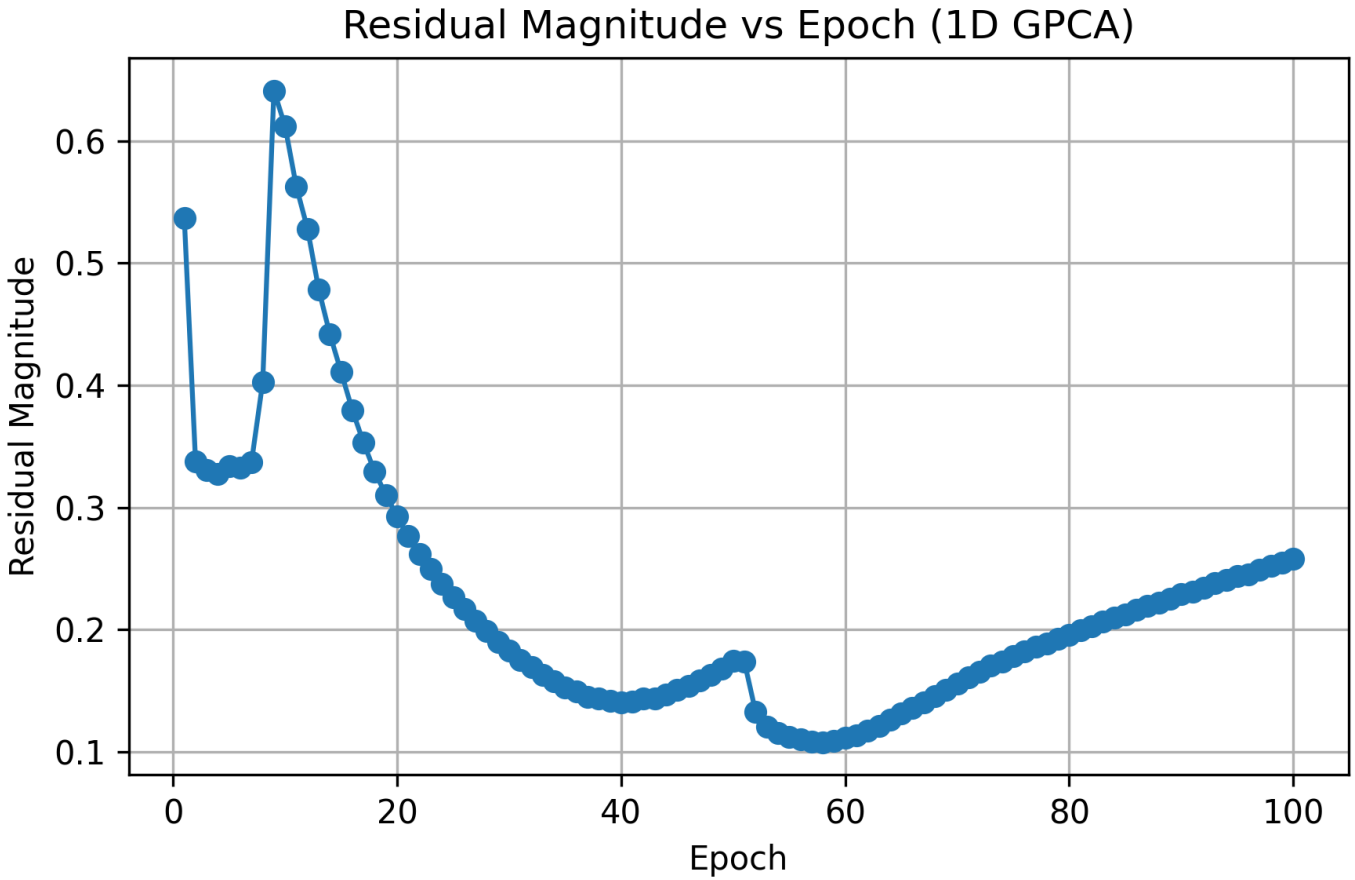


Flatten



Varying training data after 50 epochs

Since the residual near epoch 50 is rather small, we can say that even if we change training data, it still follows the geodesic line.
So whatever the data is, the weight matrix tend to follow geodesic line.



Future works

Since we found that the eigenvectors tend to follow geodesic when we train it, we can choose shrinkage point upon the geodesic.

References

- Mahoney, Michael, and Charles Martin. "Traditional and heavy tailed self regularization in neural network models." International Conference on Machine Learning. PMLR, 2019.
- Goldberg, Lisa R., and Alec N. Kercheval. "James–Stein for the leading eigenvector." Proceedings of the National Academy of Sciences 120.2 (2023): e2207046120.