PCA of High Dimensional Random Walks with Comparison to Neural Network Training

Joseph Antognini*, Jascha Sohl-Dickstein

*Work done as part of Google AI Residency, g.co/airesidency

Summary

- Deep neural networks (NNs) are high dimensional objects which makes visualizing the training process difficult. One approach some authors have proposed is to perform principal components analysis (PCA) on the trajectory of the parameters of the NNs and visualize the projection onto the lowest PCA dimensions (e.g., Lorch 2016; Lipton 2016; Li et al. 2018). When projected onto the lowest PCA components the trajectory appears smooth and contains a large fraction of the variance (typically over 80% in the first two PCA components).

- We show that when PCA is applied to a random walk in the limit of an infinite number of dimensions:
  - Approximately 60% of the variance is in the first PCA component; 80% of the variance is in the first two PCA components.
  - The projection of a random walk onto any two PCA components is a Lissajous curve.
  - These results are *independent* of the noise distribution.
  - These results also apply to a random walk with momentum
  - These results also apply to a random walk in a quadratic well (initialized near the origin) in the early stages of the trajectory.

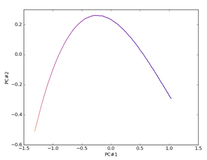
- We perform PCA on the parameter trajectories of a linear model trained on wCIFAR-10, a small NN trained on MNIST, and ResNet-50 trained on ImageNet and find that the PCA projected trajectories of the parameters closely resemble PCA-projected random walks, especially in the early stages of training.

Motivation

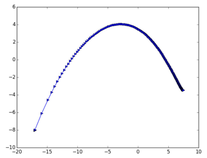
- NN training is a high dimensional stochastic process. This makes understanding and visualizing the training procedure difficult.

- Several authors have visualized NN training by performing PCA on the training trajectory and projecting onto the first few components (e.g., Lorch 2016; Lipton 2016; Li et al. 2018). These authors have found:

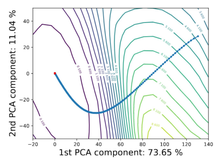
  - Although the trajectory when projected onto random dimensions looks similar to a random walk, the PCA projected trajectories are very smooth.
  - A large fraction of the variance is in the first few PCA components (between 80--90% in the first few).
  - This suggests an interpretation that although there is a stochastic component to NN training, in general the NN predominantly moves in a small number of dimensions, and does so in a consistent manner.



Lorch (2016)



Lipton (2016)



Li et al. (2016)

- *We consider what happens when you apply the same technique to a high dimensional random walk and find qualitatively the same results.* This suggests that this technique indicates that upon PCA, NN training can be effectively modeled as a biased random walk.

Core Result

Preliminaries

An n-dimensional random walk can be written as follows:

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \boldsymbol{\xi}_t, \qquad \boldsymbol{\xi}_t \sim \mathcal{P},$$

where $\mathbf{x}_0 = \mathbf{0}$ and $\mathcal{P}$ is an arbitrary probability distribution. We can also write this in matrix form:

$$\mathbf{SX} = \mathbf{R}, \qquad \mathbf{S} \equiv \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \ddots & \vdots \\ 0 & -1 & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix},$$

Performing PCA is equivalent to finding the eigenvalues and eignevectors of the following matrix:

$$\hat{\mathbf{X}}\hat{\mathbf{X}}^T = \mathbf{C}\mathbf{S}^{-1}\mathbf{R}\mathbf{R}^T\mathbf{S}^{-T}\mathbf{C}, \qquad \hat{\mathbf{X}} = \mathbf{C}\mathbf{X}, \quad \mathbf{C} \equiv \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T.$$

In the limit that $d \gg n$ we will have $\mathbf{R}\mathbf{R}^T \to \mathbf{I}$ because the off diagonal terms will be $\mathbb{E}[\xi_i]^2 = 0$ whereas the on diagonal terms will be $\mathbb{E}[\xi^2] = \sum_{i=0}^{d} \mathbb{V}[\xi_i] = 1$

Asymptotic convergence to circulant matrices

In the limit that $n \to \infty$ banded Toeplitz matrices become asymptotically convergent to circulant matrices. This implies that they have the same inverses and distribution of eigenvalues and eigenvectors as the corresponding circulant matrices.

The eigenvalues of a circulant matrix with entries $c_0, c_1, \ldots$ are

$$\lambda_{\text{circ},k} = c_0 + c_{n-1}\omega_k + c_{n-2}\omega_k^2 + \ldots + c_1\omega_k^{n-1},$$

Using this, we can derive the variance in each PCA component:

$$\lambda_{\hat{\mathbf{X}}\hat{\mathbf{X}}^T,k} = \frac{1}{2}\left[1 - \cos\left(\frac{\pi k}{n}\right)\right]^{-1}$$

This implies that the explained variance ratio of each component is:

$$\boxed{\rho_k = \frac{6}{\pi^2 k^2}}$$

Since the eigenvectors of circulant matrices are simply Fourier modes, we show that the projection onto a PCA component is given by:

$$\boxed{\mathbf{X}_{\text{PCA},k} = \sqrt{\frac{2\lambda_k}{n}}\cos\left(\frac{\pi k t}{n}\right)}$$

This implies that the projection of a high dimensional random walk onto PCA components is a Lissajous curve.

Generalizations

PCA variances of a random walk with momentum:

$$\begin{aligned} \mathbf{v}_t &= \gamma\mathbf{v}_{t-1} + \boldsymbol{\xi}_t \\ \mathbf{x}_t &= \mathbf{x}_{t-1} + \mathbf{v}_t. \end{aligned}$$

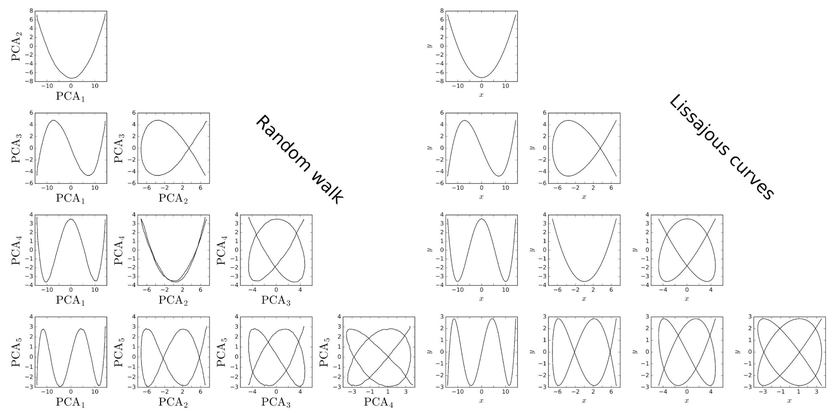
$$\lambda_k = \frac{1}{2}\left[1 + \gamma + \gamma^2 - (1+\gamma)^2\cos\left(\frac{\pi k}{n}\right) + \gamma\cos\left(\frac{2\pi k}{n}\right)\right]^{-1}$$

PCA variances of an Ornstein-Uhlenbeck process: $\mathbf{x}_t = (1-\alpha)\mathbf{x}_{t-1} + \boldsymbol{\xi}_t$,

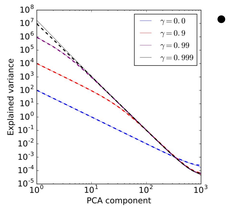
$$\lambda_{\text{OU},k} = \left[1 + (1-\alpha)^2 - 2(1-\alpha)\cos\left(\frac{2\pi k}{n}\right)\right]^{-1} \simeq \left[\frac{4\pi^2 k^2(1-\alpha)}{n^2} + \alpha^2\right]^{-1}$$

Experiments

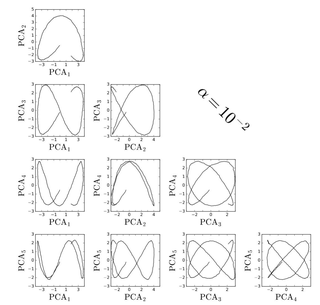
- High dimensional random walks projected onto PCA components are Lissajous curves:



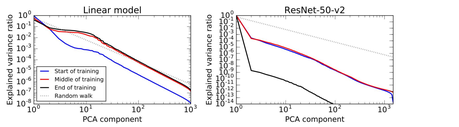
- The explained variance of each PCA component we observe closely matches our predictions for both a high dimensional random walk and a high dimensional random walk with momentum:

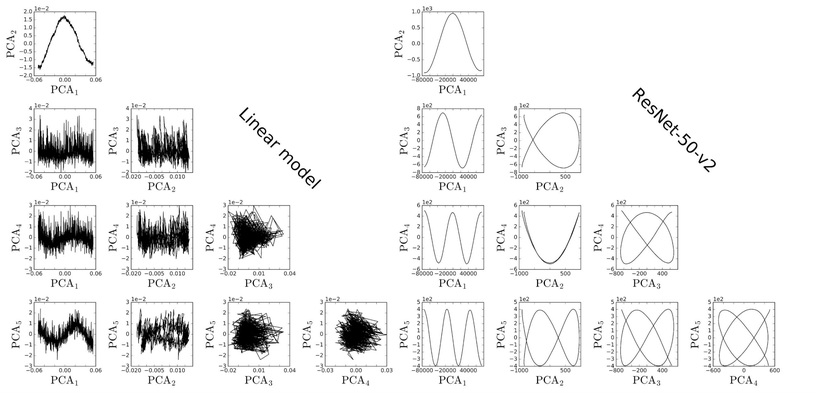


- The PCA projected trajectories of an Ornstein-Uhlenbeck process resemble perturbed Lissajous curves:



- PCA projections of the training trajectory of a large neural network resemble Lissajous curves
  - Trained a linear model on CIFAR-10 for 10,000 steps, collected parameters from the middle 1000 steps
  - Trained ResNet-50-v2 on ImageNet for 150,000 steps, collected parameters from the middle 15,000 steps
- Explained variance ratio distribution is comparable to a random walk early in training with the exception of the first component.





References

- Lorch,E., Visualizing deep network training trajectories with pca. In *ICML Workshop on Visualization for Deep Learning*, 2016
- Lipton, Z., Stuck in a what? Adventures in weight space. *arXiv preprint arXiv:1602.07320*, 2016
- Li, H., Xu, Z., Taylor, G., and Goldstein, T., Visualizing the loss landscape of neural nets. In *International Conference on Learning Representations*, 2018