Neural Network Training Dynamics: Towards Making Sense of a Confusing Mess

Razvan Ciuca, Université de Montréal

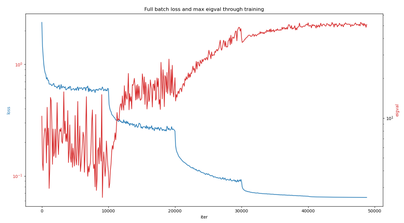
razvan.ciuca@umontreal.ca

Université de Montréal

The Training and Model Setup

- CIFAR10 no data augmentation.
- SGD with mometum, lr=1e-1, momentum=0.97, clipped gradient, batch size 512.
- Very thin ResNet with LayerNorm, 26000 parameters in total, final top-1 accuracy is 80%, so not a totally unrealistic model.

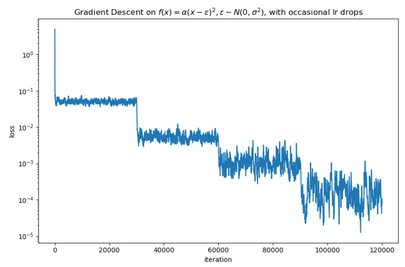
Puzzling Features of the Training graph



- The max eigenvalue (red curve) increases as we lower the learning rate: edge of stability phenomenon.
- Cliff-like decreases in loss at every learning rate decrease.
- What's going on?! How could **reducing the training speed** somehow send us down a sharp cliff in training loss?

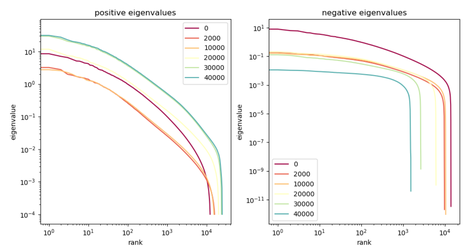
A Quadratic Stochastic Model

- A simple stochastic model: do GD on $f(x) = \lambda(x+\epsilon)^2$, where $\epsilon \sim N(0, \sigma^2)$ is a stochastic term that shifts the minimum randomly.



- We can analytically obtain the minimum achievable at a given noise level and learning rate:$E[\lambda x^2] = \frac{\alpha\lambda^2\sigma^2}{1-\alpha\lambda}$

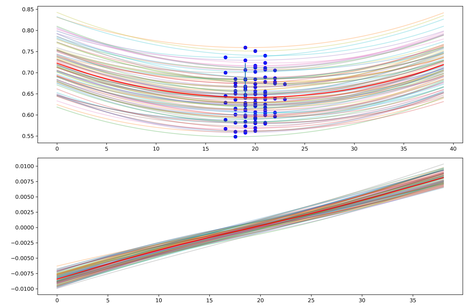
The Hessian Spectrum Through Training



- Positive Values maintain a rough power law that shifts upwards.
- Negative values have a strong cliff-drop effect. Notice that we still have lots of negative directions at end of training!

Testing The Quadratic Approximation

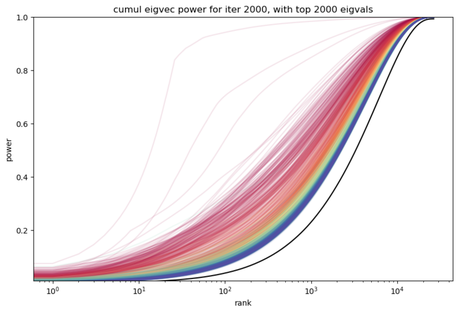
- Let's test the accuracy of the toy quadratic model by plotting the losses in a given eigendirection for multiple batches.



- Result: The toy model is remarkably accurate! Different batches have essentially perfectly quadratic losses (in this direction), with normally-distributed minimum locations.

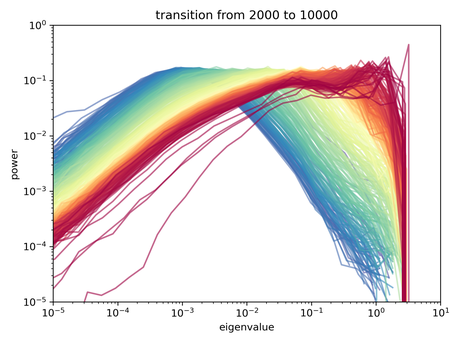
High λ eigenvectors are sharply distributed

- For a given iteration in training, let's plot the cumulative power in the sorted components of the top eigenvectors.



- Low λ vectors are mostly normally distributed, but high λ vectors are much more concentrated in a few directions.

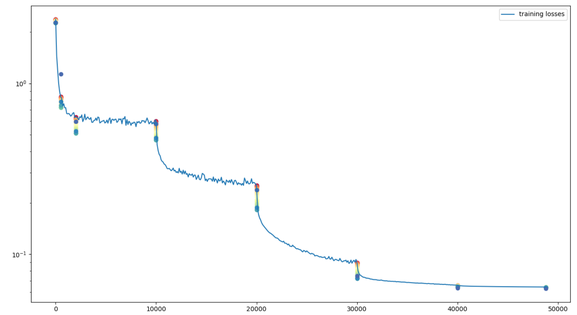
Eigenvectors are changing throughout training



- The eigenvectors are changing, but they're mostly rotating into other vectors of similar eigenvalue! The high λ subspace is remarkably consistent, but the low λ directions shift into each other a lot.

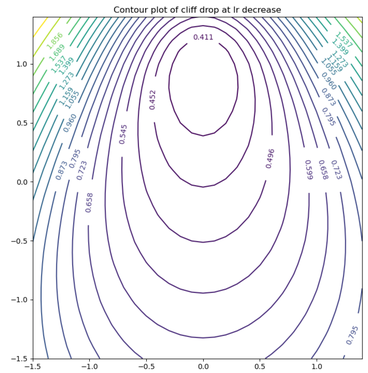
Is the quadratic approximation enough? No.

- Plotting the local quadratic minimum through training shows that only part of the loss decreases can be explained purely quadratically.

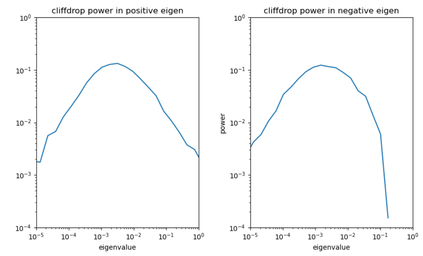


Zooming in on the "cliff drops"

- The neighborhood of the cliff drop at iteration 10000 has a peculiar "valley" structure:



- The cliff drop is mostly directed in the low λ subspace. High λ directions don't change much:



Summary And Implications

- The Stochastic Quadratic model is part of the answer of what's going on.
- High eigenvalues seem to be "gating" access to the correct low-eigenvalue directions.
- Eigenvectors with similar eigenvalues "diffuse into each other" over time, little mixing between scales occurs.