Qualitatively Characterizing Neural Network Optimization Problems

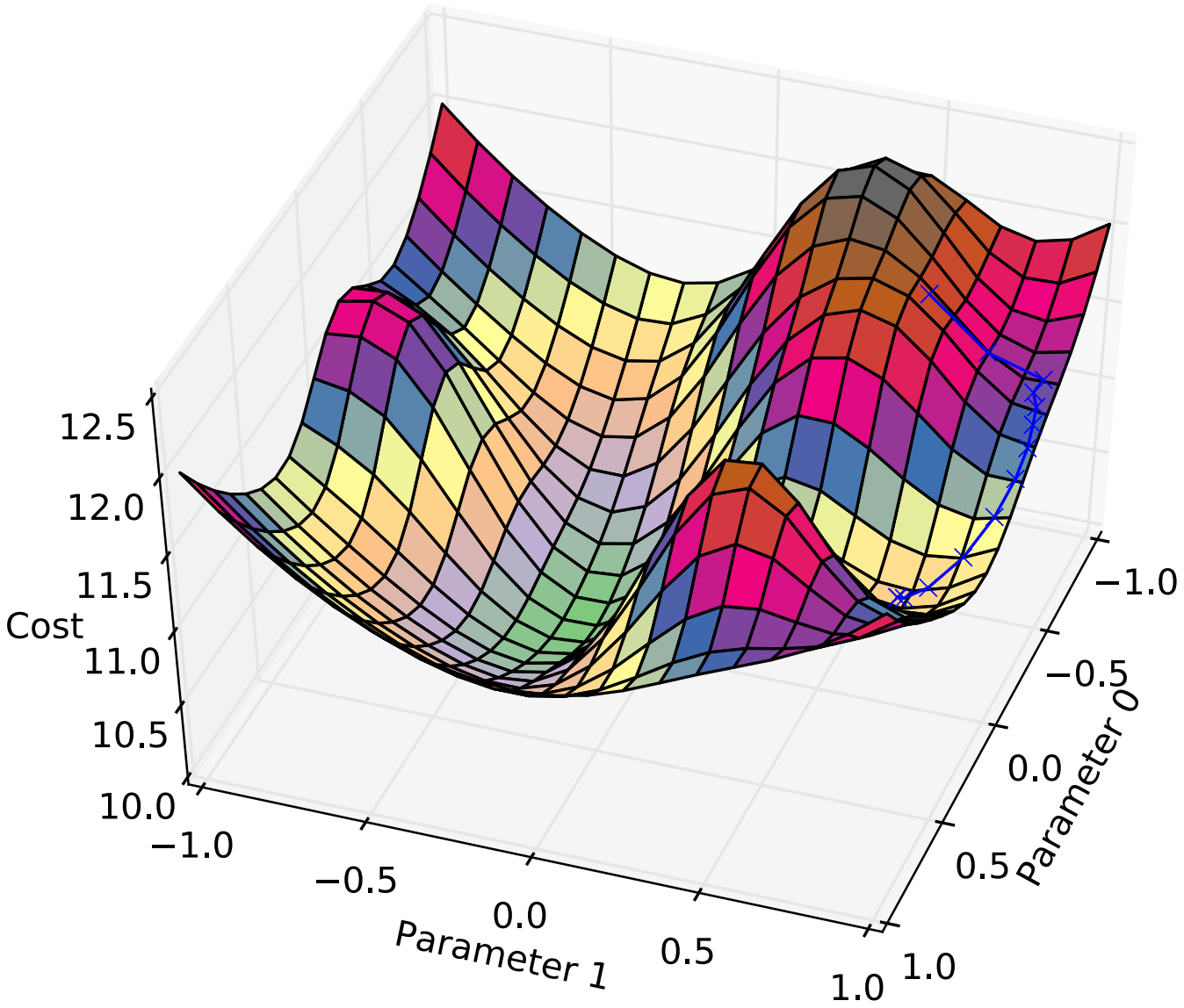
Ian Goodfellow

Oriol Vinyals

Andrew Saxe

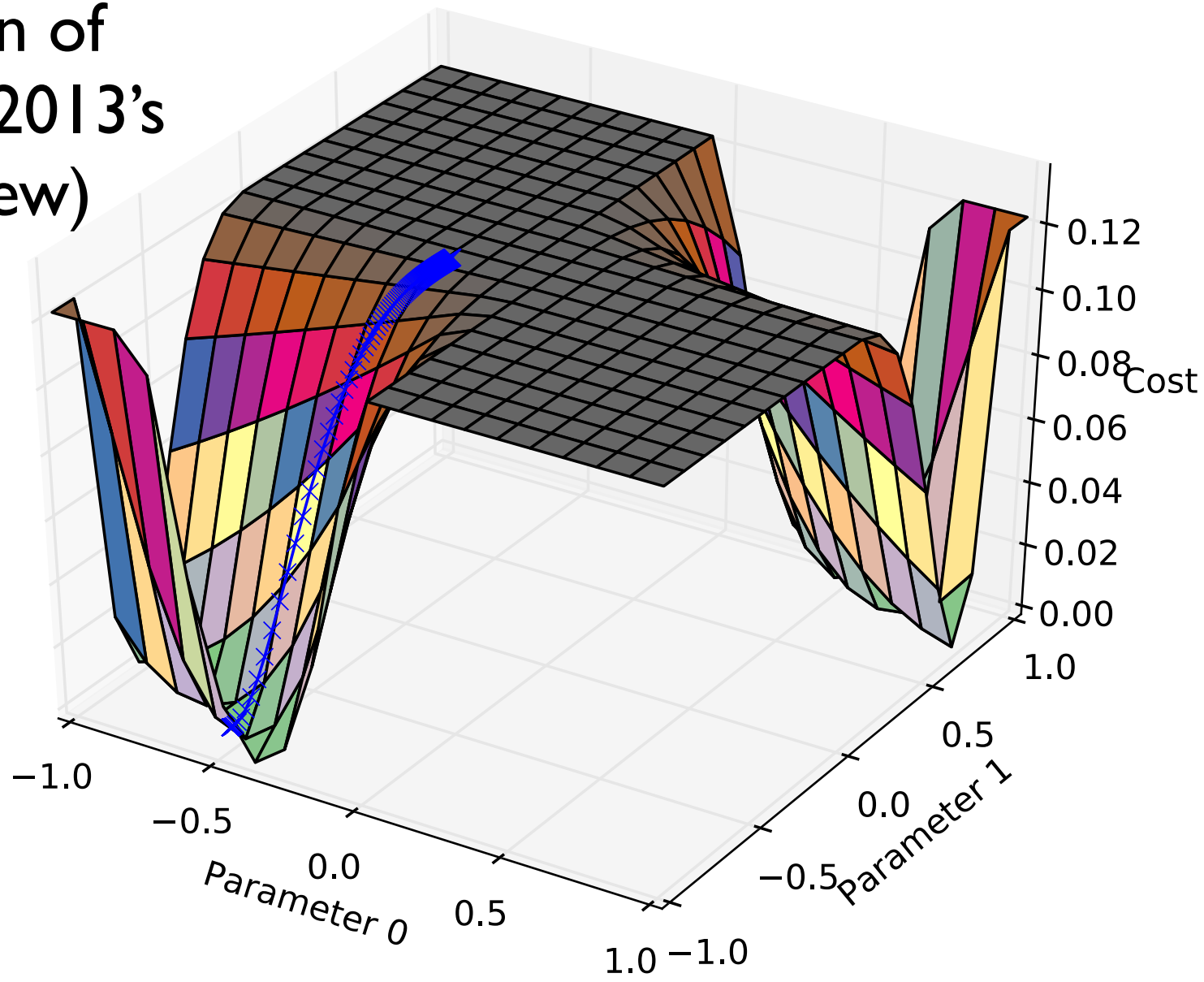
Google™

Traditional view of NN training



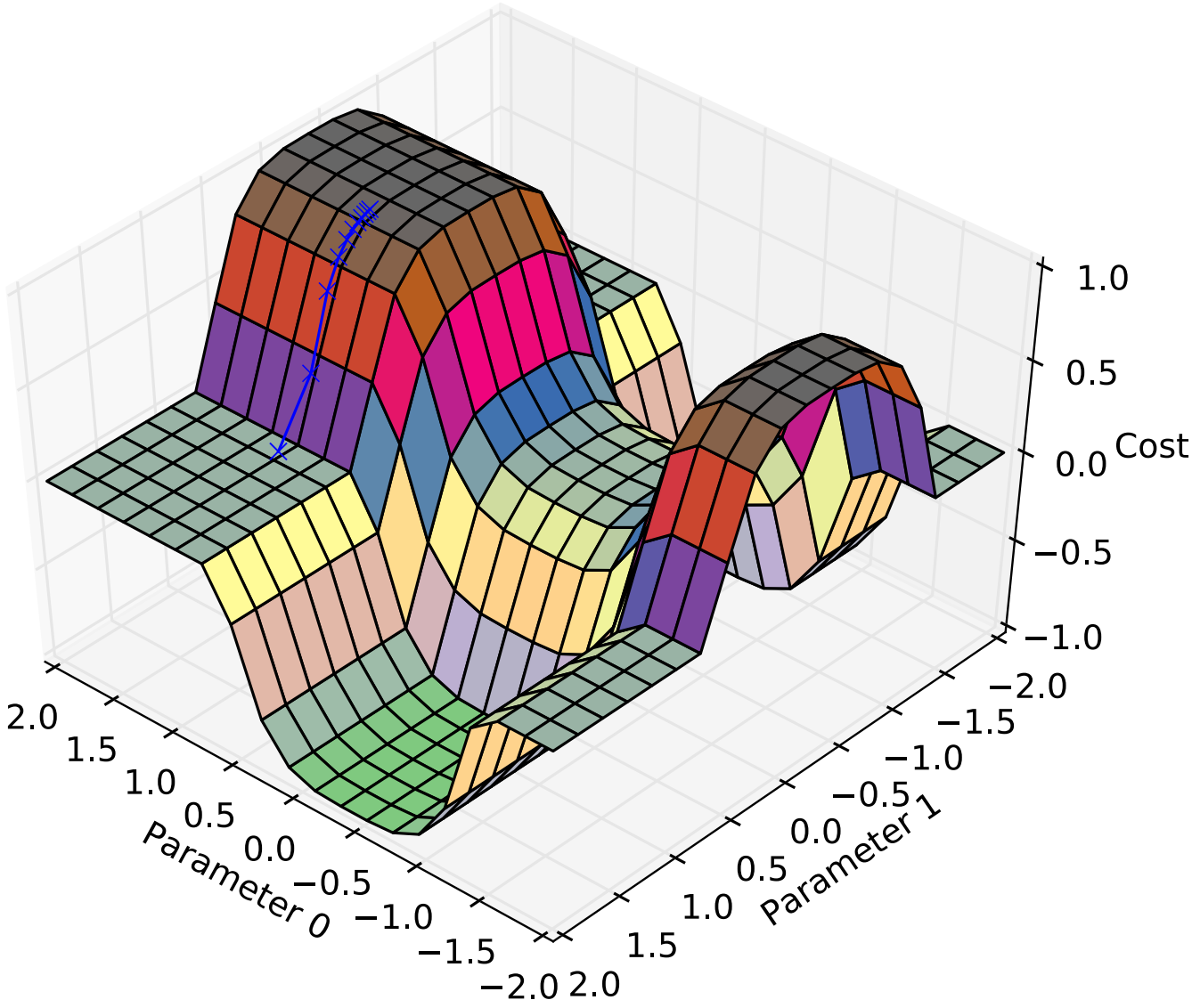
Factored linear view

(Cartoon of Saxe et al 2013's worldview)



Attractive saddle point view

(Cartoon of Dauphin et al 2014's worldview)



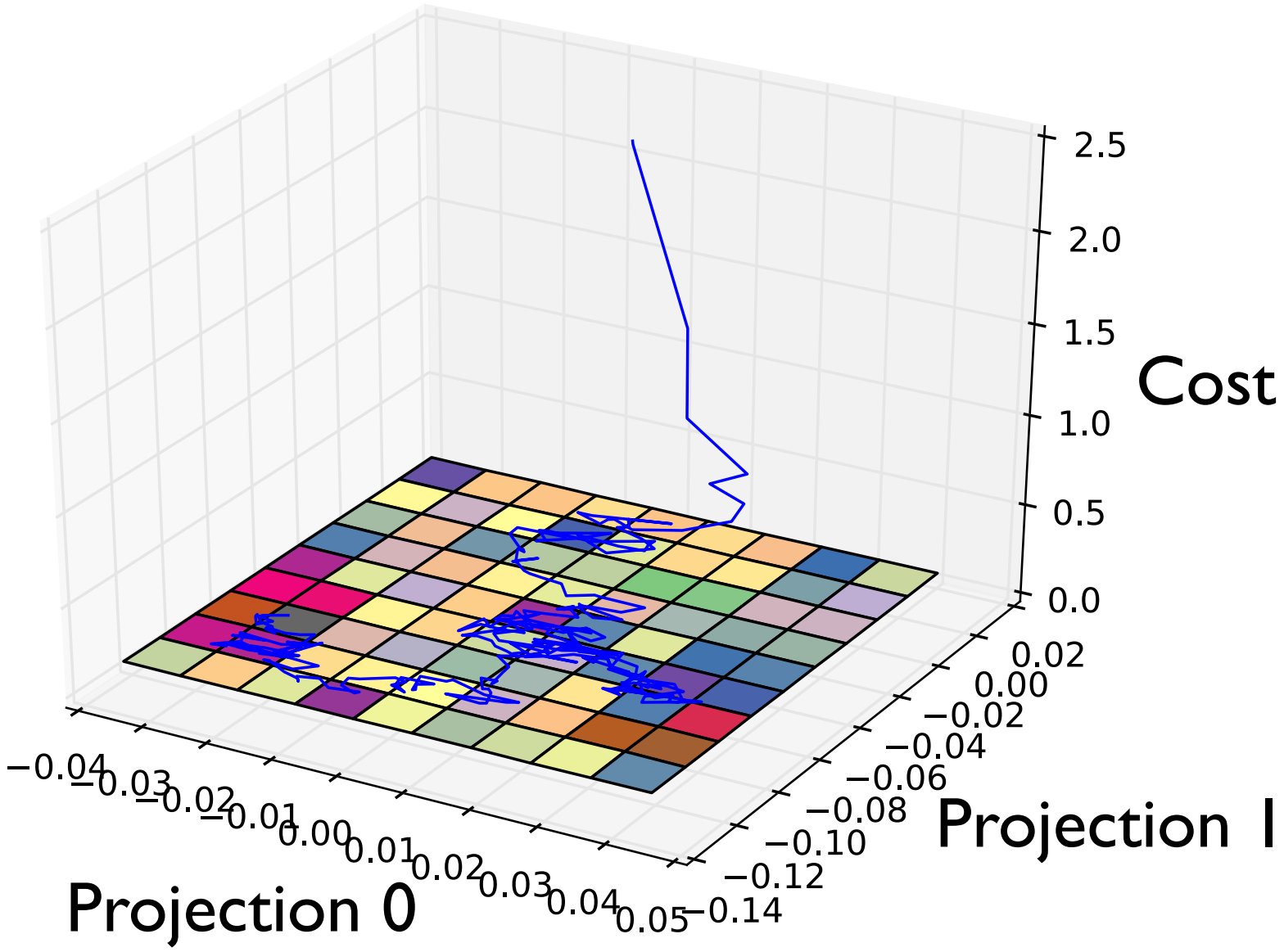
Questions

- Does SGD get stuck in local minima?
- Does SGD get stuck on saddle points?
- Does SGD wind around numerous bumpy obstacles?
- Does SGD thread a twisting canyon?

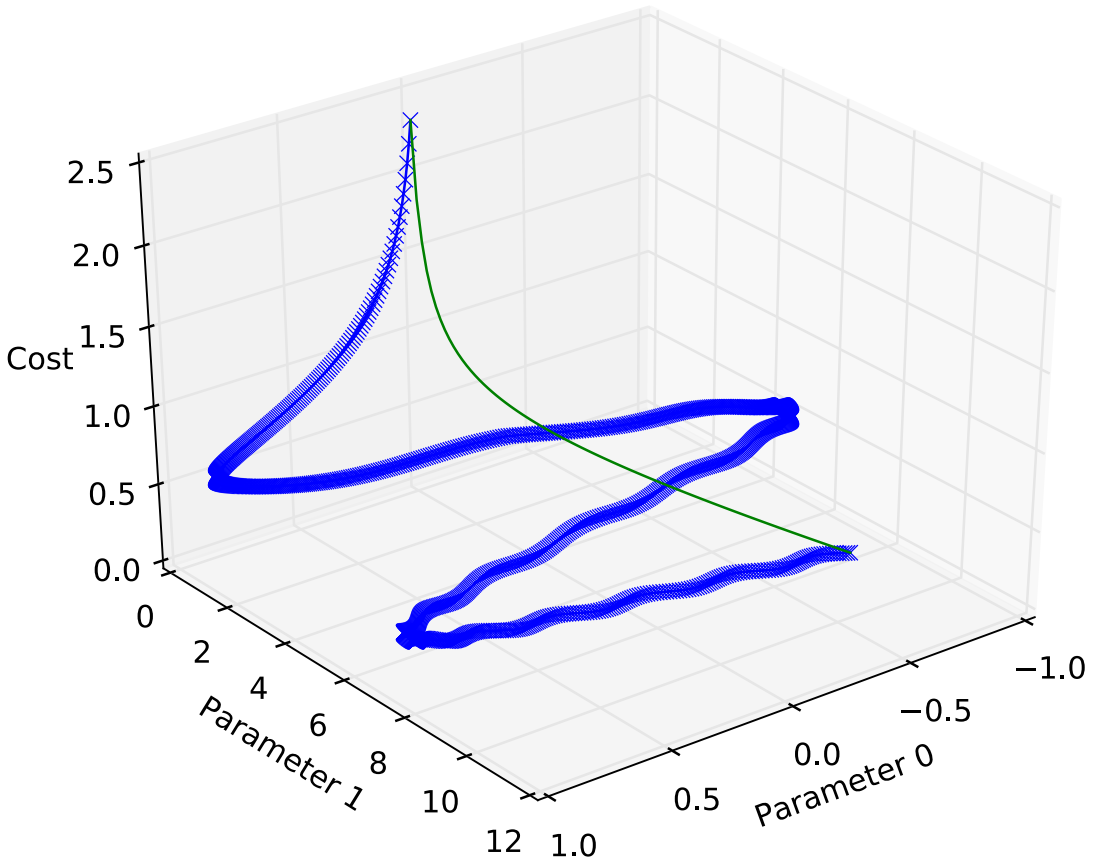
History written by the winners

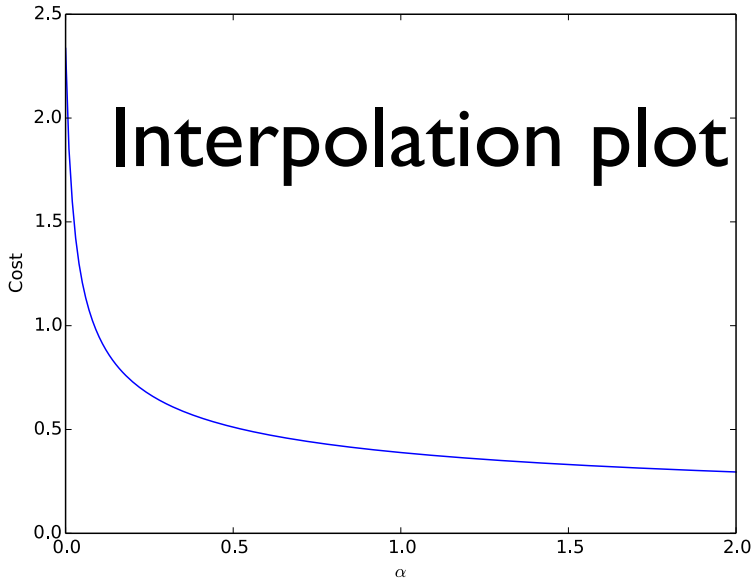
- Visualize trajectories of (near) SOTA results
- Selection bias: looking at success
- Failure is interesting, but hard to attribute to optimization
- Careful with interpretation
  - SGD never encounters X?
  - SGD fails if it encounters X?

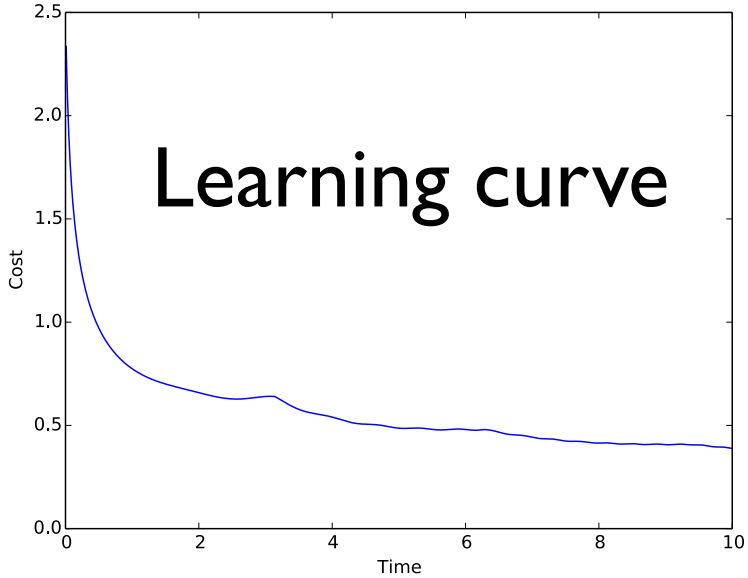
2-D subspace visualization



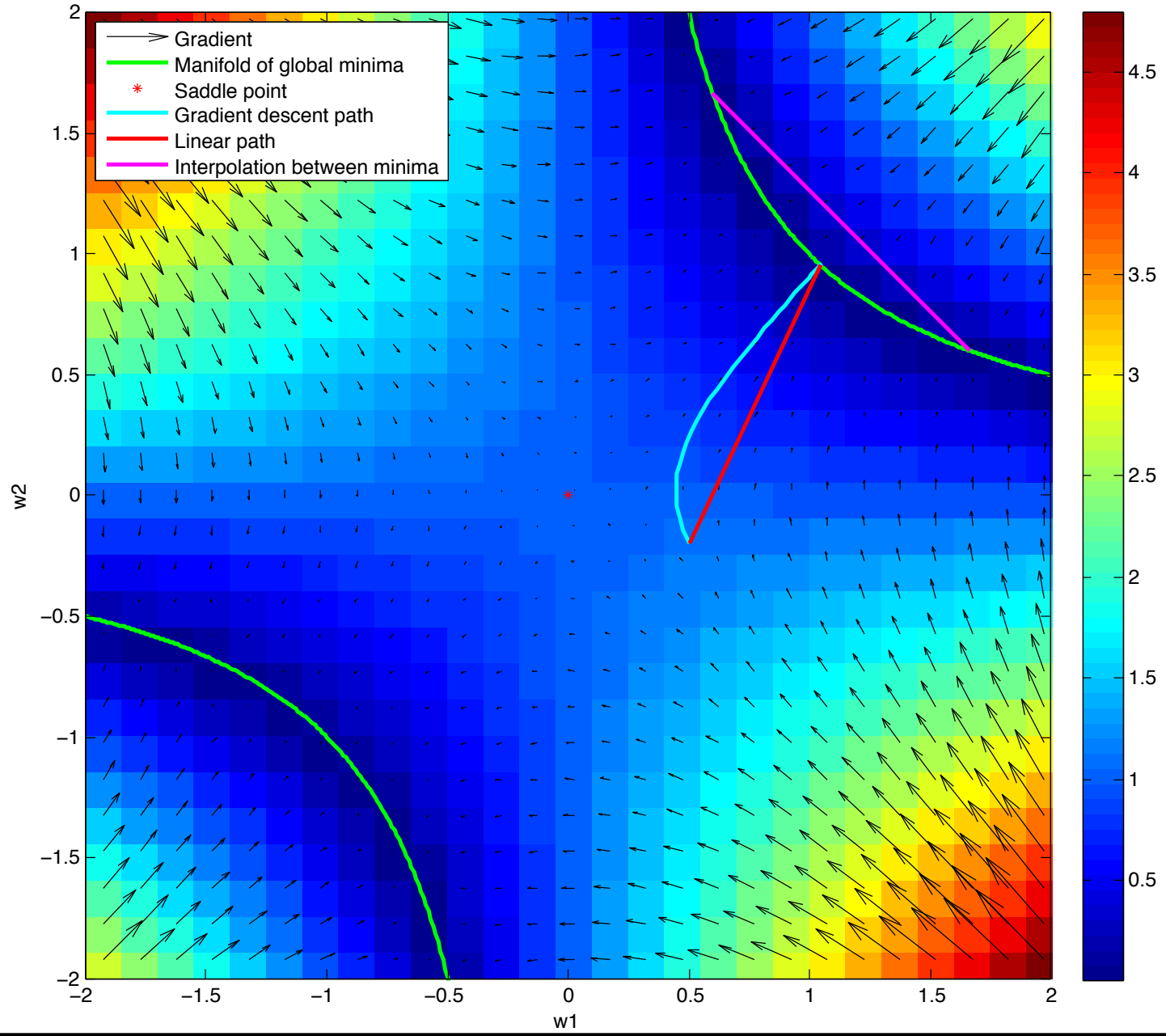
A special 1-D subspace



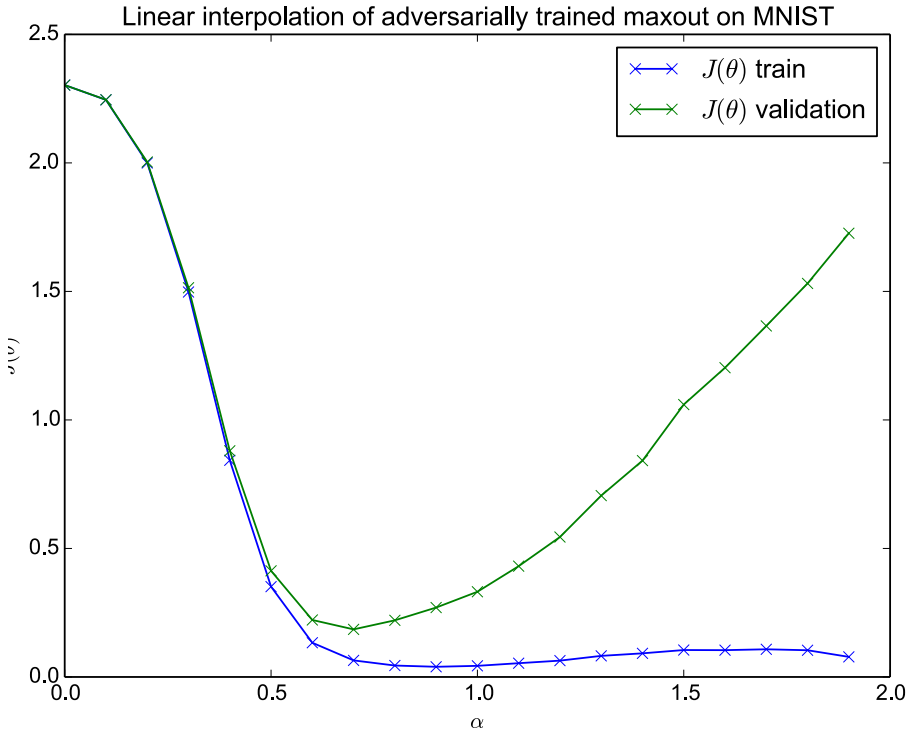


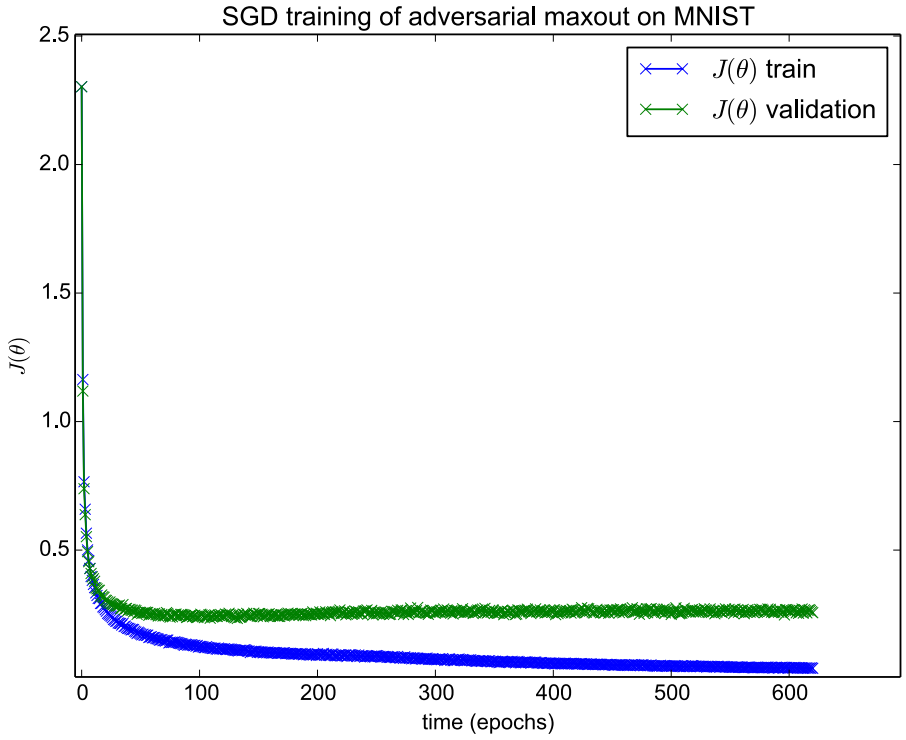


2-parameter deep linear model

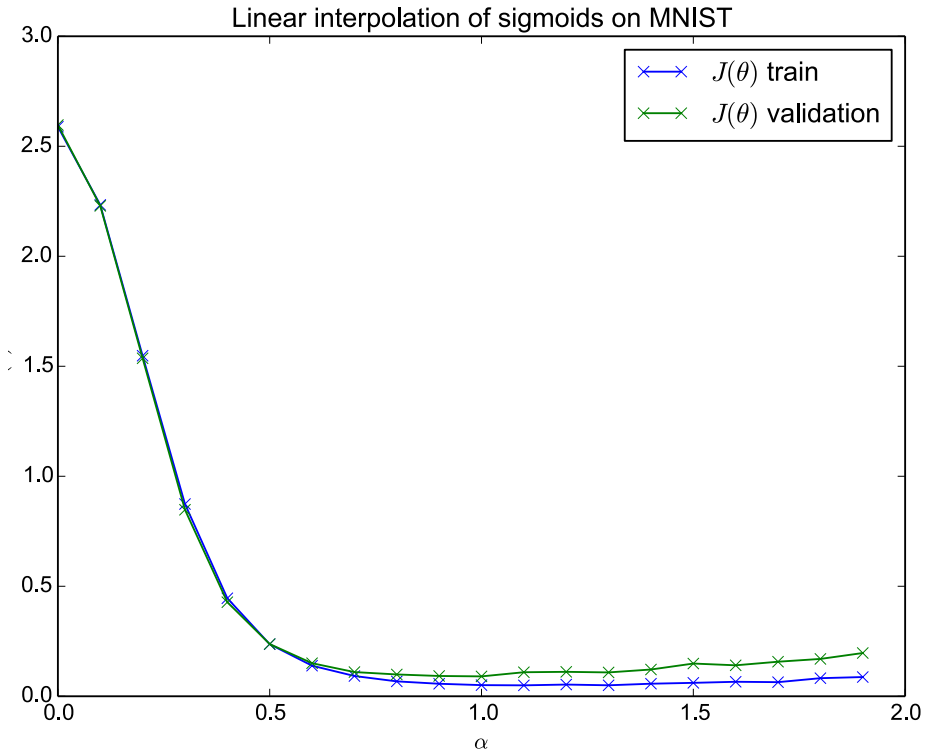


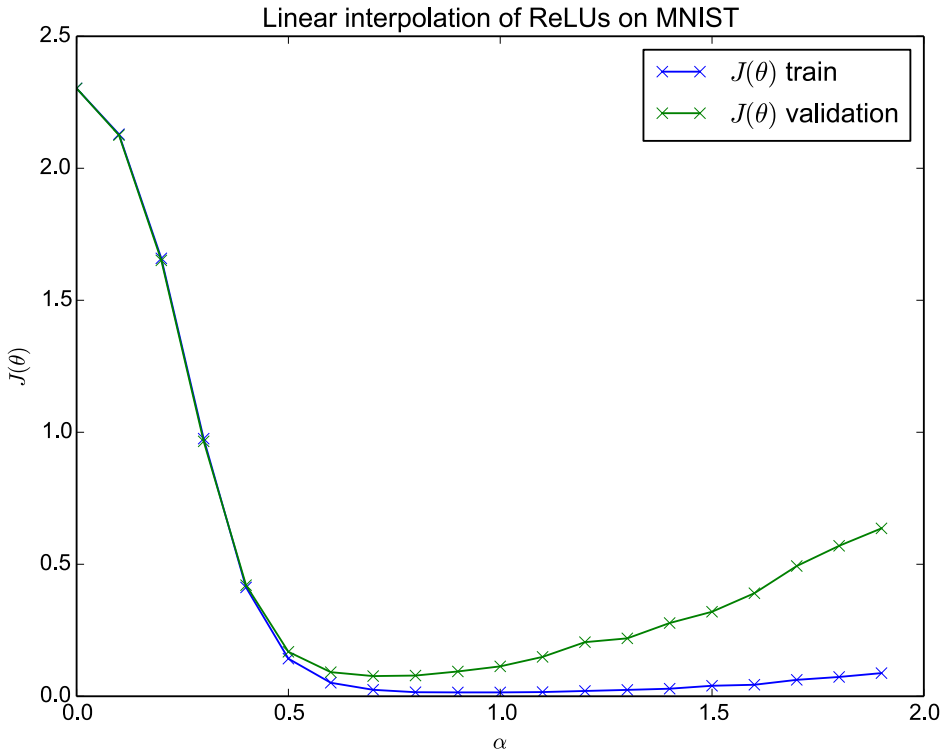
Maxout / MNIST experiment



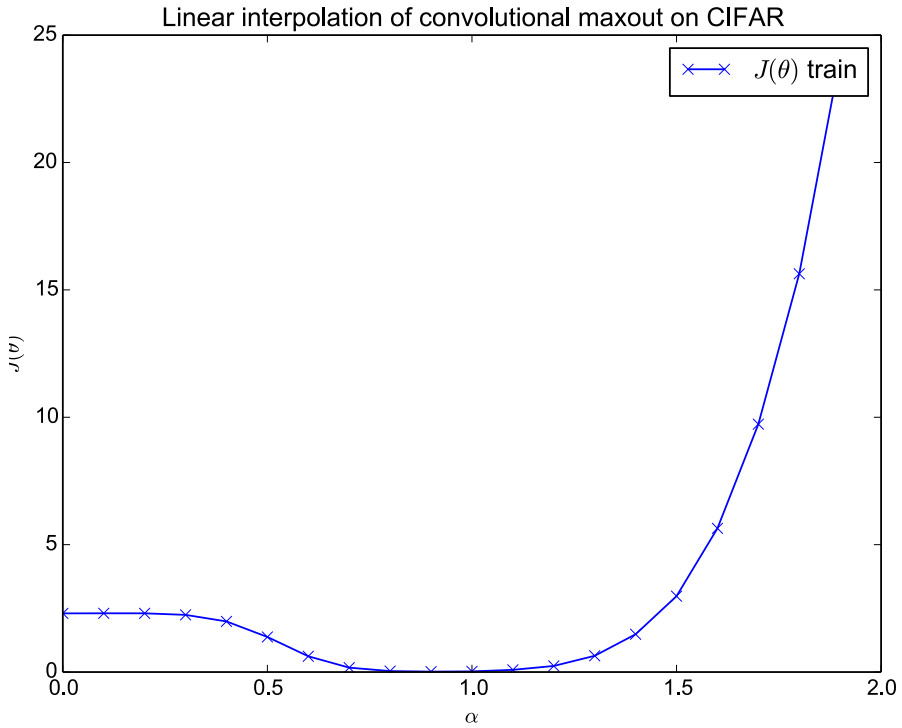


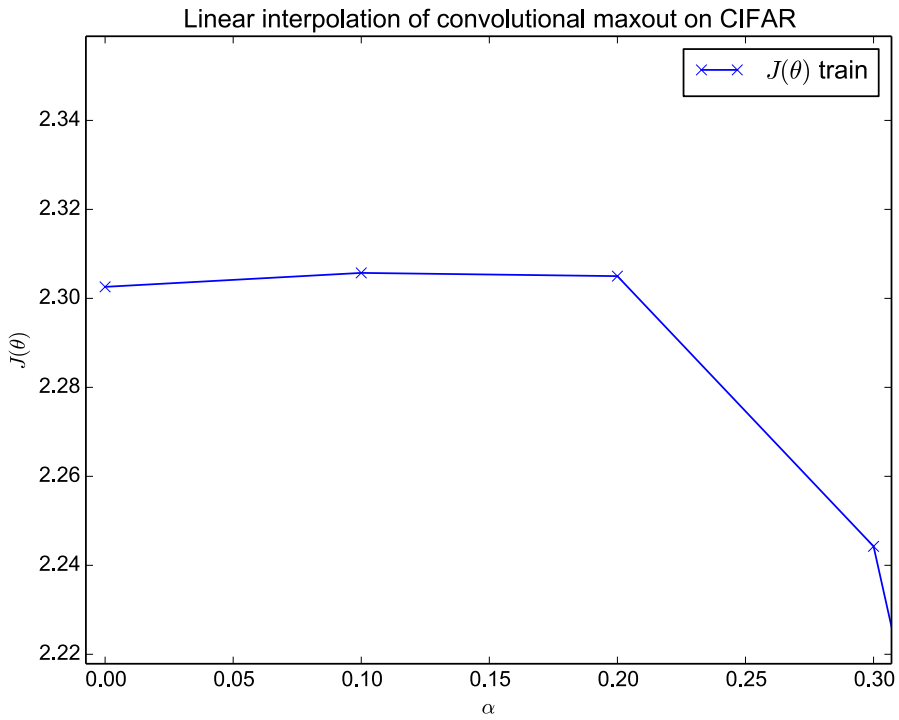
Other activation functions





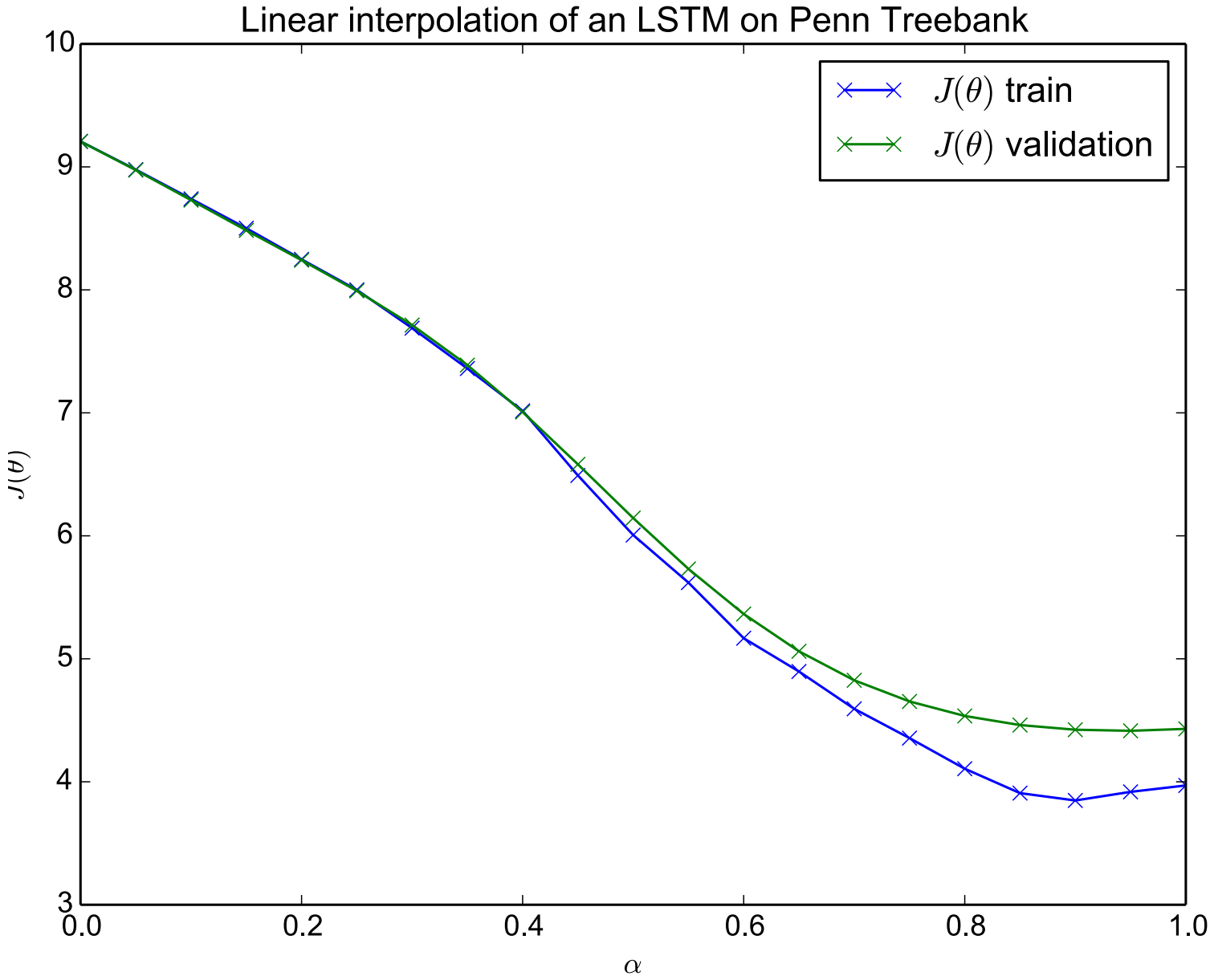
Convolutional network



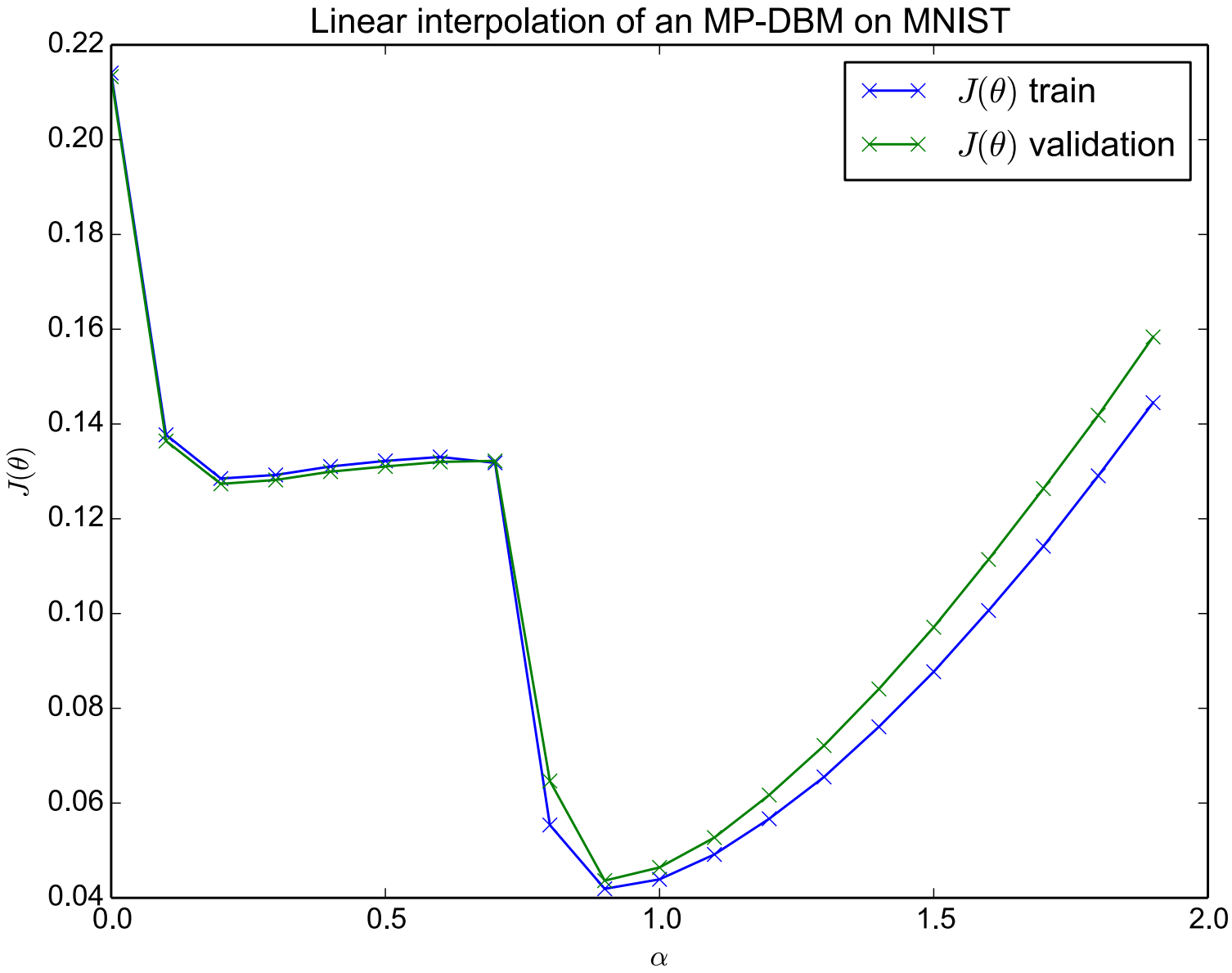


A small barrier

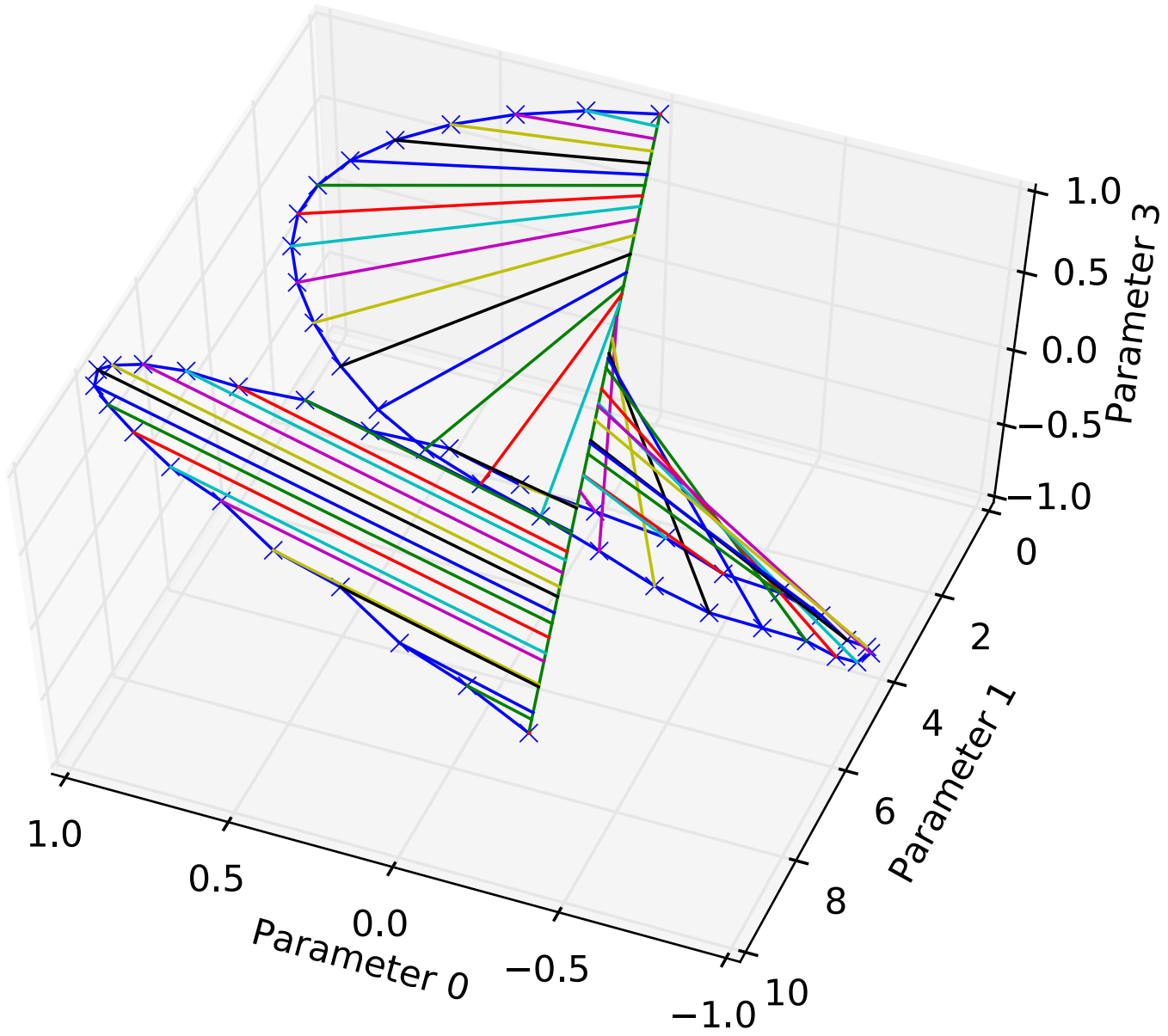
LSTM


Linear interpolation of an LSTM on Penn Treebank
$J(\theta)$ train
$J(\theta)$ validation
$J(\theta)$
α
10
9
8
7
6
5
4
3
0.0
0.2
0.4
0.6
0.8
1.0

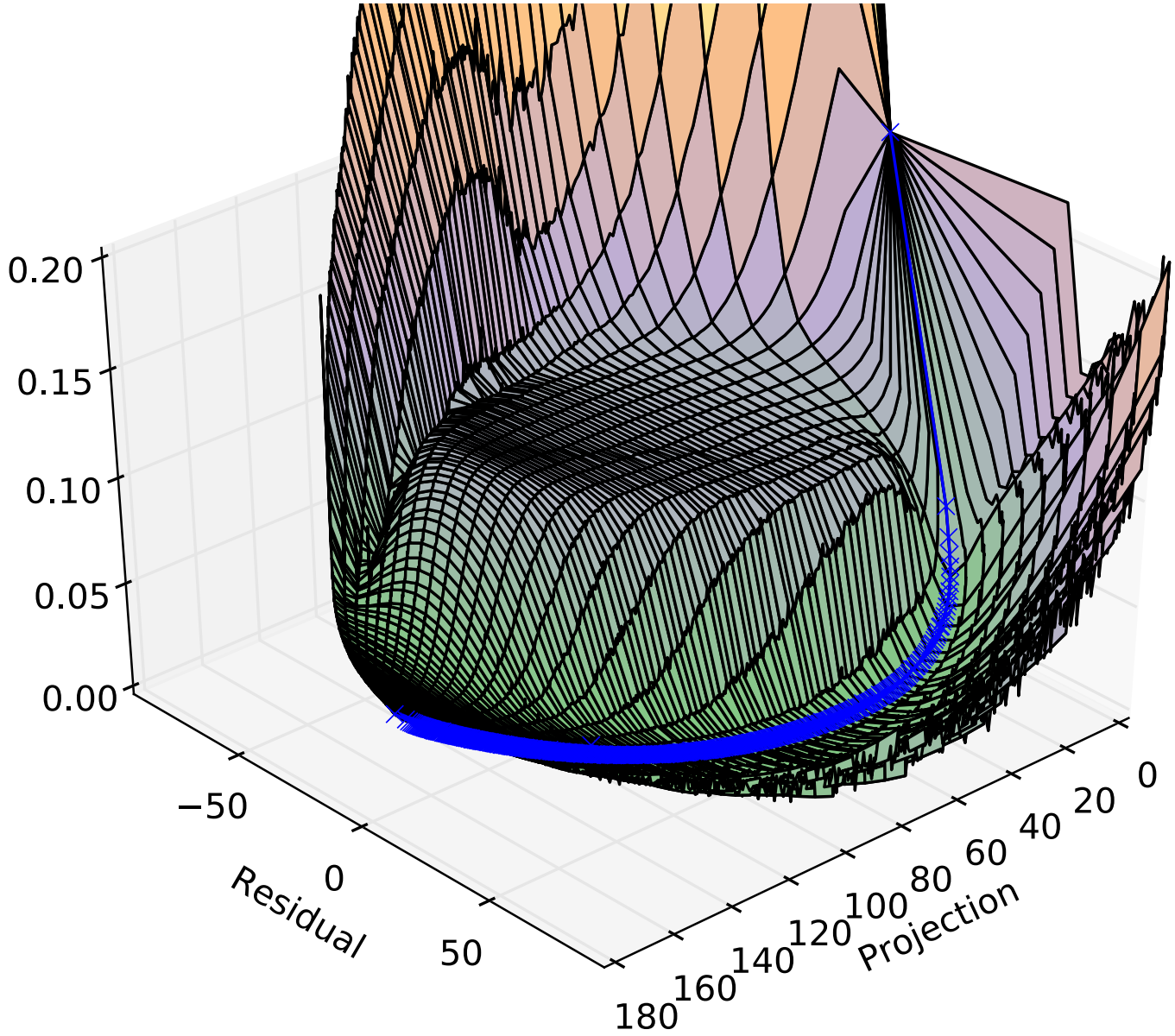
MP-DBM


Linear interpolation of an MP-DBM on MNIST
$J(\theta)$ train
$J(\theta)$ validation
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
$J(\theta)$
0.0
0.5
1.0
1.5
2.0
α

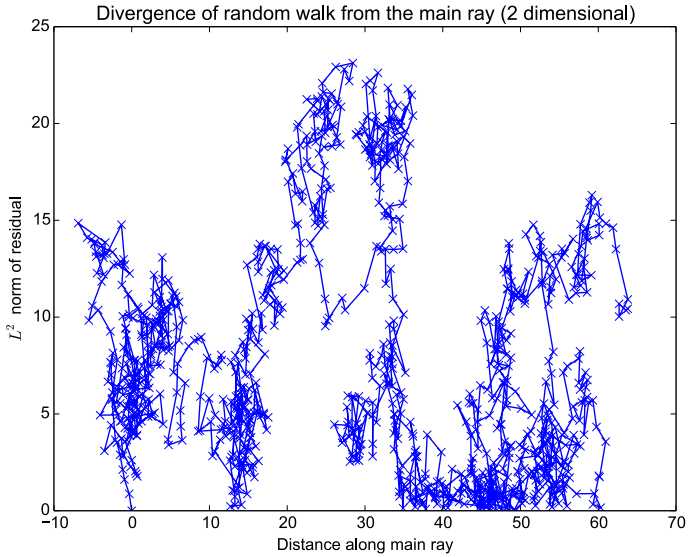
3-D Visualization

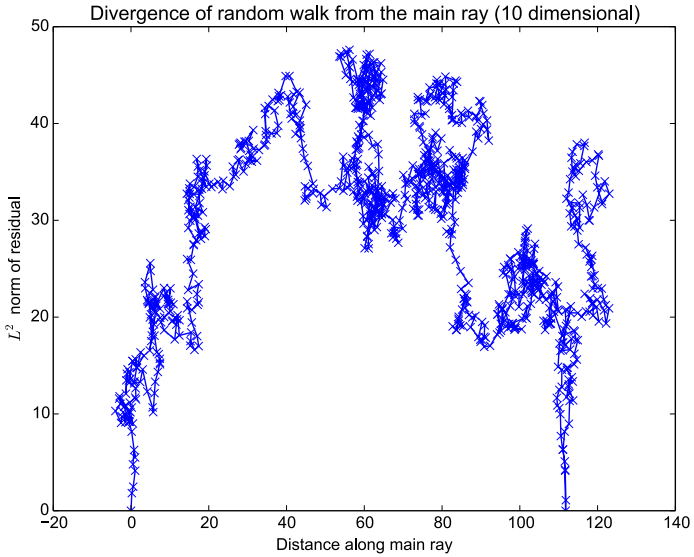


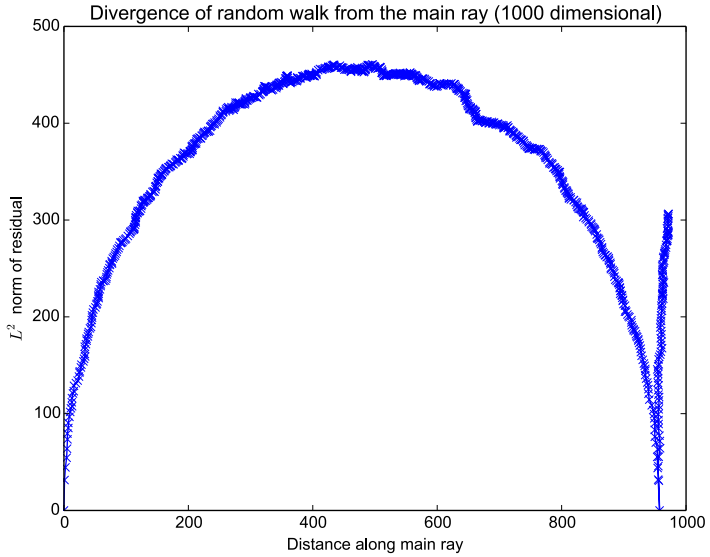
3-D MP-DBM visualization

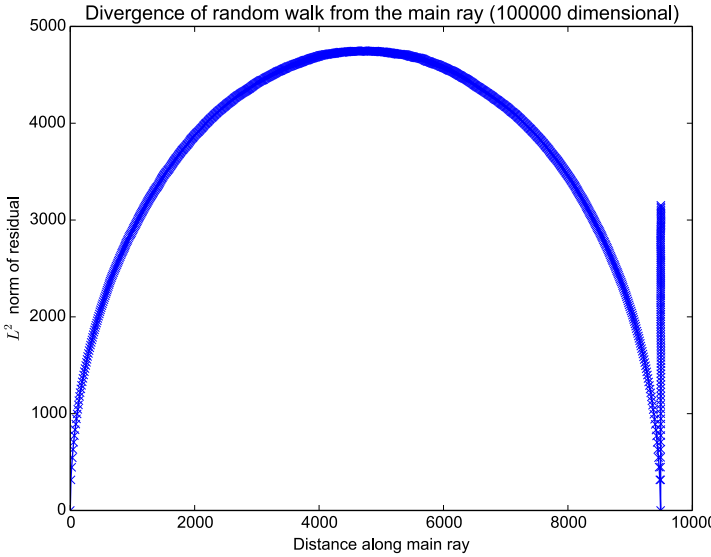


Random walk control

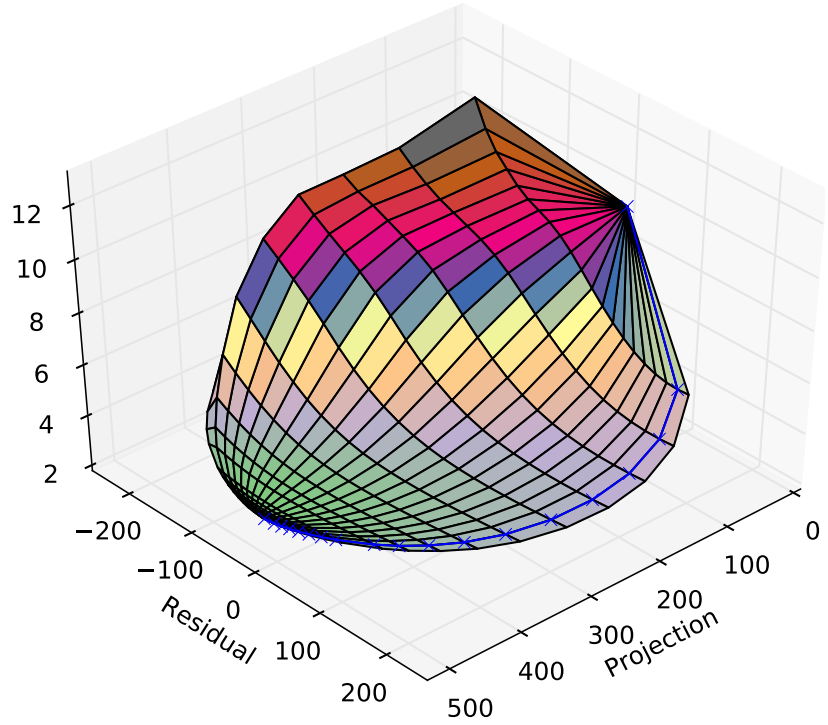




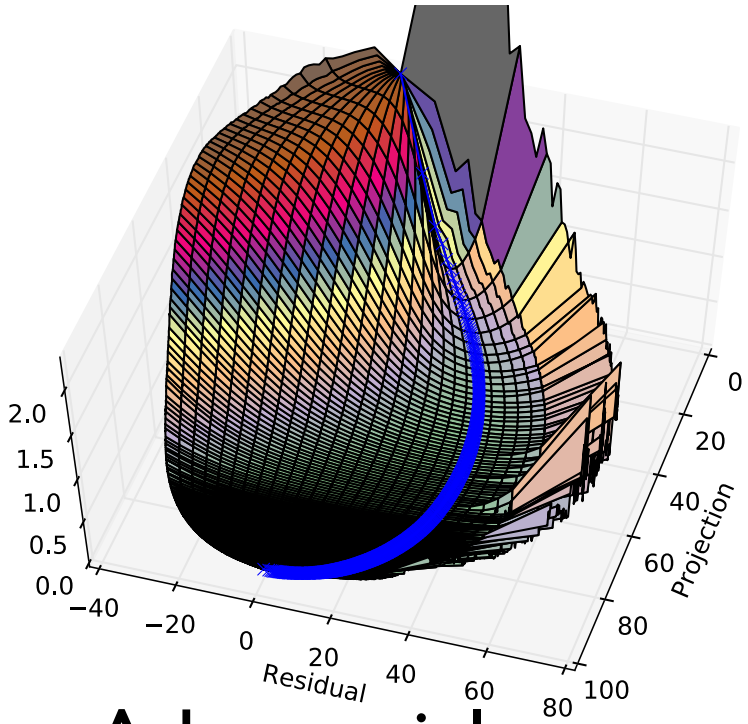




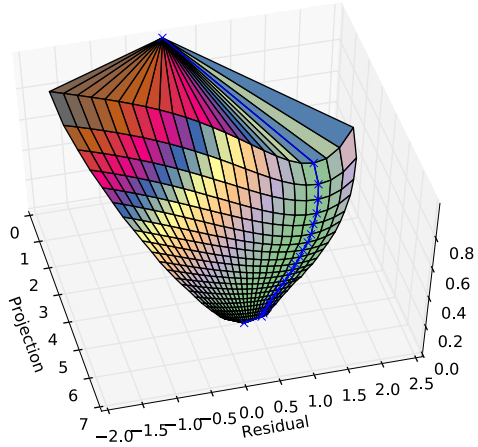
3-D Plots Without Obstacles



LSTM

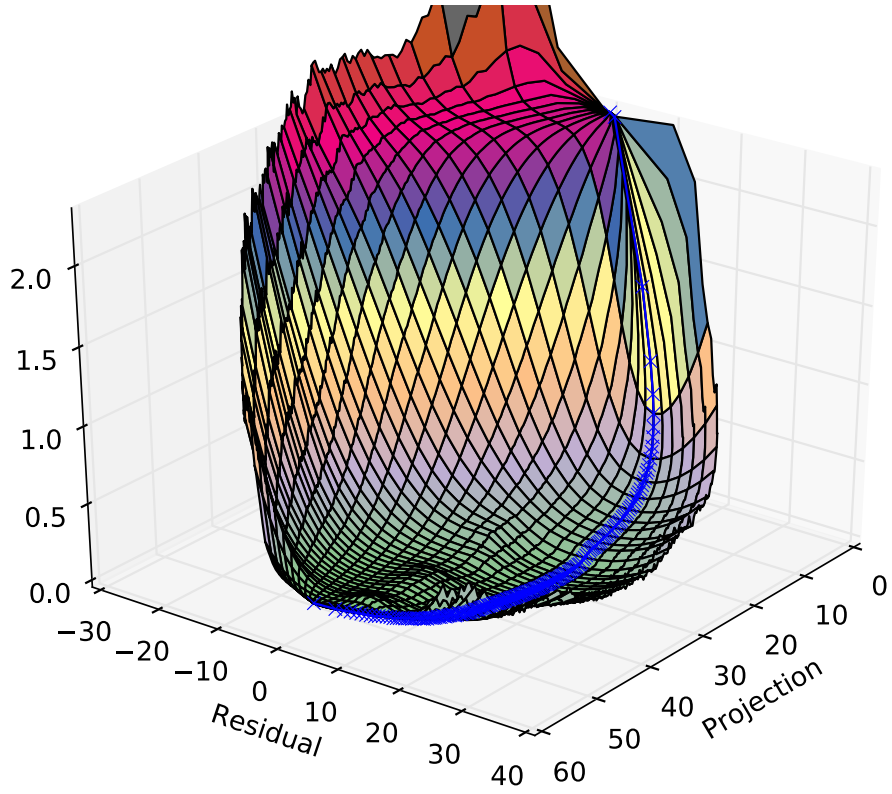


Adversarial ReLUs

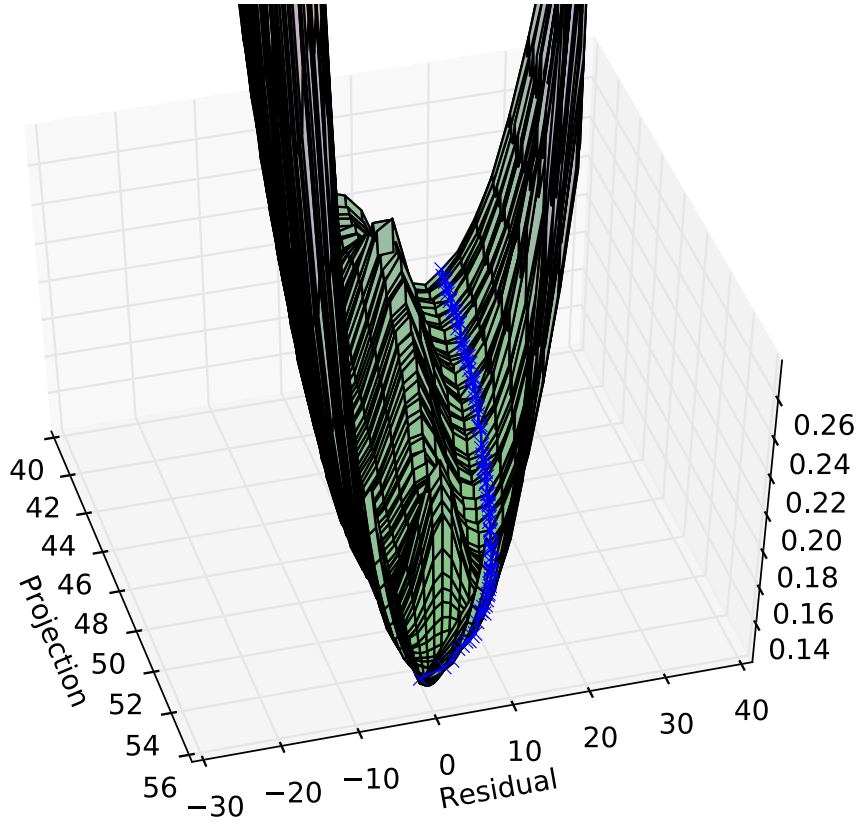


Factored Linear

3-D Plot of Adversarial Maxout



SGD naturally exploits negative curvature!



Obstacles!

Conclusion

- For most problems, there exists a linear subspace of monotonically decreasing values
- For some problems, there are obstacles between this subspace the SGD path
- Factored linear models capture many qualitative aspects of deep network training
- See more visualizations at our poster / demo / paper