Tutorial

pg 3

difference between empirical risk and loss

https://web.archive.org/web/20181024045752/http://ee104.stanford.edu/lectures/supervised.pdf pg 16 - 17

pg 4

Constant initialization

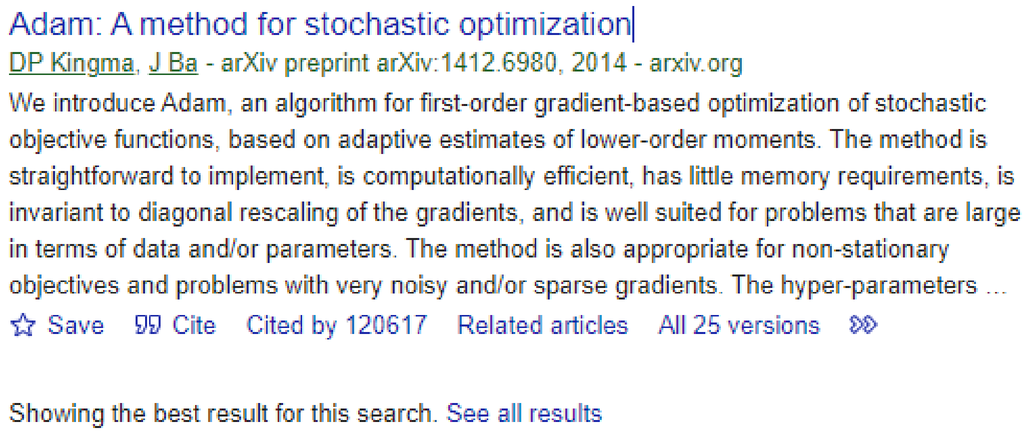
https://www.deeplearning.ai/ai-notes/initialization/

pg 9

https://www.cs.princeton.edu/courses/archive/fall18/cos597G/lecnotes/lecture3.pdf

https://zhuanlan.zhihu.com/p/82757193

oscillation



Adam: A method for stochastic optimization

DP Kingma, J Ba - arXiv preprint arXiv:1412.6980, 2014 - arxiv.org

We introduce Adam, an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. The method is straightforward to implement, is computationally efficient, has little memory requirements, is invariant to diagonal rescaling of the gradients, and is well suited for problems that are large in terms of data and/or parameters. The method is also appropriate for non-stationary objectives and problems with very noisy and/or sparse gradients. The hyper-parameters ...

Showing the best result for this search. See all results

pg 13

https://zhuanlan.zhihu.com/p/32230623

pg 19

都已经要overfitting了。那就是已经拟合了

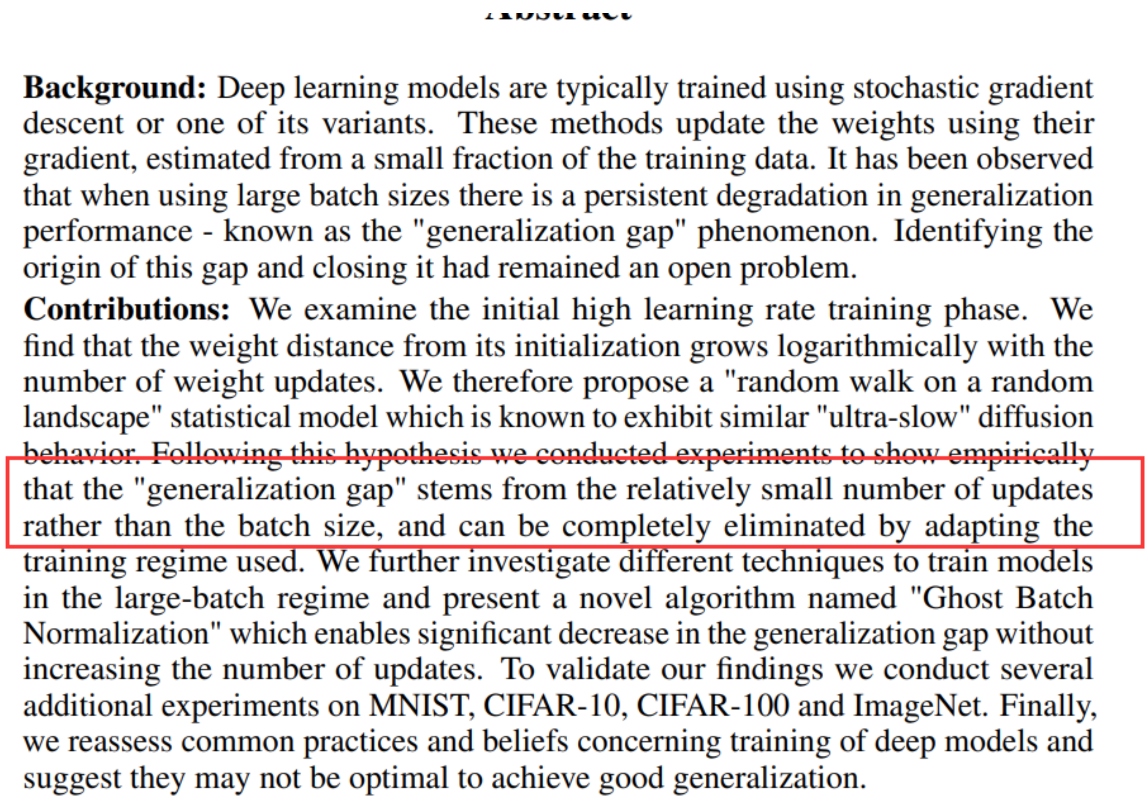
pg 20

why decay?

Anneal (decay) learning rate over time so the parameters can settle into a local minimum

pg 22

https://arxiv.org/pdf/1705.08741.pdf



Abstract

Background: Deep learning models are typically trained using stochastic gradient descent or one of its variants. These methods update the weights using their gradient, estimated from a small fraction of the training data. It has been observed that when using large batch sizes there is a persistent degradation in generalization performance - known as the "generalization gap" phenomenon. Identifying the origin of this gap and closing it had remained an open problem.

Contributions: We examine the initial high learning rate training phase. We find that the weight distance from its initialization grows logarithmically with the number of weight updates. We therefore propose a "random walk on a random landscape" statistical model which is known to exhibit similar "ultra-slow" diffusion behavior. Following this hypothesis we conducted experiments to show empirically that the "generalization gap" stems from the relatively small number of updates rather than the batch size, and can be completely eliminated by adapting the training regime used. We further investigate different techniques to train models in the large-batch regime and present a novel algorithm named "Ghost Batch Normalization" which enables significant decrease in the generalization gap without increasing the number of updates. To validate our findings we conduct several additional experiments on MNIST, CIFAR-10, CIFAR-100 and ImageNet. Finally, we reassess common practices and beliefs concerning training of deep models and suggest they may not be optimal to achieve good generalization.

https://arxiv.org/pdf/1609.04836.pdf

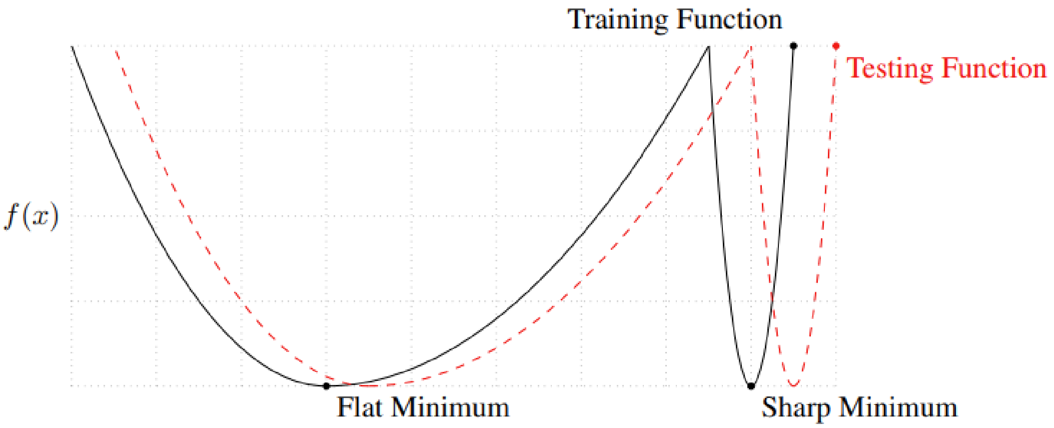
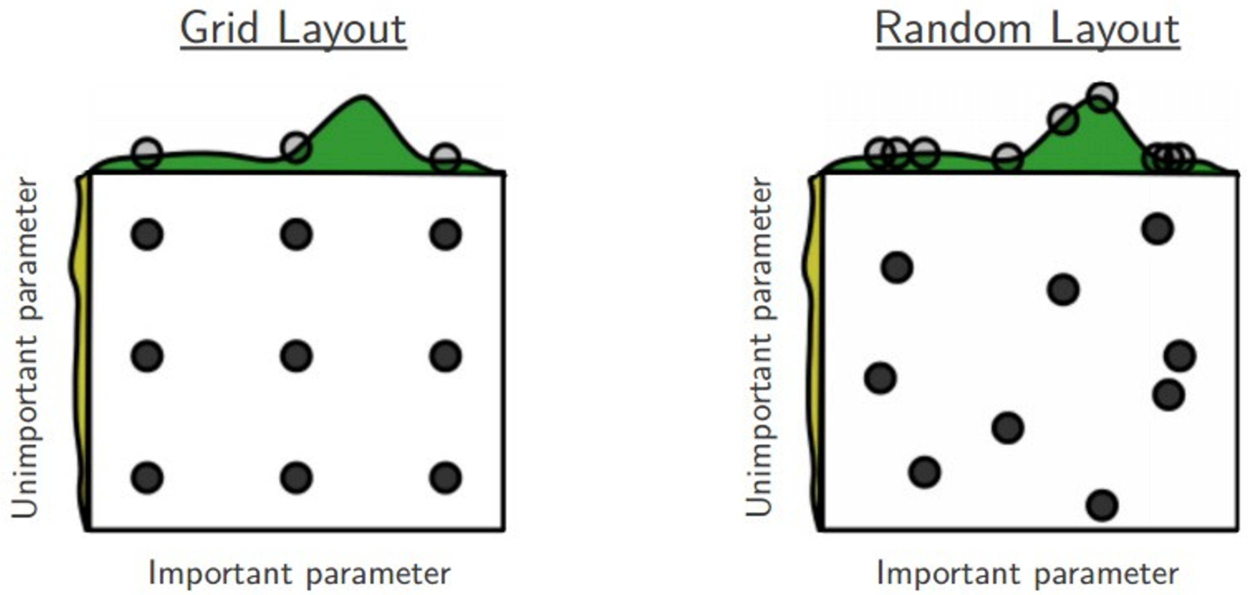


Figure 1: A Conceptual Sketch of Flat and Sharp Minima. The Y-axis indicates value of the loss function and the X-axis the variables (parameters)

pg 24



https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf