

DECOUPLING THE DATA GEOMETRY FROM THE PARAMETER GEOMETRY FOR STOCHASTIC GRADIENTS

SNOWBIRD LEARNING WORKSHOP 2012
EXTENDED ABSTRACT

TOM SCHAUL, SIXIN ZHANG, AND YANN LECUN

1. MOTIVATION

Large-scale learning problems require algorithms that scale benignly with respect to the size of the dataset and the number of parameters to be trained; leading numerous practitioners to favor the classic *stochastic gradient descent* (SGD [1, 2, 3]) over more sophisticated methods.

Besides its fast convergence, SGD has been observed to sometimes lead to significantly better generalization performance than batch gradient descent. SGD is also quicker than batch methods in adapting to non-stationary data distributions. Its Achilles heel are the inherently *sequential* updates, making it very difficult to parallelize across many machines; which is clashing with the goals of large-scale learning.

Our goals here are twofold. On the theoretical level, we want to gain a fuller understanding of how the *dynamics of stochastic updates* contrast with those of batch updates, and how they are affected by the conditioning of energy surfaces, the presence of local optima, and the properties of the data distribution. On the practical level, we want to use this knowledge to design more efficient mini-batch SGD variants (which are parallelizable), together with robust settings for their hyper-parameters.

The study of stochastic gradient methods dates back over six decades [1, 2, 3, 4, 5, 6, 7], but to our knowledge, the present questions remain understudied. The most similar viewpoint is found in approaches based on *natural gradients* and information geometry [8, 9, 10].

2. CONJECTURE

A distinct feature of stochastic updates is that they capture (geometric) properties of the data distribution. This is impossible for batch methods, because these aspects are lost during gradient averaging, allowing them only to retain information about the geometry of the energy function. Those geometries may or may not be related, but decoupling their effects can improve our understanding.

We claim that using the data geometry is precisely what enables SGD to obtain better generalization performance.

3. ANALYSIS

We take a closer look at the dynamics of stochastic gradient methods (with batch sizes ranging from one sample to the full training set), in cases when the energy function exhibits *high curvature* or multiple *local optima*, and when the data distribution is *not homogeneous*.

For this, we first study the continuous-time diffusion process corresponding to SGD, on a class of convex learning problems (where each sample makes a quadratic contribution to the loss). We characterize the expected difference between training and generalization error, and find an analytic formulation of the entire stochastic process for this simple case. The stationary distribution (for each parameter θ) is the Gaussian

$$P(\theta) = \frac{1}{\sqrt{2\pi s}} \exp\left(-\frac{(\theta - \frac{\sigma}{D})^2}{2s^2}\right)$$

with variance

$$s^2 = \frac{\eta\lambda\sigma^2(D-1)}{2mD}$$

where m is the mini-batch size, D is the dataset size, η is the learning rate, λ is the curvature and σ^2 is the variance of the data. There is no difference in expected generalization performance between SGD and batch gradient descent, if the learning rate η decays to zero. We also derive the optimal settings at each iteration for the mini-batch size and the learning rates, given an estimate of the local curvature and covariance matrices.

Second, we analyze a prototypical multi-modal energy surface, where some of the local optima exhibit much more robust generalization properties than others. Preliminary results show that the higher the stochasticity of the updates (i.e., smaller batch-size),

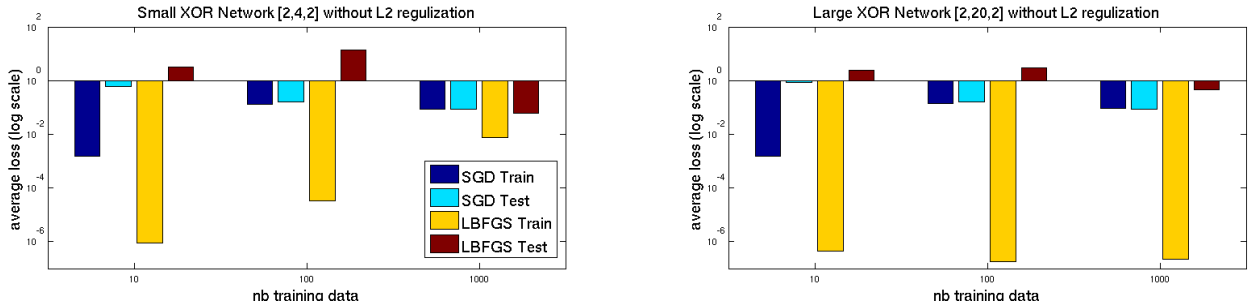


FIGURE 1. Illustration of generalization differences. Plotted are median training and test errors on the generalized XOR task, after training a one hidden layer perceptron until convergence (10k total data samples seen, 10 independent runs per setting). We vary the size of the training set and compare SGD with the batch method LBFGS, for two different network sizes (4 or 20 hidden neurons, on the left and right plots respectively), without using regularization. We observe that LBFGS overfits more prominently to small datasets than SGD.

the higher the likelihood of converging to a robust optimum.

4. EXPERIMENTS

Our analytical results make simplified assumptions on the problem structure. Thus, in order to demonstrate that the results hold qualitatively on real-world problems, we conducted two sets of neural-network training experiments.

Using small multilayer perceptrons on the generalized XOR task, we can do extensive comparisons (on convergence speed and generalization) by varying all of the following parameters: regularization coefficient, size of training dataset, network size, size of mini-batch and learning rate.

In a second batch of experiments, we apply our new settings to training convolutional networks on the MNIST digit recognition dataset, and determine the speedups in computation time gained from adaptive mini-batch sizes. We also determine the impact on generalization of maintaining SGD-levels of stochasticity in the mini-batch updates.

5. CONCLUSION

Motivated by scalability and parallelization issues, study how stochastic gradient updates can improve generalization performance, due to better capturing the underlying geometry of the data distribution. Our approach builds on analytically tractable, prototypical problem classes, and the conclusions are then validated on two classical neural network learning problems.

REFERENCES

- [1] H. Robbins and S. Monro, A stochastic approximation method, *Ann. Math. Stat.*, Vol. 22, pp. 400-407. 1951.
- [2] J. Wolfowitz, On the stochastic approximation method of Robbins and Monro, *Ann. Math. Stat.*, Vol. 23, pp. 457-461. 1952.
- [3] R.D. Martin and C.J. Masreliez. Robust estimation via stochastic approximation. *IEEE Trans. Inform. Theory*, 21, pp. 263-271. 1975.
- [4] B.T. Polyak and A.B. Juditsky. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization* 30, No. 4. pp. 838-855. 1992.
- [5] A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal of Optimization* 19, No. 4. pp. 1574-1609. 2009.
- [6] W. Xu. Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent. Technical report. 2010.
- [7] F. Bach and E. Moulines. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. *Advances in Neural Information Processing Systems (NIPS)*. 2011.
- [8] S. I. Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10, 251-276. 1998.
- [9] J. Peters, and S. Schaal. Natural Actor-Critic. *Neurocomputing*, 71, 1180-1190. 2008.
- [10] Y. Sun, D. Wierstra, T. Schaul, and J. Schmidhuber. Stochastic Search using the Natural Gradient. *International Conference on Machine Learning (ICML)*. 2009.