Advanced Machine Learning
Project Proposal

Si Kai Lee
sl3950

Introduction and Motivation

For the project, I will be working on trying to demonstrate the superiority of Adam over AdaGrad in the settings described in Kawaguchi 2016[1]. In modern machine learning, we are extremely reliant on gradient methods to search for optima. However, we usually treat such methods as black-box optimisers and do not fully understand them. As researchers, it is important for us to understand the advantages and disadvantages of such methods and their assumptions so that we can pick the right method for the problem at hand.

If the above idea does not work out, I will switch to a survey paper on recent advances in stochastic gradient methods for fast convergence namely AdaGrad, AdaDelta and Adam and if time permits, I will also elucidate why parametric models trained with such methods with few iterations have vanishing generalisation error[2] .

Paper Summaries

AdaGrad[3]

AdaGrad is a stochastic gradient optimisation algorithm that uses the geometry learned from the data observed in earlier iterations to speed up the gradient descent optimisation process. Essentially, it attempts to, by adaptively changing the rates of gradient descent for different dimensions, reshape the usually elliptical manifold of the data into a spherical manifold. AdaGrad does so by downweighting frequently seen and less predictive features and upweighting rarer and more predictive features. Hence, AdaGrad does well with sparse datasets.

[1] Kawaguchi, K., 2016. Deep Learning without Poor Local Minima. arXiv preprint arXiv:1605.07110.

[2] Hardt, M., Recht, B. and Singer, Y., 2015. Train faster, generalize better: Stability of stochastic gradient descent. arXiv preprint arXiv:1509.01240.

[3] Duchi, J., Hazan, E. and Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12(Jul), pp.2121-2159.

Adam[4]

Adam is designed to combine the advantages of AdaGrad and RMSProp for stochastic gradient descent. It uses the moving averages of the gradient and the squared gradient to estimate the mean and variance of the gradient and corrects for bias during initialisation. Empirically, it outperforms both AdaGrad and RMSProp.

Deep Learning without Poor Local Minima

Kawaguchi shows, with a set of assumptions and under certain conditions, that in both deep linear and non-linear networks, the squared-loss function is non-convex and non-concave, every local minimum is a global minimum, every critical point that is not a global minimum is a saddle point, and there exist "bad" saddle points (where the Hessian has no negative eigenvalue) for the deeper networks (with more than three layers), whereas there are no "bad" saddle points for the shallow networks (with at most three layers) [5].

Approach

By looking at the above methods from a stochastic optimisation viewpoint, it is clear that the use of approximations to estimate first and second order gradient methods instead of inverting either the Jacobian or the Hessian has been very successful. However, with just the online learning framework proposed by Zinkevich[6] it is unclear why Adam works better than AdaGrad in practice. By analysing the above methods with the framework and settings used in Kawaguchi[7], I hope to demonstrate that Adam has a high lower bound compared to AdaGrad.

[4] Kingma, D. and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[5] Kawaguchi, K., 2016. Deep Learning without Poor Local Minima. arXiv preprint arXiv:1605.07110.

[6] Zinkevich, M., 2003. Online convex programming and generalized infinitesimal gradient ascent.

[7] Kawaguchi, K., 2016. Deep Learning without Poor Local Minima. arXiv preprint arXiv:1605.07110.