Chapter 1

Deep Learning Theory

Notes based on blog by Desh Raj https://desh2608.github.io/.

1.1 Optimization

Neural networks can be viewed as trying to minimize a loss function. The reason this is difficult is because the loss function is non-convex. Non-convex functions can have many local minima, saddle points, flat regions, and varying curvature which makes non-convex optimization at least NP-Hard (https://www.cs.cornell.edu/courses/cs6787/2017fa/Lecture7.pdf). If this is the case, why can deep networks find approximately reasonable solutions?

Almost all neural networks are trained by moving the parameters in the direction of a non-zero gradient. The goals of such descent are to find a critical point $\nabla = 0$ and to find local optimum $\nabla = 0$ and $\nabla^2 > 0$ (aka Hessian is positive semi-definite).

Critical Points Parameter θ, usually update in the following form: $\theta_{n+1} = \theta_n - \eta \nabla f(\theta_n)$. The hyperparameter in the update equation is the learning rate, η. We know that we want it to be small but how do we know that η is small enough? We utilize the Hessian aka ∇^2. Suppose there exists some β such that $-\beta I \leq \nabla^2 f(\theta) \leq \beta I$. Then a higher β means ∇^2 varies more so the learning rate should be lower. It can be more formally proved that in $O(\frac{\beta}{\epsilon^2})$, θ will arrive at a critical point.

Saddle Points One issue that such a descent may run into is the issue of saddle points. Several results (http://proceedings.mlr.press/v40/Ge15.pdf, https://arxiv.org/pdf/1602.04915.pdf, https://arxiv.org/pdf/1703.00887.pdf) show that gradient descent can be done by evading saddle points. The main ideas from these papers are that there are many saddle points but it is hard to converge to one and while Hessians can be used to avoid the saddle points, we do not actually have to because a noisy form of gradient descent converges to a local optimum in a polynomial number of steps. The last paper discusses how perturbed gradient descent does a better job of escaping saddle points than regular gradient descent.