CS 165: Foundations of Machine Learning **Winter 2019**

Lecture 2: January 10

Lecturer: Anima Anandkumar *Scribes: Nikola Kovachki, Jialin Song*

2.1 Optimality conditions

We consider unconstrainted optimization on $\mathbb{R}^d$. At the strict local minumum $x \in \mathbb{R}^d$, we expect

1. $\nabla_x f(x) = 0$ - stationary or critical points.
2. $\nabla_x^2 f(x) \succ 0$ - positive-definite Hessian.

Consider the Taylor series

$$f(x + \Delta x) = f(x) + \langle \Delta x, \nabla f(x) \rangle + \frac{1}{2}\langle \Delta x, \nabla^2 f(x) \Delta x \rangle + \frac{1}{6}\langle \Delta x, \nabla^3 f[\Delta x, \Delta x] \rangle + \dots$$

Positive definiteness of the Hessian insures the function increases around a neighborhood of the minima. For a saddle point, the positive definetness condition is violated. Look at the eigenvalues of the Hessian to determine this. Further, we need to look at higher order terms if the minima are degenerate.

Consider $f(x_1, \dots, x_m)$ which is permutation invarient. Then any convex combinations of critical point are also critical points. This tells us that symmetries in the function create degenerate critical points. The hidden layers of deep neural networks naturally induce symmetry. For example, consider a neural network with two hidden layers and uses ReLU as its activation function. Then we can scale the weights of the first layer by a factor of $k > 0$ and the weights of the second layer by a factor of $\frac{1}{k}$. The behavior of this network will not change due to the threshold nature of ReLU. Thus we have obtained two set of weights producing a same model.

If x is a saddle point, gradient descent

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

and Newton's method

$$x_{t+1} = x_t - (\nabla^2 f(x_t) + \gamma I)^{-1} \nabla f(x_t)$$

will both get stuck at x.

Let x be a saddle and consider the Hessian $\nabla^2 f(x)$ letting $v_1, \dots, v_k$ be the eigenvector and $\lambda_1, \dots, \lambda_k$ be the eigenvalues. We can use a modified Newton's method

$$x_{t+1} = x_t - (\nabla^2 f(x_t) + \gamma I)^{-1} \nabla f(x_t)$$

where $\gamma > -\min\{\lambda_1, \dots, \lambda_k\}$, i.e. $\nabla^2 f(x_t) + \gamma I$ is positive-definite.

2.2 Cubic Regularization

Taylor expanding,

$$f(x + \Delta x) = f(x) + \langle \Delta x, \nabla f(x) \rangle + \frac{L}{2}\|\Delta x\|^2$$

where $\|\nabla^2 f(x)\| \leq L$ or equivalently $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. Gradient decent is equivalent to optimizing this upper bound with η set to $\frac{1}{L}$. Expanding to one more term we have

$$f(x + \Delta x) = f(x) + \langle \Delta x, \nabla f(x) \rangle + \frac{1}{2}\langle \Delta x, \nabla^2 f(x) \Delta x \rangle + \frac{M}{6}\|\Delta x\|^3$$

where now we assume a Lipshitz Hessian

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq M\|x - y\|.$$

Optimizing this upper bound is the supposedly the "natural" extension to gradient decent. This yields essentially to a regularized Newton's method.

2.3 Stochastic Gradient Decent

Model SGD by the dynamic

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \xi_t$$

where $\xi_t \sim U(B(r))$. For larger enough r, the noise will push away from saddles.