Exercises
Deep Learning
Fall 2018

Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Fernando Pérez-Cruz
Web http://www.da.inf.ethz.ch/teaching/2018/DeepLearning/

Series Monday, Nov 12, 2018
(Deep Learning, Exercise series 7)

Problem 1 (Gradient Descent):

Consider the following unconstrained optimization problem

$$\min_{x \in \mathbb{R}^d} f(x). \tag{1}$$

Given a starting point $x_0 \in \mathbb{R}^d$ and a step-size $\alpha \in \mathbb{R}_+$ Gradient Descent (GD) updates its iterates as follows

$$x_{k+1} = x_k - \alpha \nabla f(x_k), \;\; k = 0, 1, 2, ... \tag{2}$$

until some stopping criterion is satisfied (usually $\|\nabla f(x_k)\| \leq \epsilon$, where $\epsilon > 0$ is some user specified accuracy).

1. How do the cost of computing $\nabla f(x_k)$ scale in d?
2. Show that for GD with optimal step size two consecutive steps are orthogonal to each other, i.e.

$$\nabla f(x_k)^\intercal \nabla f(x_{k+1}) = 0 \tag{3}$$

Now, suppose that f is convex and Lipschitz-smooth ($\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|, \; \forall x, y$).

3. Show that this implies

$$0 \leq f(y) - f(x) - \nabla f(x)^\intercal (y - x) \leq \frac{L}{2}\|x - y\|^2 \tag{4}$$

   Hint: Use the fundamental theorem of calculus as well as Cauchy-Schwarz inequality.

4. Using the result of 3., show that the GD updates satisfy

$$f(x_{k+1}) - f(x_k) \leq -\alpha(1 - \frac{L}{2}\alpha)\|\nabla f(x_t)\|^2 \tag{5}$$

5. For what values of $\alpha > 0$ does GD strictly decrease f in every step. What's the best possible value of α?

Problem 2 (Stochastic Gradient descent):

The most common way of training Neural Networks involves some variant of Stochastic Gradient Descent (SGD). In its most basic form, a datapoint i is sampled uniformly at random (thus $\mathbb{E}(\nabla f_i(x) = \nabla f(x))$ in each iteration $k = 1, 2, ...$ and the iterates are updates as

$$x_{k+1} = x_k - \alpha_k \nabla f_i(x_k). \tag{6}$$

Assume the function f is smooth, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall \, x, y. \tag{7}$$

1. Show that, given x_k and a constant step size $\alpha_k = 1/(2L)$, SGD does not converge to a critical point x^*, i.e. that

$$\mathbb{E}\left(\|x^{k+1} - x^*\|_2^2\right) \geq \frac{1}{(2L)^2}\mathbb{E}\|\nabla f_i(x_k)\|_2^2 \geq \frac{1}{(2L)^2}\mathsf{Var}[\nabla f_i] \tag{8}$$

2. Name two possibilities to retain convergence.