Exercises
Deep Learning
Fall 2017

Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Prof. Dr. Thomas Hofmann
Web http://www.da.inf.ethz.ch/teaching/2017/DeepLearning/

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

Solution 1 (Gradient Descent):

(1) The cost sale as $O(d)$ since d partial derivatives need to be computed for the gradient.
(2) We want to choose α such that the objective f is minimized along the search direction $-\nabla f(x_k)$, i.e.

$$\min_{\alpha\in\mathbb{R}} f(x_k - \alpha\nabla f(x_k)) := f(x_{k+1}). \tag{1}$$

A *necessary* optimality condition is that the first derivative of $f(x_{k+1})$ w.r.t. α is zero, i.e.

$$\begin{aligned} & \tfrac{\partial f(x_{k+1})}{\partial\alpha} = \tfrac{\partial f(x_{k+1})}{\partial x_{k+1}}\tfrac{\partial x_{k+1}}{\partial\alpha} = 0 && (2)\\ \Leftrightarrow\quad & \nabla f(x_{k+1})^\intercal(-\nabla f(x_k)) = 0, && (3)\end{aligned}$$

which proves the assertion.
(3) The left hand side of Eq. (7) follows directly from the convexity of f, which implies

$$f(y) \geq f(x) + \nabla f(x)^\intercal(y-x),\ \ \forall(x,y)\in\mathbb{R}^{d\times d}. \tag{4}$$

The fundamental theorem of Calculus states that

$$f(y) - f(x) = \int_x^y \nabla f(\tau)d\tau. \tag{5}$$

By substituting $\tau := (1-t)x + ty$ we have $\frac{d\tau}{dt} = y - x$ and thus $d\tau = (y-x)dt$. Hence we can rewrite (5) as

$$f(y) - f(x) = \int_0^1 \nabla f(x + t(y-x))^\intercal(y-x)dt. \tag{6}$$

Thus

$$\begin{aligned} f(y) - f(x) - \nabla f(x)^\intercal(y-x) &= \int_0^1 \nabla f(x+t(y-x))^\intercal(y-x)dt - \nabla f(x)^\intercal(y-x)\\ &= \int_0^1 [\nabla f(x+t(y-x)) - \nabla f(x)]^\intercal(y-x)dt\\ &\leq \int_0^1 \|\nabla f(x+t((y-x)) - \nabla f(x)\|\|y-x\|dt \text{ (by Cauchy-Schwarz)}\\ &\leq \int_0^1 L\|t(y-x)\|\|y-x\|dt \text{ (by L-smoothness)}\\ &= L\|y-x\|^2\int_0^1 tdt\\ &= \frac{L}{2}\|y-x\|^2. \end{aligned} \tag{7}$$

(4) By (7) and the definition of x_{k+1} we directly have

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \frac{L}{2}\|x_{k+1} - x_k\|^2 + \nabla f(x_k)^\intercal(x_{k+1} - x_k)\\ &= \frac{L}{2}\|-\alpha\nabla f(x_k)\|^2 - \alpha\|\nabla f(x_k)\|^2\\ &= -\alpha(1 - \frac{L\alpha}{2})\|\nabla f(x_k)\|^2 \end{aligned} \tag{8}$$

(5) Requiring (8) to be negative gives $(1 - \frac{L\alpha}{2}) \overset{!}{>} 0$, which yields $\alpha \in (0, \frac{2}{L})$. To maximize the function decrease we solve

$$\max_{\alpha > 0} g(\alpha) := \alpha(1 - \frac{L\alpha}{2}). \tag{9}$$

The first- and second derivative write as follows

$$g'(\alpha) = 1 - L\alpha, \text{ and } g''(\alpha) = -L. \tag{10}$$

Since $L \geq 0$ we know that g is concave and hence setting $g' = 0$ yields the global maximizer $\alpha^* = \frac{1}{L}$.

Solution 2 (Stochastic Gradient Descent):

1. As stated in the exercise, we assume x_k as given (for simplicity) and only consider the gradients $\nabla f_i(x_k)$ as random variable (due to sampling).

$$\begin{aligned}
\mathbb{E}(\|x^{k+1} - x^*\|_2^2) &= \mathbb{E}(\|x_k - \alpha \nabla f_i(x_k) - x^*\|_2^2) \\
&= \mathbb{E}(\|x_k - x^*\|_2^2 - 2\alpha \nabla f_i(x_k)^\intercal (x_k - x^*) + \alpha^2 \|\nabla f_i(x_k)\|_2^2) \\
&= \|x_k - x^*\|_2^2 - 2\alpha \nabla f_i(x_k)^\intercal (x_k - x^*) + \alpha^2 \mathbb{E}(\|\nabla f_i(x_k)\|_2^2) \\
&\geq \|x_k - x^*\|_2^2 - 2\alpha \|\nabla f_i(x_k) - \nabla f(x*)\| \|x_k - x*\| + \alpha^2 \mathbb{E}(\|\nabla f_i(x_k)\|_2^2) \text{ (CS)} \\
&\geq \|x_k - x^*\|_2^2 - 2\alpha L \|(x_k - x^*)\|^2 + \alpha^2 \mathbb{E}(\|\nabla f_i(x_k)\|_2^2) \text{ (Lip.)} \\
&= \alpha^2 \mathbb{E}(\|\nabla f_i(x_k)\|_2^2) \text{ (step size).}
\end{aligned} \tag{11}$$

   Finally, the last inequality follow from the fact that Var$[\nabla f_i] = \mathbb{E}[\nabla f_i^\intercal \nabla f_i] + \mathbb{E}[\nabla f_i]^2$.

2. Employ a step-size shedule with decreasing stepsize in k or perform some kind of variance reduction either via increasing the batch size over time or by the use of so-called control variates (see e.g. SVRG).