Optimization Theory and Algorithm II	**October 8, 2022**

Homework 3

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

HW 1 *Prove Theorem 1, 3 and 4 in Lecture 7.*

HW 2 *The problem of finding the shortest distance from a point* $\mathbf{x}_0$ *to the hyperplane* $\{\mathbf{x}|A\mathbf{x} = \mathbf{b}\}$*, where* A *has full row rank, can be formulated as the quadratic program*

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}\|^2$$
$$s.t.\ A\mathbf{x} = \mathbf{b}.$$

(i) Show that optimal solution is

$$\mathbf{x}^* = \mathbf{x}_0 + A^\top(AA^\top)^{-1}(A\mathbf{x}_0 - \mathbf{b}).$$

(ii) Using above results, provide the projected gradient descent algorithm for

$$\min_{\mathbf{x}} \ f(\mathbf{x}) \tag{1}$$
$$s.t.\ A\mathbf{x} = \mathbf{b}. \tag{2}$$

HW 3 *We consider the following optimization problem:*

$$\min_{\mathbf{x}} \ f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}).$$

And assume that f *is* β*-smooth and* α*-strong convex. Using mini-bath SGD with fixed learning rate to solve it as*

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \frac{s}{n_b}\sum_{i_t \in D_t} \nabla f_{i_t}(\mathbf{x}^t),$$

where $D_t \subset \{1, 2, \ldots, m\}$ *are drawn randomly and* $|D_t| = n_b$ *is the size of* D_t*. We further suppose that*

(1) The index D_t *does not depended from the previous* $D_0, D_1, \ldots, D_{t-1}$*.*

(2) $\mathbb{E}_{i_t \in D_t}[\nabla f_{i_t}(\mathbf{x}^t)] = \nabla f(\mathbf{x}^t)$ *(Unbiased Estimation).*

(3) $\mathbb{E}_{i_t \in D_t}[\|\nabla f_{i_t}(\mathbf{x}^t)\|^2] \leq \sigma^2 + \|\nabla f(\mathbf{x}^t)\|^2$ *(control the variance).*

Prove

(i) $\mathbb{E}_{D_t}\|g^t\|^2 = \frac{\sigma^2}{n_b} + \|\nabla f(\mathbf{x}^t)\|^2$*, where* $\mathbf{g}^t = \frac{1}{n_b}\sum_{i \in D_t} \nabla f_i(\mathbf{x}^t)$*.*

(ii)

$$\mathbb{E}_{D_t}[f(\mathbf{x}^{t+1})] \leq f(\mathbf{x}^t) - s\nabla f(\mathbf{x}^t)^\top \mathbb{E}_{D_t}[\mathbf{g}^t] + \frac{\beta s^2}{2}\mathbb{E}_{D_t}[\|\mathbf{g}^t\|^2].$$

(iii)

$$\mathbb{E}_{D_t}[f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t)] \leq -(s - \frac{\beta s^2}{2})\|\nabla f(\mathbf{x}^t)\|^2 + \frac{\beta s^2}{2n_b}\sigma^2.$$

(iv) Then

$$\mathbb{E}[f(\mathbf{x}^{t+1}) - f^*] - \frac{\beta s}{2n_b\alpha(2-\beta s)}\sigma^2 \leq (1-\alpha s(2-\beta s))\left[\mathbb{E}[f(\mathbf{x}^t) - f^*] - \frac{\beta s}{2n_b\alpha(2-\beta s)}\sigma^2\right].$$

HW 4 *Read Textbook Page 470. And select one of the data set to implement*

(1) SGD for Logistic Regression with fixed learning rate.

(2) SGD for Logistic Regression with decreasing learning rate.

(2) SVRG for Logistic Regression.

References