Solution to Homework 8

Total 30 points

Problem 1. (5 points) Suppose $F(\mathbf{x}) \triangleq \mathbb{E}_\xi[f(\mathbf{x};\xi)]$ is L-smooth and μ-strongly convex, $g(\mathbf{x}_t,\xi_t)$ is an unbiased estimator of $\nabla F(\mathbf{x}_t)$, with bounded variance σ^2. Show that the stochastic gradient method with fixed step size $\eta \le 1/(2L)$ achieves

$$\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] \le (1-2\eta\mu)^t (F(\mathbf{x}_0) - F(\mathbf{x}^*)) + \frac{\eta\sigma^2 L}{4\mu}.$$

Solution. By the smoothness, we have

$$\begin{aligned} F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) \le& \nabla F(\mathbf{x}_t)^\top(\mathbf{x}_{t+1}-\mathbf{x}_t) + \frac{L}{2}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2^2 \\ =& -\eta\nabla F(\mathbf{x}_t)^\top g(\mathbf{x}_t,\xi_t) + \frac{L}{2}\eta^2\left\|g(\mathbf{x}_t,\xi_t)\right\|_2^2 \end{aligned}$$

Then, we can take expectation and get

$$\begin{aligned} \mathbb{E}_t[F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)] \le& -\eta\nabla F(\mathbf{x}_t)^\top \mathbb{E}[g(\mathbf{x}_t,\xi_t)] + \frac{L}{2}\eta^2\mathbb{E}[\|g(\mathbf{x}_t,\xi_t)\|_2^2] \\ \le& -\eta\left\|\nabla F(\mathbf{x}_t)\right\|_2^2 + \frac{L}{2}\eta^2\sigma^2 \\ \le& -2\eta\mu(F(\mathbf{x}_t) - F(\mathbf{x}^*)) + \frac{L}{2}\eta^2\sigma^2, \end{aligned}$$

where the last inequality comes from the mu-strong convexity. Thus,

$$\mathbb{E}_t[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)] \le (1-2\eta\mu)(F(\mathbf{x}_t) - F(\mathbf{x}^*)) + \frac{L}{2}\eta^2\sigma^2.$$

By taking expectation over all randomness, we have

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)] \le (1-2\eta\mu)\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] + \frac{L}{2}\eta^2\sigma^2,$$

which indicates

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) - \frac{L}{4\mu}\eta\sigma^2] \le (1-2\eta\mu)\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] - \frac{L}{4\mu}\eta\sigma^2.$$

Thus, we have

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)] \le (1-2\eta\mu)^t (F(\mathbf{x}_0) - F(\mathbf{x}^*)) + \frac{L}{4\mu}\eta\sigma^2.$$

Problem 2. (5 points) In this problem, we study a stochastic gradient method with a projection step. Let $F : \mathbb{R}^d \to \mathbb{R}$ be differentiable and μ-strongly convex, and let $\mathcal{C}$ be a closed, convex set. Consider the projected stochastic gradient method

$$\mathbf{x}_{t+1} = \mathcal{P}_{\mathcal{C}}(\mathbf{x}_t - \eta_t G(\mathbf{x}_t)),$$

where $G(\mathbf{x}_t)$ is an unbiased estimate of $\nabla F(\mathbf{x}_t)$. Assume that the randomness in $G(\mathbf{x}_t)$ is independent of all past randomness in the algorithm. Letting $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{C}} F(\mathbf{x})$, prove that the iterates satisfy the bound

$$\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2] \le (1 - 2\eta_t\mu)\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] + \eta_t^2 B^2$$

where $B^2 = \sup_{\mathbf{x}\in\mathcal{C}} \mathbb{E}\,\|G(\mathbf{x})\|_2^2$.
Solution. We use non-expansiveness of the projection operator and the fact that $\mathbf{x}_t \in \mathcal{C}$ to obtain

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 &= \|\mathcal{P}_{\mathcal{C}}(\mathbf{x}_t - \eta_t G(\mathbf{x}_t)) - \mathcal{P}_{\mathcal{C}}(\mathbf{x}^*)\|_2^2 \\
&\le \|\mathbf{x}_t - \mathbf{x}^* - \eta_t G(\mathbf{x}_t)\|_2^2 \\
&= \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2\,\|G(\mathbf{x}_t)\|_2^2 - 2\eta_t\langle G(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle \\
&\le \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2\,\|G(\mathbf{x}_t)\|_2^2 - 2\eta_t\langle G(\mathbf{x}_t) - G(\mathbf{x}^*), \mathbf{x}_t - \mathbf{x}^*\rangle
\end{aligned}$$

where the last inequality follows from optimality of $\mathbf{x}^*$. Now taking the expectations on both sides conditioned on $\mathbf{x}_t$, we have

$$\begin{aligned}
\mathbb{E}_t[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2] &\le \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2 B^2 - 2\eta_t\langle\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}^*), \mathbf{x}_t - \mathbf{x}^*\rangle \\
&\le (1 - 2\eta_t\mu)\,\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2 B^2
\end{aligned}$$

where the second line follows by μ-strong convexity of F. By taking expectation on both side, we can get the conclusion.

Problem 3. (5 points) Let $F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$, where $f_i(\mathbf{x})$ is differentiable and L-smooth. Suppose j is uniformly sampled from $\{1, 2, \dots, n\}$. Show that

$$\mathbb{E}[\|\nabla f_j(\mathbf{x})\|_2^2] \le L^2\mathbb{E}[\|\mathbf{x} - \mathbf{x}^*\|_2^2] + \mathbb{E}[\|\nabla f_j(\mathbf{x}) - \nabla F(\mathbf{x}))\|_2^2]$$

where $\mathbf{x}^*$ is a minimizer of $F(\mathbf{x})$.
Solution.

$$\begin{aligned}
\mathbb{E}[\|\nabla f_j(\mathbf{x})\|_2^2] &= \mathbb{E}[\|\nabla f_j(\mathbf{x}) - \nabla F(\mathbf{x}) + \nabla F(\mathbf{x})\|_2^2] \\
&= \mathbb{E}[\|\nabla f_j(\mathbf{x}) - \nabla F(\mathbf{x})\|_2^2] + \mathbb{E}[(\nabla f_j(\mathbf{x}) - \nabla F(\mathbf{x}))^\top\nabla F(\mathbf{x})] + \mathbb{E}[\|\nabla F(\mathbf{x})\|_2^2] \\
&= \mathbb{E}[\|\nabla f_j(\mathbf{x}) - \nabla F(\mathbf{x})\|_2^2] + \mathbb{E}[\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{x}^*)\|_2^2] \\
&\le L^2\mathbb{E}[\|\mathbf{x} - \mathbf{x}^*\|_2^2] + \mathbb{E}[\|\nabla f_j(\mathbf{x}) - \nabla F(\mathbf{x}))\|_2^2]
\end{aligned}$$

The last equation is due to $\nabla F(\mathbf{x}^*) = 0$ and $\mathbb{E}[\nabla f_j(\mathbf{x})] = \nabla F(\mathbf{x})$.

Problem 4. (15 points) In this problem, you are required to use stochastic gradient method to solve the following quadratic problem:

$$f(\mathbf{x}) = \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{a}_i^\top \mathbf{x} - b_i)^2 = \frac{1}{2n}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2,$$

where $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{b} \in \mathbb{R}^n$. The homework ZIP file contains two text files, labeled `A.txt` and `b.txt`, that contains an $n \times d$ matrix $\mathbf{A}$ and an n-dimensional vector $\mathbf{b}$, with $n = 500$, $d = 50$.

(a) (2 points) Compute the closed form of $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$.

(b) (4 points) Implement the stochastic gradient method for minimizing f with constant step size and diminishing step size.

(c) (3 points) Plot the error $\|\mathbf{x}_t - \mathbf{x}^*\|_2$ versus the iteration number, where $\mathbf{x}^*$ is computed by (a).

(d) (3 points) Now suppose that after every $T = 10$ iterations, you are allowed to evaluate the exact gradient $f(\mathbf{z})$, where $\mathbf{z}$ is the current iterate. Construct a better stochastic gradient estimate and implement it.

(e) (3 points) Plot the error $\|\mathbf{x}_t - \mathbf{x}^*\|_2$ and compare it to the naive scheme from (b).