Exam Stochastic Optimization – part I
31 January 2023

duration 1.5 hours, paper documents allowed

Exercise 1. We consider the following optimization problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N} f_i(x) + g(x)$$

where for all $i \in \{1,\dots,N\}$, f_i is a differentiable convex function whose gradient is L-Lipschitz and g is a convex, lower-semi-continuous function whose proximal operator is easy to compute. We shall denote $f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$ and $F(x) = f(x) + g(x)$.

Notation: $U(\{1,\dots,N\})$ is the uniform distribution over the set $\{1,\dots,N\}$ and $B(p)$ is the Bernoulli distribution with mean p. We denote the conditional expection with $\mathbb{E}_k[X_k] = \mathbb{E}[X_k|i_0, b_0, \dots, i_k, b_k]$.

The goal of this exercise is to study the convergence of the following prox-SVRG algorithm

$$\begin{aligned}
&x_0 \in \mathbb{R}^d,\, w_0 = x_0, \gamma > 0, p \in]0,1[\\
&\forall k \in \mathbb{N}:\\
&\qquad i_{k+1} \sim U(\{1,\dots,N\})\\
&\qquad G_{k+1} = \nabla f(w_k) + \nabla f_{i_{k+1}}(x_k) - \nabla f_{i_{k+1}}(w_k)\\
&\qquad x_{k+1} = \mathrm{prox}_{\gamma g}\left(x_k - \gamma G_{k+1}\right)\\
&\qquad b_{k+1} \sim B(p)\\
&\qquad w_{k+1} = (1-b_{k+1})w_k + b_{k+1}x_k
\end{aligned}$$

We shall not assume the strong convexity of F, only its convexity. We will proceed by proving the following points.

1. Show that

$$f(x_{k+1}) \le f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

2. Show that for any $x_* \in \arg\min_x F(x)$

$$g(x_{k+1}) + \langle G_{k+1}, x_{k+1} - x_k\rangle + \frac{1}{2\gamma}\|x_{k+1} - x_k\|^2 \le g(x_*) + \langle G_{k+1}, x_* - x_k\rangle + \frac{1}{2\gamma}\|x_k - x_*\|^2 - \frac{1}{2\gamma}\|x_{k+1} - x_*\|^2$$

3. Show that

$$\mathbb{E}_k[\langle G_{k+1}, x_* - x_k\rangle] = \langle \nabla f(x_k), x_* - x_k\rangle$$

4. Show that for all $\alpha > 0$

$$\langle G_{k+1} - \nabla f(x_k), x_k - x_{k+1} \rangle \leq \frac{\alpha}{2}\|G_{k+1} - \nabla f(x_k)\|^2 + \frac{1}{2\alpha}\|x_k - x_{k+1}\|^2$$

5. Show that

$$\|G_{k+1} - \nabla f(x_k)\|^2 \leq 2\|\nabla f_{i_{k+1}}(x_k) - \nabla f_{i_{k+1}}(x_*)\|^2 + 2\|\nabla f_{i_{k+1}}(w_k) - \nabla f_{i_{k+1}}(x_*)\|^2$$

6. Denote $\mathcal{D}_k = \frac{1}{N}\sum_{i=1}^{N} \|\nabla f_i(w_k) - \nabla f_i(x_*)\|^2$. Show that

$$\mathbb{E}[\mathcal{D}_{k+1} | i_0, b_0, \ldots, i_k] = p\mathbb{E}_k\Big[\|\nabla f_{i_{k+1}}(x_k) - \nabla f_{i_{k+1}}(x_*)\|^2\Big] + (1-p)\mathcal{D}_k$$

7. Show that

$$\mathbb{E}_k\Big[\|\nabla f_{i_{k+1}}(x_k) - \nabla f_{i_{k+1}}(x_*)\|^2\Big] \leq 2L\big(f(x_*) - f(x_k) - \langle \nabla f(x_k), x_* - x_k \rangle\big)\text{‘}$$

8. Chose values for the proof constants α and β that ensure

$$\mathbb{E}_k\big[F(x_{k+1}) + \beta\mathcal{D}_{k+1} + \frac{1}{2\gamma}\|x_{k+1} - x_k\|^2\big] \leq F(x_*) + \beta\mathcal{D}_k + \frac{1}{2\gamma}\|x_k - x_*\|^2 + \big(\frac{L}{2} - \frac{1}{2\gamma} + \frac{1}{2\alpha}\big)\|x_{k+1} - x_k\|^2$$

9. What range of step sizes ensures that $\big(\frac{L}{2} - \frac{1}{2\gamma} + \frac{1}{2\alpha}\big)\|x_{k+1} - x_k\|^2 \leq 0$?

10. Suppose γ satisfies the condition of the previous question and denote $\bar{x}_K = \frac{1}{K}\sum_{k=1}^{K}$. Show that

$$\mathbb{E}[F(\bar{x}_K) - F(x_*)] \leq \frac{\beta\mathcal{D}_0 + \frac{1}{2\gamma}\|x_0 - x_*\|^2}{K}$$

11. Compare with the convergence rate of other algorithms we have seen during the course.