Exam Stochastic Optimization – part I
30 January 2024

duration 1.5 hours, paper documents allowed

The goal of this exercise is to propose a unification of the proofs of stochastic gradient descent and randomized coordinate descent.
We consider the optimization problem

$$\min_{x \in \mathbb{R}^d} F(x)$$

where F is a convex and differentiable function whose gradient is L-Lipschitz continuous and we suppose that $\exists x^* \in \arg\min F$.
We assume that there exists a function $g : \mathbb{R}^d \times \xi$ and a random variable $\xi \sim \mathcal{D}$ such that

$$\mathbb{E}_{\xi\sim\mathcal{D}}[g(x, \xi)] = \nabla F(x) \ .$$

We shall also assume that there exists real numbers A, B and C such that

$$\mathbb{E}_{\xi\sim\mathcal{D}}[\|g(x, \xi)\|^2] \leq 2A(F(x) - F(x_*)) + B\|\nabla F(x)\|^2 + C \tag{1}$$

PART I

1. Suppose that $F(x) = \mathbb{E}_{\xi\sim D}[f(x, \xi)]$, $g(x, \xi) = \nabla f(x, \xi)$, $\mathbb{E}_{\xi\sim\mathcal{D}}[g(x, \xi)] = \nabla F(x)$ and $\mathbb{E}_{\xi\sim\mathcal{D}}[\|g(x, \xi) - \nabla F(x)\|^2] \leq \sigma^2$.

   Show that g satisfies (1) with $B = 1$ and values of A and C to be determined.

2. Suppose that $\xi \sim U(\{1, \ldots, d\})$ and $g(x, \xi) = d\nabla_\xi F(x)e_\xi$, which means it is a vector with one nonzero element equal to d times the partial derivative of F.

   Show that g satisfies (1) with $C = 0$.

3. Suppose that $F(x) = \mathbb{E}_{\xi\sim D}[f(x, \xi)]$ where $\forall\xi$, $\nabla f(x^*, \xi) = 0$ and $(x \mapsto f(x, \xi))$ is convex with a L-Lipschitz gradient. Show that $g = \nabla f$ satisfies (1) with $C = 0$.

PART II

Our goal now is to propose a unified analysis of the following generic stochastic gradient method where the convergence result would depend on A, B and C.

$$\begin{aligned} &\xi_{k+1} \sim \mathcal{D} \\ &x_{k+1} = x_k - \gamma_k g(x_k, \xi_{k+1}) \end{aligned}$$

where (γ_k) is a deterministic sequence of positive step sizes. We shall denote the conditional expection knowing the past as $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|\xi_1, \ldots, \xi_k]$.

1. Explain why $F(x^*) \geq F(x_k) + \langle \nabla F(x_k), x^* - x_k \rangle + \frac{1}{2L}\|\nabla F(x_k)\|^2$.
2. Show that $\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\gamma_k \langle g(x_k, \xi_{k+1}), x_k - x^* \rangle + \|x_{k+1} - x_k\|^2$.
3. Show that

   $$\mathbb{E}_k[\|x_{k+1}-x^*\|^2] \leq \|x_k-x^*\|^2+2(\gamma_k-\gamma_k^2 A)(F(x^*)-F(x_k))+(\gamma_k^2 B-\frac{\gamma_k}{L})\|\nabla F(x_k)\|^2+\gamma_k^2 C$$

4. Under what conditions on (γ_k) do we have

   $$\mathbb{E}\Big[\sum_{k=0}^{K-1} 2(\gamma_k - \gamma_k^2 A)(F(x_k) - F(x^*))\Big] \leq \|x_0 - x^*\|^2 + \sum_{k=0}^{K-1} \gamma_k^2 C \ ?$$

5. Give a point $\bar{x}_K$ and a sequence (γ_k) for which $F(\bar{x}_K) - F(x^*)$ converges to 0

PART III

Particularize the bound found in PART II to the 3 cases proposed in PART I. Comment on the quality of the bound found.