Continuous Optimization **Spring 2025**

Exercise 7: Stochastic Optimization

Lecturer: Aurelien Lucchi

Problem 1 (Stochastic Gradient Descent):

Consider an objective function with the following finite-sum structure:

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}, \xi_i) \tag{1}$$

where ξ_i is the $i-$th random variable (e.g. a datapoint in a given dataset in a machine learning setting). In this case, the computational cost of one GD step scales as $\mathcal{O}(d)$. One, obviously cheaper alternative is to only compute the update based on the gradient of one specific datapoint. This is the updated of *stochastic* gradient descent (SGD), which is arguably the most widely used optimizer in machine learning:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta\nabla f_i(\mathbf{x}_k), \quad i \in \{1, \dots, n\};\ \eta > 0. \tag{2}$$

In each iteration, the datapoint i is chosen uniformly at random such that $\mathbb{E}[\nabla f_i(\mathbf{x}_k)] = \nabla f(\mathbf{x}_k)$. We assume that the loss function f is smooth and μ-strongly convex.

1. In this regime, how many samples are left unseen in expectation after one epoch (n iterations)?
2. Show that given $\mathbf{x}_k$ and a constant step size $\eta = \frac{1}{2L}$, SGD does not converge to a critical point $\mathbf{x}^*$, i.e.

$$\mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2\right] \geq \frac{1}{(2L)^2}\mathbb{E}\left[\|\nabla f_i(\mathbf{x}_k)\|_2^2\right] \tag{3}$$

3. Name two possibilities to retain convergence.

Problem 2 (Convergence of the gradient norm):

Under the same finite-sum setting discussed in Problem 1, assume that f is μ-strongly-convex with L-Lipschitz continuous gradients, H-Lipschitz continuous Hessians, and bounded gradients ($\mathbb{E}_i \|\nabla f_i(\mathbf{x}_k)\|^2 \leq B^2$).

1. Show that $g(\mathbf{x}) \coloneqq \|\nabla f(\mathbf{x})\|^2$ is $\widetilde{L}$-smooth with $\widetilde{L} \coloneqq 2HB + 2L^2$.
2. Find the expression for the gradient of $g(\mathbf{x})$
3. Show that

$$g(\mathbf{x}_{k+1}) \leq g(\mathbf{x}_k) - 2\eta\langle\nabla^2 f(\mathbf{x}_k)\nabla f(\mathbf{x}_k), \nabla f_i(\mathbf{x}_k)\rangle + \frac{\widetilde{L}}{2}\eta^2\|\nabla f_i(\mathbf{x}_k)\|^2. \tag{4}$$

4. Using $\mathbb{E}_i \|\nabla f_i(\mathbf{x}_k)\|^2 \leq B^2$, show that

$$\mathbb{E}[g(\mathbf{x}_{k+1})] \leq (1 - 2\eta\mu)^{k+1} g(\mathbf{x}_0) + \underbrace{\sum_{j=0}^{k}(1 - 2\eta\mu)^j\frac{\widetilde{L}}{2}\eta^2 B^2}_{\text{Noise}}. \tag{5}$$

5. Bound the noise term and conclude that

$$\mathbb{E}[g(\mathbf{x}_k)] \leq (1 - 2\mu\eta)^k g(\mathbf{x}_0) + \frac{\widetilde{L}\eta}{4\mu}B^2. \tag{6}$$

Problem 3 (Convergence for PL functions):

Under the same finite-sum setting discussed in Problem 1, assume that f is μ-PL with L-Lipschitz continuous gradients and bounded gradients ($\mathbb{E}_i \|\nabla f_i(\mathbf{x}_k)\|^2 \leq B^2$).

1. Prove that

$$\mathbb{E}[f(\mathbf{x}_{k+1}) - f^*] \leq (1 - 2\eta_k \mu)[f(\mathbf{x}_k) - f^*] + \frac{LB^2\eta_k^2}{2}.$$

2. Let $\delta_f(k) \equiv k^2\mathbb{E}[f(\mathbf{x}_k) - f^*]$. Using $\eta_k = \frac{2k+1}{2\mu(k+1)^2}$, show that

$$\delta_f(k+1) \leq \delta_f(k) + \frac{LB^2}{2\mu^2},$$

3. Conclude that

$$\mathbb{E}[f(\mathbf{x}_{k+1}) - f^*] \leq \frac{LB^2}{2\mu^2(k+1)}.$$

Problem 4 (Programming exercise):

Write simple SGD code on least-square problem. You should compute the derivatives on paper, then implement them and run the algorithm, finally check the results by plotting the convergence curves. Also use a constant step size so as to see that SGD does not convergence to the minimum. Then compare with decreasing step size.