Labs
Optimization for Machine Learning
Spring 2024

Saarland University
CISPA Helmholtz Center for Information Security
Sebastian Stich
TAs: Yuan Gao & Xiaowen Jiang
https://cms.cispa.saarland/optml24/

Problem Set 3, April 30, 2024 (Stochastic Gradient Descent)

Stochastic Gradient Descent

Exercise 1. Sigma-star Assumption

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a μ-strongly convex function of the form $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$, $n \geq 1$, where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is a L-smooth and convex function. Let $\mathbf{x}^\star = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$.

1. *Show that, for any $i \in \{1, 2, ..., n\}$, and any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have:*

$$||\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})||^2 \leq 2L(f_i(\mathbf{x}) - f_i(\mathbf{y}) - \langle \nabla f_i(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle)$$

   (Hint: Let $g_i(\mathbf{x}) := f_i(\mathbf{x}) - \langle \mathbf{x}, \nabla f_i(\mathbf{y})\rangle$. Show that $\mathbf{y}$ is the minimizer of g_i and thus we have $g_i(\mathbf{y}) \leq g_i(\mathbf{x}) - \frac{1}{2L}||\nabla g_i(\mathbf{x})||^2$.)

2. *Consider the stochastic gradient oracle $\nabla f_i(\mathbf{x})$, where the index i is sampled uniformly at random with probability $\frac{1}{n}$. Prove that its second moment can be bounded as follows:*

$$\mathbb{E}_i||\nabla f_i(\mathbf{x})||^2 \leq 4L(f(\mathbf{x}) - f(\mathbf{x}^\star)) + 2\sigma_\star^2$$

   where $\sigma_\star^2 := \mathbb{E}||\nabla f_i(\mathbf{x}^\star)||^2$. (Hint: $||\nabla f_i(\mathbf{x})||^2 = ||\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^\star) + \nabla f_i(\mathbf{x}^\star)||^2$)

3. *Consider the stochastic gradient algorithm $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \nabla f_{i_t}(\mathbf{x}_t)$, where the index i_t is sampled uniformly at random. By using the result stated in the previous question, prove that (under appropriate conditions on γ) the one step progress of SGD can be upper bounded as follows (recall that f is assumed to be μ-strongly convex):*

$$\mathbb{E}_{i_t}||\mathbf{x}_{t+1} - \mathbf{x}^\star||^2 \leq (1 - \mu\gamma)||\mathbf{x}_t - \mathbf{x}^\star||^2 - \gamma(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) + 2\gamma^2\sigma_\star^2$$

4. *Compare this result with what you have seen in the course.*

Practical Implementation of SGD (Strongly Recommended)

In this exercise, you will train a linear regression model under a least square loss to predict a person's weight from their height. You will implement mini-batch stochastic gradient descent with various learning rates from step to step. Follow the Python notebook provided here:

colab.research.google.com/github/epfml/OptML_course/blob/master/labs/ex02/template/Lab 5-Stochastic Gradient Descent.ipynb