535520 Fall 2024: Optimization Algorithms **(Due: 2024/10/29, 21:00)**

Homework 1: GD, Momentum, and SGD

Submission Guidelines: Your deliverables shall consist of 2 separate files – (i) A PDF file: Please compile all your write-ups and your report into one .pdf file (photos/scanned copies are acceptable; please make sure that the electronic files are of good quality and reader-friendly); (ii) A zip file: Please compress all your source code into one .zip file. Please submit your deliverables via E3.

Problem 1 (Strong Convexity) (15 points)

As discussed in Lecture 4, the following are three equivalent characterization of strong convexity for a continuously differentiable function $f : X \to \mathbb{R}$:

- Condition 1: There exists some $\mu > 0$ such that for any $x, y \in X$,

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2. \tag{1}$$

- Condition 2: There exists some $\mu > 0$ such that for any $x, y \in X$,

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \mu \|x - y\|^2. \tag{2}$$

- Condition 3: If f is twice continuously differentiable, then there exists some $\mu > 0$ such that for any $x \in X$,

$$\nabla^2 f(x) - \mu I \succ 0. \tag{3}$$

Please verify the above conditions for strong convexity. (Hint: Note that you need to show that each condition holds if and only if f is strongly convex.)

Problem 2 (First-Order Methods for 1-Dimensional Optimization) (5+5+15=25 points)

Let us consider the following function

$$f(x) = \begin{cases} 25x^2, & x < 1 \\ x^2 + 48x - 24, & 1 \leq x \leq 2 \\ 25x^2 - 48x + 72, & x > 2. \end{cases} \tag{4}$$

(a) Show that f is 2-strongly convex and L-smooth. What is the smoothness constant L?

(b) Please find the global minimizer of f.

(c) Write a short Python program (use the filename **gd_hb_nag.py**) to run the following three first-order methods:

- Gradient descent (GD) with a constant step size $\eta = 1/50$.
- Heavy-ball momentum (HB) with gradient step size $\eta = 1/18$ and momentum step size $\theta = 4/9$.
- Nesterov's accelerated gradient (NAG) with gradient step size $\eta = 1/50$ and momentum step size $\theta = 2/3$.

Suppose the initial point $x_0 = 3$ and the initial momentum is 0. (i) Plot the sub-optimality gap in function values (i.e., $f(x_t) - f(x^*)$) vs number of iterations t for the above three methods. (ii) Plot the distance from the optimal solution (i.e., $|x_t - x^*|$) vs number of iterations t for the above three methods. For GD and HB, please

also plot the upper bounds on $|x_t - x^*|$ discussed in the lectures. (iii) Discuss whether the actual performance is closely aligned with the upper bounds.

Problem 3 (Nesterov's Accelerated Gradient) (5+5+5+5+5=25 points)

Recall from Lecture 9 that we have reformulated Nesterov's accelerated gradient as an ordinary differential equation and established the convergence. In this problem, let us formally prove the convergence rate of the original form of the Nesterov's method step by step. Recall that in each iteration, Nesterov's method with a step size $\eta_t = 1/L$ proceeds as follows: Given any initial points x_1, y_1 that satisfy $x_1 = y_1$,

$$x_{t+1} = y_t - \frac{1}{L}\nabla f(y_t) \tag{5}$$

$$y_{t+1} = x_{t+1} - \frac{1-\theta_t}{\theta_{t+1}}(x_{t+1} - x_t) \tag{6}$$

$$\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}, \quad \theta_0 = 0. \tag{7}$$

Note that (7) implies that $\theta_{t+1}^2 - \theta_{t+1} - \theta_t^2 = 0$.

(a) Show that for any $x, y \in \mathbb{R}^d$, we have

$$f\big(y - \frac{1}{L}\nabla f(y)\big) - f(x) \le -\frac{1}{2L}\|\nabla f(y)\|^2 - \nabla f(y)^\top (x - y). \tag{8}$$

(Hint: Use convexity and smoothness)

(b) By leveraging the result in (a), show that

$$f(x_{t+1}) - f(x_t) \le -\frac{L}{2}\|x_{t+1} - y_t\|^2 + L(x_{t+1} - y_t)^\top (x_t - y_t). \tag{9}$$

Similarly, by leveraging the result in (a), show that

$$f(x_{t+1}) - f(x^*) \le -\frac{L}{2}\|x_{t+1} - y_t\|^2 + L(x_{t+1} - y_t)^\top (x^* - y_t). \tag{10}$$

(c) By adding $\theta_t(\theta_t - 1)$ times of (9) and θ_t times of (10), show that

$$\theta_t^2 \Delta_{t+1} - \theta_{t-1}^2 \Delta_t \le -\frac{L}{2}\big(\|\theta_t(x_{t+1} - y_t)\|^2 + 2\theta_t(x_{t+1} - y_t)^\top \underbrace{(\theta_t y_t - (\theta_t - 1)x_t - x^*)}_{=:\phi_t}\big) \tag{11}$$

(d) By completing the square for the RHS of (11) and using the update rule of Nesterov's method, show that

$$\theta_t^2 \Delta_{t+1} - \theta_{t-1}^2 \Delta_t \le \frac{L}{2}\Big(\|\phi_t\|^2 - \|\phi_{t+1}^2\|\Big) \tag{12}$$

(e) Finally, by using an induction argument, show that the convergence rate of Nesterov's method is

$$f(x_t) - f(x^*) \le \frac{2L\|x_1 - x^*\|^2}{t^2}. \tag{13}$$

Problem 4 (Gurobi Optimizer) (10 points)

In this problem, you have the nice opportunity to play with the useful Gurobi optimization solver. You can install the Python API of Gurobi by following the instruction at:
https://support.gurobi.com/hc/en-us/articles/4534161999889-How-do-I-install-Gurobi-Optimizer.

- Please first take a look at the attached Gurobi example **QP.ipynb**. More examples can be found at

https://www.gurobi.com/documentation/current/examples/example_code.html.

- Based on the example code **QP.py**, try to write a Python script to leverage Gurobi to solve the following *Markowitz Portfolio Optimization* problem:

$$\text{Minimize}_x \quad -p^\top x + \mu \cdot x^\top \Sigma x \tag{14}$$
$$\text{subject to} \quad \mathbf{1}^\top x = 1, x \succeq 0, \tag{15}$$

where $x \in \mathbb{R}^4$, $\mathbf{1} = [1,1,1,1]^\top$, $p = [0.12, 0.1, 0.07, 0.03]^\top$, and $\Sigma \in \mathbb{R}^{4\times 4}$ is a symmetric matrix defined as

$$\Sigma = \begin{pmatrix} 0.2 & -0.03 & 0 & 0 \\ -0.03 & 0.1 & -0.02 & 0 \\ 0 & -0.02 & 0.05 & 0 \\ 0 & 0 & 0 & 0.01 \end{pmatrix} \tag{16}$$

Please report the optimal function values $f(x^*)$ and optimal solutions x^* under different parameters $\mu \in \{0, 0.1, 1.0, 2.0, 5.0, 10.0\}$.

Problem 5 (SGD and SVRG) (35 points)

In Lectures 6-7, we learned two useful gradient-based algorithms, SGD and SVRG, as well as their convergence analysis. Let us compare these two algorithms empirically in terms of convergence behavior. Specifically, we will evaluate SGD and SVRG in a way similar to Figure 2 of the SVRG paper (https://papers.nips.cc/paper/2013/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf) on MNIST dataset under both convex and non-convex loss functions. To facilitate gradient computation, you may write your code in either PyTorch or TensorFlow (though the sample code presumes PyTorch framework). If you are a beginner in learning the deep learning framework, please refer to the following tutorials:

- PyTorch: https://pytorch.org/tutorials/
- Tensorflow: https://www.tensorflow.org/tutorials

For the deliverables, please submit the following:

- Technical report: Please summarize all your experimental results in 1 single report (and please be brief)
- All your source code

(a) We start from the logistic regression with the convex loss (see Section 5 of SVRG paper for more details).

- Read through **sgd.py**, **svrg.py**, **train.py**, and **utils.py** and then implement the member functions of several classes (e.g., **SVRG**) as well as several other helper functions (e.g., **MNIST_logistic**).
- Moreover, plot figures similar to Figures 2(a)-(b) in the SVRG paper (Note: The x-axis of Figures 2(a)-(b) is the computational cost measured by the number of gradient computations divided by the size of the dataset). To create a figure like Figure 2(b) in the SVRG paper, you would need to find the primal optimal value (e.g., by running GD for sufficiently many iterations).

(b) Based on (a), redo the same things under a single-layer neural network classifier.

Please briefly summarize your results in the report and document all the hyperparameters (e.g. learning rates and batch size) of your experiments.