535526 Fall 2022: Optimization Algorithms (**Due: 2022/11/25, 21:00**)

Homework 1: Duality, SGD, PGD, and Variance Reduction

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 (Slater's Condition) (5+5+5=15 points)

In this problem, let us manage to prove the well-known Slater's condition for strong duality. For ease of exposition, we focus on the constrained problems with one inequality constraint (and the proof can be extended to multiple constraints with some additional technicalities). Specifically, consider the following problem:

$$\min_{x\in\mathcal{D}} f(x), \quad \text{subject to } g_1(x) \leq 0, \tag{1}$$

where both $f(x)$ and $g_1(x)$ are convex functions. Let $p^* > -\infty$ denote the primal optimal value. Define two helper sets as follows:

$$\mathcal{A} := \{(u,t)|\, \exists x \in \mathcal{D} \,\text{such that } f(x) \leq t, g_1(x) \leq u\}, \tag{2}$$

$$\mathcal{B} := \{(0,s)|\, s < p^*\}. \tag{3}$$

(a) Show that $\mathcal{A}$ and $\mathcal{B}$ are two disjoint non-empty convex sets.

(b) Based on (a), we know by Separating Hyperplane Theorem, there exists real numbers $\tilde{\lambda}$, μ, and α such that

$$\tilde{\lambda}u + \mu t \geq \alpha, \quad \forall (u,t) \in \mathcal{A}, \tag{4}$$

$$\tilde{\lambda}u + \mu t \leq \alpha, \quad \forall (u,t) \in \mathcal{B}. \tag{5}$$

Show that $\mu \geq 0$, $\tilde{\lambda} \geq 0$, and $\tilde{\lambda}g_1(x) + \mu f(x) \geq \alpha \geq \mu p^*$.

(c) Based on (b), prove that strong duality holds under the Slater's condition. (Hint: Consider the two cases $\mu > 0$ and $\mu = 0$ separately)

Problem 2 (Duality Gap and Strong Duality) (5+5+5=15 points)

In this problem, let us practice how to derive the duality gap and verify that strong duality does not necessarily hold for convex problems. Consider the following optimization problem:

$$\min_{x,y} \quad \exp(-x), \quad \text{subject to } x^2/y \leq 0, \tag{6}$$

with variables x and y, and domain $\mathcal{D} = \{(x,y) : y > 0\}$.

(a) Verify that this is a convex optimization problem and find the optimal value.

(b) Write down the Lagrange dual problem (with Lagrange multiplier λ), and then find the optimal dual solution λ^* and the corresponding dual optimal value d^*. What is the duality gap?

(c) Does Slater's condition hold for this problem?

Problem 3 (Convergence of PGD for Convex and Smooth Problems) (5+5+5+5+5=25 points)

As discussed in Lec 7, we mention that for a convex and L-smooth function, the convergence rate of PGD is

$$f(x_t) - f(x^*) \leq \frac{3L\|x_0 - x^*\| + \big(f(x_0) - f(x^*)\big)}{t+1}. \tag{7}$$

In this problem, let us prove this result formally in a step-by-step manner.
(a) To begin with, let us show the following lemma: For any $x, z \in C$, let $\bar{x} = \prod_C(x - \frac{1}{L}\nabla f(x))$ and $g_C(x) = L(x - \bar{x})$ (note that here we basically reuse the same notation as in our lecture slides). Then, we have

$$f(z) \geq f(\bar{x}) + g_C(x)^\top (z - x) + \frac{1}{2L}\|g_C(x)\|^2. \tag{8}$$

(Hint: Consider $f(z) - f(\bar{x}) = (f(z) - f(x)) - (f(\bar{x}) - f(x))$ and then utilize the convexity and smoothness conditions)
(b) By using the result in (a), show that the one-step improvement can be written as

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

(c) Next, let us connect $\|g_C(x_t)\|$ to the objective function $f(x_t)$: Show that

$$\|g_C(x_t)\| \geq \frac{f(x_{t+1}) - f(x_t)}{\|x_t - x^*\|}. \tag{10}$$

(Hint: Find a proper way to apply the result in (a) and use Cauchy-Schwarz inequality)

(d) Define the sub-optimality gap at the t-th iteration as $\Delta_t := f(x_t) - f(x^*)$. By using the results in (a)-(c), show that $\Delta_{t+1} - \Delta_t \leq \frac{-\Delta_{t+1}^2}{2L\|x_0 - x^*\|^2}$.

(e) Finally, by using the result in (d), use an induction argument to show the convergence rate of PGD in (7).

Problem 4 (SGD and SVRG) (50 points)

In Lec 5-7, 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.