CSE 597: LARGE-SCALE MACHINE LEARNING
MATHEMATICAL FOUNDATIONS AND APPLICATIONS

FINAL EXAM

FALL 2021

`Name:...........................` `PSU ID:.................. @psu.edu`

There are four problems with total score of 60 with 10 bonus points (exam will be graded for 50 points). Please write all the details in your derivations and be as rigorous as possible.

Q1: ______ / 10

Q2: ______ / 20

Q3: ______ / 20

Q4: ______ / 10

Total ______ / 50

Result:

Question 1 (10 pts). Consider the binary classification problem, where we are given a training set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$. In SVM for binary classification we would like to find the minimizer of the following problem

$$\min_{\boldsymbol{w}} \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{(\boldsymbol{x},y)\in S} \ell(\boldsymbol{w}; (\boldsymbol{x}, y))$$

where

$$\ell(\boldsymbol{w}; (\boldsymbol{x}, y)) = \max\{0, 1 - y\langle \boldsymbol{w}, \boldsymbol{x}\rangle\}$$

is the Hinge loss.

By examining the dual form of the SVM problem and using the strong duality theorem (i.e., if $\boldsymbol{w}_*$ and $\boldsymbol{\alpha}_*$ are primal and dual solutions, respectively, then it holds that $\boldsymbol{w}_* = \sum_{i=1}^n \alpha_{*,i} y_i \boldsymbol{x}_i$), show that $\|\boldsymbol{w}_*\| \leq 1/\sqrt{\lambda}$.

Question 2 (20 pts)**.** Consider a distributed optimization setting where p machines aim to jointly optimize the following ERM objective

$$f(\boldsymbol{w}) = \frac{1}{p}\sum_{i=1}^{p} f_i(\boldsymbol{w}),$$

under the orchestration of a central server where $f_i(\boldsymbol{w})$ is the training (empirical) loss over data at ith device.

A simple distributed optimization algorithm (synchronous SGD) to optimize above objective is as follows.Starting at an initial solution $\boldsymbol{w}_1$, at each iteration t, the server sends the global model $\boldsymbol{w}_t$ to all p devices, each device computes local unbiased stochastic gradient $\boldsymbol{g}_{i,t}$ at $\boldsymbol{w}_t$ ($\mathbb{E}[\boldsymbol{g}_{i,t}] = \nabla f(\boldsymbol{w}_t)$) and sends it back to the server. Then server aggregates the stochastic gradients from all devices by averaging $\boldsymbol{g}_t = (1/p)\sum_{i=1}^{p} \boldsymbol{g}_{i,t}$ and updates the global model by $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{g}_t$. And this proceeds for T iterations. Since at each step, p stochastic gradients are computed in parallel, we can show that for $O(T)$ communication rounds, we can achieve an $O(\frac{1}{pT})$ convergence rate for smooth and strongly convex functions (follows from the analysis of mini-batch SGD in lecture with batch size $B = p$).

In this question we consider a variant of above distributed optimization algorithm and would like to establish its convergence rate. In the modified algorithm, instead of using a single sample (i.e., fixed batch size) to compute local stochastic gradients at all iterations, the devices are allowed to utilize growing mini-batches with growing coefficient $\rho > 1$ at different steps. Specifically, consider the following distributed optimization algorithm with growing (dynamic) mini-batches:

inputs: p, T, $\boldsymbol{w}_1$, initial mini-batch size B_1, mini-batch growing coefficient $\rho > 1$
initialize: $t = 1$
while $\sum_{s=1}^{t} B_s \le T$ **do**
    Each device i samples B_t samples IID and computes unbiased stochastic gradient $\boldsymbol{g}_{i,t}$ at $\boldsymbol{w}_t$
    Server aggregates gradients from all devices via averaging: $\boldsymbol{g}_t = \frac{1}{p}\sum_{i=1}^{p} \boldsymbol{g}_{i,t}$
    Server updates the model by $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{g}_t$ and broadcast to all devices
    Set $B_{t+1} = \lfloor \rho^t B_1 \rfloor$ (growing the batch size)
    Update $t \leftarrow t + 1$

For above algorithm,

(a) Assume the global objective function $f(\cdot)$ is α-strongly convex and β-smooth. Show that if we choose $\eta < \frac{1}{\beta}$ in above algorithm, then for all $t \in \{1, 2, \ldots, \}$, we have

$$\mathbb{E}\left[f\left(\boldsymbol{w}_{t+1}\right) - f^*\right] \le (1-\nu)\mathbb{E}\left[f\left(\boldsymbol{w}_t\right) - f^*\right] + \frac{\eta(2-\beta\eta)}{2pB_t}\sigma^2$$

where f^* is the global minimum, σ is the variance of stochastic gradients, and $\nu = \frac{1}{2}\eta\alpha(1-\beta\eta)$ satisfies $0 < \nu < 1$

(b) Use above result and show that with $t = O(\log T)$ communication rounds we can achieve an $O\left(\frac{1}{pT}\right)$ convergence rate (i.e., using dynamic mini-batches we can reduce the number of communications from $O(T)$ in vanilla distributed SGD with $B = 1$ to $O(\log T)$ while achieving the same convergence rate).

Question 3 (20 pts)**.** This problem is mostly a reading exercise. Consider a dataset $S = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$. Assume that the data is linearly separable by margin $\gamma \in (0, 1]$ defined as $\gamma = \max_i y_i \frac{\langle \boldsymbol{w}_*, \boldsymbol{x}_i \rangle}{\|\boldsymbol{w}_*\| \|\boldsymbol{x}_i\|}$ where $\boldsymbol{w}_*$ is the optimal classifier.

Let $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ be a random Gaussian matrix with $\boldsymbol{M}_{ij} = \frac{1}{\sqrt{m}} \mathcal{N}(0, 1)$. Show that for any $\delta, \varepsilon \in (0, 1)$ if $m > \mathcal{O}\left(\frac{1}{\varepsilon^2} \ln \frac{n}{\delta}\right)$, the dataset $S' = \{(\boldsymbol{M}\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{M}\boldsymbol{x}_n, y_n)\}$ is linearly seperable with margin $\gamma - \frac{2\varepsilon}{1-\varepsilon}$. (What you need to do is to read this paper "Is margin preserved after random projection?", and simplify/rewrite the proof.)

Question 4 (10 pts). Consider the function $f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x})$. Let $\nabla f_S(\boldsymbol{x}) = \frac{1}{|S|}\sum_{i \in S} \nabla f_i(\boldsymbol{x})$ be the stochastic gradient of function computed at a subset $S \subseteq \{1, 2, \cdots, n\}$ sampled uniformly at random. By using concentration inequalities, show that what would be the size of S (batch size) to guarantee that

$$\mathbb{P}\left[\|\nabla f_S(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|_2^2 \leq \varepsilon\right] \geq 1 - \delta$$

for any $\varepsilon, \delta \in (0, 1)$ and $\boldsymbol{x} \in \mathbb{R}^d$.

Good luck.