CS7545, Spring 2023: Machine Learning Theory - Practice Exam

Spring 2023 Due: ——

Your Name: ________________________________

GT Account Username: ________________________________

Note: you may skip ONE problem on this exam and still receive full credit (problems with multiple parts count as a single problem). Some extra credit will be given if all 6 problems are solved.

1) **Deviations.** Let n, m be positive integers. Assume you have $n \times m$ independent random variables, $X_{i,j}$ for $i \in [n]$ and $j \in [m]$, bounded in the range $[-1, 1]$, and assume for all i and j that $X_{i,j}$ is distributed according to D_i with mean μ_i, assume that the μ_i's are unique, and for some known $\gamma > 0$ assume that $|\mu_i - \mu_j| \geq \gamma$ for all $i \neq j$.

(a) Find a reasonable bound, in terms of n, m, and γ, on the probability of the event that

$$\arg\max_{i \in [n]} \frac{1}{m} \sum_{j=1}^{m} X_{i,j} \neq \arg\max_{i \in [n]} \mu_i$$

(b) Choose any value for m (as a function of n, γ) so that the above event occurs with probability no more than $\frac{1}{n^2}$. (You may assume n is sufficiently large.)

2) **Combinatorics.** Let $\mathcal{X}$ be the set $\{0,1\}^n$, let $\mathcal{Y}$ be the set $\{0,1\}$ and Let $\mathcal{F}$ be some set of functions mapping $\mathcal{X} \to \mathcal{Y}$ which *depend only on the total number of 1's in the input.* Let k be an arbitrary positive integer, and imagine we have k points $x_1, \ldots, x_k \in \mathcal{X}$. Also, assume $k \gg n$.

Give a bound on the cardinality of the set of k-length vectors $U = \{(f(x_1), \ldots, f(x_k)) : f \in \mathcal{F}\}$. (This bound should be significantly better than the trivial bounds of 2^k or $|\mathcal{F}|$.)

3) **A lot of experts!.** Let's say we want to predict whether the stock market will go up or down on each of the next T days. You've learned about the Exponential Weights Algorithm for combining the advice of experts. Here's an idea: let's create an expert for every T-length binary sequence. That is, we have a very large pool of experts, and any given expert might say something like "the stock market goes up on day 1, down on day 2, down on day 3, ..." Indeed, one of these experts will be correct on all T days.

Claim: Imagine we run EWA on this pool of experts. We would be able to get sublinear regret on predicting the stock market, and hence we can get an edge on our investments!

Is the above claim correct? Argue why or why not.

4) **Vanilla Optimization via Online Gradient Descent.** We are given a convex and bounded set $\mathcal{K} \subset \mathbb{R}^d$. In Online Convex Optimization we observe a sequence of convex and lipschitz loss functions $\ell_1, \ell_2, \ldots, \ell_T$ from $\mathcal{K} \to \mathbb{R}$, and we must choose a sequence of points $x_1, x_2, \ldots$ online in order to minimize regret, defined as $\sum_{t=1}^{T} \ell_t(x_t) - \min_{x \in \mathcal{K}} \ell_t(x)$. Recall the Online Gradient Descent (OGD) algorithm: it selects $x_1 \in \mathcal{K}$ arbitrarily and then, at every time, t performs the update:

$$\tilde{x}_{t+1} = x_t - \eta \nabla \ell_t(x_t) \qquad \text{and} \qquad x_{t+1} = \mathrm{Proj}_{\mathcal{K}}(\tilde{x}_{t+1}),$$

where $\mathrm{Proj}_{\mathcal{K}}(y) := \arg\min_{y' \in \mathcal{K}} \|y' - y\|_2$. Recall that OGD has a regret bound of $O(\sqrt{T})$.

Let us consider a simpler (non-online) problem: finding an (almost-)optimal solution of the objective $\min_{x \in \mathcal{K}} G(x)$, where $G(x)$ is some convex and lipschitz function on $\mathcal{K}$. Here is a typical algorithm for solving this problem: first pick a point $x_1 \in \mathcal{K}$, and then do gradient descent updates,

$$\text{For } t = 1, \ldots, T, \qquad \tilde{x}_{t+1} = x_t - \eta \nabla G(x_t) \qquad \text{and} \qquad x_{t+1} = \mathrm{Proj}_{\mathcal{K}}(\tilde{x}_{t+1}).$$

Use the regret bound for OGD to prove that the average iterate, $\bar{x}_T := \frac{1}{T}\sum_{t=1}^{T} x_t$, is a $O(\frac{1}{\sqrt{T}})$-approximate solution to the objective $G(\cdot)$. (That is, prove that $G(\bar{x}_T) - \min_{x \in \mathcal{K}} G(x) = O(1/\sqrt{T})$).

5) **VC Dimension.** What is the VC dimension of axis aligned squares in the plane $\mathbb{R}^2$? Here, a "square" is a function $h(x)$ parameterized by four values a_1, a_2, b_1, b_2 such that $a_2 - a_1 = b_2 - b_1$ which outputs $h(x) = 1$ on $x = (x_1, x_2)$ when both $a_1 \leq x_1 \leq a_2$ and $b_1 \leq x_2 \leq b_2$, and outputs 0 otherwise.

6) **Tuning Parameters (Continued).** In Problem 1 of HW3, we tuned the parameters of different upper bounds to make them to be tightest possible. In the following, we will explore an additional example of this process.

(a) Performance$(\mathcal{A}; N, T, \eta, \epsilon) \leq \frac{\log N}{\eta} + \frac{\eta T}{\epsilon^2} + \epsilon T$.