CS7545, Spring 2023: Machine Learning Theory - Practice Exam Solutions

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

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

Solution: This is very similar to the HW 1 problem about biased coin. Without loss of generality assume μ_n is the maximum. Consider a fixed $k \neq n$. Consider the random variables $Z_j = X_{k,j} - X_{n,j}$ for $j \in [m]$. We can see that these are independent random variables between bounded in the range $[-2, 2]$ and $\mathbb{E}\left[\sum_{j=1}^{m} Z_j\right] = (\mu_k - \mu_n)m$

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{m}\sum_{j=1}^{m} X_{k,j} > \frac{1}{m}\sum_{j=1}^{m} X_{n,j}\right] &= \mathbb{P}\left[\sum_{j=1}^{m} Z_j > 0\right] \\
&= \mathbb{P}\left[\sum_{j=1}^{m} Z_j - (\mu_k - \mu_n)m > (\mu_k - \mu_n)m\right] \\
&= \mathbb{P}\left[\sum_{j=1}^{m} Z_j - (\mu_k - \mu_n)m > -(\mu_n - \mu_k)m\right] \\
&\leq \exp\left(\frac{-2(\mu_n - \mu_k)^2 m^2}{16m}\right) \\
&\leq \exp\left(\frac{-\gamma^2 m}{8}\right)
\end{aligned}$$

Now

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

As we chose a fix $k \neq n$, we can union bound over all values of $k \neq n$ to get

$$\mathbb{P}\left[\max_{i \neq n} \frac{1}{m}\sum_{j=1}^{m} X_{i,j} > \frac{1}{m}\sum_{j=1}^{m} X_{n,j}\right] \leq (n-1)\exp\left(\frac{-\gamma^2 m}{8}\right)$$

(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.)

Solution: We want

$$n \exp\left(\frac{-\gamma^2 m}{8}\right) < \frac{1}{n^2}$$

which gives us

$$m > \frac{24 \log n}{\gamma^2}$$

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 $|\mathcal{F}|$.)

Solution: There are two solutions to this problem.

Solution 1: Since every hypothesis in the hypothesis set $\mathcal{F}$ depends only on the number of 1's in the input, it's easy to see that the VC Dimentsion of $\mathcal{F}$ is $n+1$. To show that a set of size $n+1$ can be shattered, select the set $x_0, x_1, \cdots, \mathbf{x}_n$ such as x_i has i 1's in it. Whatever labelling we have of this set, we can have an $f \in \mathcal{F}$ which generates that labelling. To show that that VC dimension is $n+1$, we have to show there is no set of size greater than $n+1$ points which can be shattered by $\mathcal{F}$. We can see that by pigeon hole principle, any set of size $n+2$ or greater will have at least two points such that the number of 1's in the two points are equal. Hence if we label one them as 0 and the other as !, there is no function $f \in \mathcal{F}$ which can generate this labelling. Hence the VC dimension of $\mathcal{F}$ is $n+1$.

Now clearly maximum value of size of the set U is the growth function of $\mathcal{F}$ defined at k, i.e $\Pi_{\mathcal{F}}(k)$. By Sauer's Lemma,

$$\Pi_{\mathcal{F}}(k) \leq (\frac{ek}{n+1})^{n+1}$$

This bound is much smaller than 2^k when $k \gg n$

Solution 2: Even though the problem asks you to get something better than $|\mathcal{F}|$, it's easy to see that since the function value depends only on the number of $1's$, the number of unique functions in $|\mathcal{F}|$ can be atmost 2^{n+1}. Hence the size of the set U can be atmost 2^{n+1} which is a better bound that the one proved using Sauer's lemma when $ke > 2n+2$.

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.

Solution: A key observation for this problem is: the performance of the market can be up or down everyday. So to ensure that there always exists one expert which is correct in all T days, there should be at least $N = \Omega(2^T)$ experts. Considering the fact that the regret of EWA is $O(\sqrt{T \ln N})$, we have

$$\text{Regret}_T = O(\ln N + \sqrt{L_T^* \ln N}) = \Omega(\ln 2^T + \sqrt{T \ln(2^T)}) = O(T),$$

which is not sub-linear.

Note: it was pointed out by a student in class that this is only an upper bound... which is true! What we are saying here is that the bound is vacuous, although it is of course possible that some other bound could be better. But indeed, one can show that there is a lower bound in this case, and the regret will be linear in T regardless of the analysis, but that is beyond the intended scope of this problem!

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:

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

where $\text{Proj}_{\mathcal{K}}(y) := \arg\min_{y' \in \mathcal{K}} \|y' - y\|_2$. 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,

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

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 $G(\bar{x}_T) - \min_{x \in \mathcal{K}} G(x) = O(1/\sqrt{T})$).

Solution:

To use OGD to solve convex optimization, in each round t, we pass $\ell_t = G$ to OGD. Since G is convex and lipschitz, the regret bound for OGD holds. Using the regret bound of OGD, we get:

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x \in \mathcal{K}} (\sum_{t=1}^{T} \ell_t(x)) \leq O(\sqrt{T})$$

$$\sum_{t=1}^{T} G(x_t) - \min_{x \in \mathcal{K}} (\sum_{t=1}^{T} G(x)) \leq O(\sqrt{T})$$

$$\sum_{t=1}^{T} G(x_t) - T \cdot \min_{x \in \mathcal{K}} (G(x)) \leq O(\sqrt{T})$$

$$\frac{1}{T} \sum_{t=1}^{T} G(x_t) - \min_{x \in \mathcal{K}} (G(x)) \leq O(\frac{1}{\sqrt{T}})$$

$$G(\frac{1}{T} \sum_{t=1}^{T} x_t) - \min_{x \in \mathcal{K}} (G(x)) \leq O(\frac{1}{\sqrt{T}})$$

Where the last step is using Jensen's inequality as G is convex.

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.

Solution: The hypothesis set is axis aligned squares. We can easily see that any three points that form a triangle can be easily shattered by this set. We claim that the VC Dimension is 3. To show that VC dimension is 3, we need to show that any set of size 4 or more cannot be shattered by this hypothesis set. Observe that if a set of 4 points cannot be shattered, then a set of more than 4 points cannot be shattered as well.

Let $ABCD$ be the 4 points, A is left most point, B is top most, C is right most, D is bottom (assume all x's y's are different for the 4 points, it's easier to see that when two or more points lie on the same x or y coordinate, it becomes more difficult to shatter that set). If the vertical distance between B and D, i.e $|y_B - y_D|$ is more than the horizontal distance of A and C, i.e $|x_A - x_C|$, then any square that has B and D must have at least one of A or C. Hence we cannot generate a labelling with B and D as 1 and both A and C as 0. Similarly if horizontal distance between $|x_A - x_C|$ is more than $|y_B - y_D|$, then every square containing A and C must contain at least one of B or D. Hence in this case we cannot generate a labelling with A and C as 1 and both B and D as 0. Hence we cannot shatter 4 points and the VC dimension of the hypothesis class is 3.

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

Solution: For simplicity, we let $f(\eta, \epsilon) = \frac{\log N}{\eta} + \frac{\eta T}{\epsilon^2} + 2\epsilon T$, which is an upper bound of the original objective. Taking the derivative and setting to zero, we have

$$\frac{\partial f}{\partial \epsilon} = -2(\eta T)\epsilon^{-3} + 2T = 0,$$

which implies $\epsilon^* = \eta^{\frac{1}{3}}$. It is obvious that this is the global minimum. Plugging into f, we have

$$f(\eta; \epsilon^*) = \frac{\log N}{\eta} + 3\eta^{\frac{1}{3}} T.$$

We can again take the derivative w.r.t. η, then we have

$$\frac{\partial f}{\partial \eta} = -\frac{\log N}{\eta^2} + \eta^{-\frac{2}{3}} T = 0$$

and we get

$$\eta^* = \left(\frac{\log N}{T}\right)^{\frac{3}{4}}.$$

We have

$$f(\eta^*, \epsilon^*) = 4(\log N)^{\frac{1}{4}} T^{\frac{3}{4}} = O\left((\log N)^{\frac{1}{4}} T^{\frac{3}{4}}\right),$$

which is an upper bound of our original objective.