Name: _____

Question:	1	2	3	4	Total
Points:	20	10	10	5	45
Score:					

E1 245 – Online Prediction and Learning, Aug-Dec 2019 – Final Exam

Instructions

- Write your name on top of this question sheet, attach your solution sheets to it and return everything together.
- The total time for this exam is 3 hours. The exam has 4 questions, for a total of 45 points. Partial points will be awarded, so please attempt as much as you can.
- Feel free to use your notes for this exam.
- You can be liberal with manipulating universal constants; errors in this respect will be tolerated.
- Academic dishonesty will not be tolerated.

Useful facts and definitions:

- $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ for $p > 0$, and $a \geq b \geq 1 \Rightarrow \|x\|_a \leq \|x\|_b \, \forall x \in \mathbb{R}^d$.
- The function $x \mapsto \|x\|_p^2$ is $2(p-1)$-strongly convex with respect to the $\|\cdot\|_p$ norm for any $1 < p \leq 2$.
- A convex, differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ satisfying $\|\nabla f(x)\|_q \leq \ell$ is ℓ-Lipschitz continuous with respect to $\|\cdot\|_p$, with $1 < p < \infty$ and $q = p/(p-1)$.
- Chi-square upper bound for KL divergence: For probability distributions p and q over a finite alphabet $[n]$, $\mathrm{KL}(p||q) \leq D_{\chi^2}(p, q) = \sum_{i=1}^n \frac{(p_i - q_i)^2}{q_i}$.
- Bernoulli KL divergence: For $0 < p, q < 1/2$, $D(\mathrm{Ber}(p)||\mathrm{Ber}(1-q)) \geq \log \frac{1}{2.4 \max\{p,q\}}$.
- A useful change of measure inequality (alternative to the one we saw in class): For probability distributions P and Q on the same space and an event E, $P[E] + Q[E^c] \geq \frac{1}{2} \exp(-KL(P||Q))$.

1. *Worst case regret for Explore-Then-Commit*
   Consider the Explore-Then-Commit bandit algorithm[1], that we studied in class, run on a 2-armed bandit with Bernoulli-distributed rewards and parameters (means) $\mu_1, \mu_2 \in [0, 1]$, a time horizon of T and an initial exploration phase of ϵT rounds with $\epsilon \in [0, 1]$. Let $\Delta = \mu_1 - \mu_2 > 0$.

   (a) **(5 points)** Show that there is a choice of ϵ, depending only on the time horizon T and not depending on Δ, under which the regret of the algorithm is bounded above by $c(\Delta + T^{2/3})$, where $c > 0$ is a universal constant.[2]

   (b) **(5 points)** Now suppose the commitment time is allowed to be data-dependent, which means the algorithm explores each arm alternately until some condition based on the observations is met, after which it commits to a single arm for the remainder. Design a condition such that the regret of the resulting algorithm can be bounded by $c'\left(\Delta + \frac{\log T}{\Delta}\right)$ where c' is a universal constant. Note: Your condition should only depend on the observed rewards and the time horizon, and not on μ_1, μ_2 or Δ.

   (c) **(10 points)** Lower bound for general non-adaptive-explore-then-exploit algorithms. Consider a general two-phase algorithm that operates as follows: The first phase lasts for $f(T)$ rounds for a fixed function f (i.e., its length depends only on T and not on Δ). The algorithm pre-decides the arm to pull at each of the $f(T)$ rounds (possibly with internal randomness) in this phase, depending on only T again. Based on only the data D collected in phase 1, the algorithm picks an arm I which is played in each of the remaining rounds $f(T) + 1, \ldots, T$ (phase 2).

   Show that for any such two-phase algorithm, there exists a 2-armed stochastic bandit instance with Bernoulli-distributed rewards for which the algorithm incurs at least $cT^{2/3}$ regret where c is some universal constant. [Hint: You can try considering two instances with arm means $\left(\frac{1}{2}, \frac{1}{2} - \epsilon\right)$ and $\left(\frac{1}{2}, \frac{1}{2} + \epsilon\right)$. Either the usual change of measure inequality from class or the alternative inequality provided here, along with the explicit form of the regret, should give suitable ways of adjusting ϵ for getting large regret on one of these two instances.]

2. *A new algorithm for prediction with expert advice*
   This problem presents an alternative to the Exponential Weights algorithm for prediction with experts with the same optimal worst case regret.

   Suppose Follow The Regularized Leader (FTRL) is run with the decision space being the m-simplex $\Delta_m = \{x \in \mathbb{R}^m : \forall i\, x_i \geq 0, \sum_i x_i = 1\}$, the regularizer being $R(x) = \frac{1}{\eta}\|x\|_p^2$ where $p = \frac{\log(m)}{\log(m)-1}$, and a sequence of T linear loss functions on $\mathbb{R}^m$ with coefficients bounded in $[0, 1]$.

   (a) **(2 points)** Write a regret bound for the algorithm (w.r.t. the single best point in Δ_m in hindsight) assuming that each loss function is ℓ-Lipschitz continuous w.r.t. some norm $\|\cdot\|$ and the regularizer R is σ-strongly convex w.r.t. the same norm $\|\cdot\|$.

[1] The algorithm simply explores round-robin in an initial exploration phase and commits to the best-looking arm for the remainder of time.

[2] This is known to be the best problem-independent regret rate with T that non-data (and non-problem) dependent exploration with commitment can buy.

(b) **(8 points)** Find a suitable norm $\|\cdot\|$ and values for σ, ℓ, η to get the best possible regret in terms of T and m. [Hint: Use the given facts about properties of the $\|\cdot\|_p$ norm.]

3. *Studying Stochastic Gradient Descent using Online Gradient Descent*
Suppose you want to minimize a differentiable convex function $f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ over $\mathcal{X}$. Instead of being able to observe gradients at every chosen point x, you can only receive a *stochastic*, independent unbiased gradient $G(x)$, with $\mathbb{E}\left[G(x) \;\middle|\; x\right] = \nabla f(x)$ and $\mathbb{E}\left[\|G(x)\|_2^2 \;\middle|\; x\right] \leq b^2 \; \forall x \in \mathcal{X}$. (Note: The expectation is over the independent, internal randomness of the subroutine generating the stochastic gradient.) Let $\max_{u,v\in\mathcal{X}} \|u - v\|_2 \leq D$.

Consider running the online gradient descent algorithm with the successive stochastic gradients received as follows: $x_1 = \arg\min_{x\in\mathcal{X}} \|x\|_2^2$, $\forall s \geq 1 : x_{s+1} = \Pi_{\mathcal{X}}\left(x_s - \eta G(x_s)\right)$ where $\Pi_{\mathcal{X}}$ denotes projection onto $\mathcal{X}$ w.r.t. the $\|\cdot\|_2$ norm. Suppose we run t such iterations and output the average iterate $\tilde{x}_t = \frac{1}{t}\sum_{s=1}^{t} x_s$ as a candidate minimizer for f. Let $x^* \in \mathcal{X}$ be a true minimizer of f over $\mathcal{X}$.

(a) **(2 points)** Viewing the sequence of (noisy) gradients $G_t = G(x_t)$, $t = 1, 2, \ldots$, as parameterizing a sequence of linear loss functions ($f_t(x) = G_t^T x \; \forall x \in \mathcal{X}$), write a (non-stochastic) regret bound for the algorithm's choices $x_1, x_2, \ldots$ against this loss function sequence.

(b) **(8 points)** Using the conclusion above, bound the final error $\mathbb{E}\left[f\left(\tilde{x}_t\right)\right] - f(x^*)$. [Hint: Use the convexity of f and the first and second moment information for $G(x)$.]

4. **(5 points)** *Highest Lower Confidence Bound algorithm for bandits*
Consider the following 'conservative' variant of the upper confidence bound (UCB) algorithm for stochastic multi-armed bandits with rewards in $[0, 1]$. The algorithm plays, at each time t after an initial round-robin phase, the arm with highest lower confidence bound on its mean reward:

$$I_t = \arg\max_{i\in[K]} \left(\hat{\mu}_i(t) - \sqrt{\frac{2\log t}{N_i(t)}}\right),$$

where $\hat{\mu}_i(t)$ and $N_i(t)$ denote the observed reward sample mean and number of plays from arm i upto (and not including) time t, respectively. What kind of regret[3] (in terms of the time horizon T) does this algorithm get and why? (Argue as explicitly as you can.)

[3] expected pseudo-regret as usual