E1 245 - Online Prediction and Learning, Aug-Dec 2019
Homework #4

1. *Hoeffding's inequality*
   Prove the following inequality for independent random variables $X_1, \ldots, X_n$, $n \in \mathbb{N}$, with values in $[0,1]$.
   $$\forall \varepsilon \geq 0 \quad \mathbb{P}\left[\frac{\sum_{t=1}^{T} X_t}{T} - \frac{\sum_{t=1}^{T} \mathbb{E}[X_t]}{T} \geq \varepsilon\right] \leq e^{-2T\varepsilon^2}.$$
   Hint: For any $\lambda > 0$ and a non-negative random variable Z, $\mathbb{P}[Z \geq \varepsilon] = \mathbb{P}\left[e^{\lambda Z} \geq e^{\lambda \varepsilon}\right]$; use Markov's inequality, Hoeffding's lemma and optimize over $\lambda > 0$.

2. *Bandit algorithms*
   Consider the iid[1] stochastic bandit problem with K Bernoulli-reward arms and total time T. Recall that if μ_i denotes the expected reward of the ith arm, then the regret of a bandit algorithm that plays an arm $I_t \in [N]$ at each time $1 \leq t \leq T$, and observes only the (random) reward from the chosen arm, is defined to be $R(T) := T \cdot \max_i \mu_i - \sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}]$.

   Explain briefly which of the following algorithms will/will not always achieve sublinear (pseudo-) regret with time horizon T (Recall: $R(T)$ is sublinear $\Leftrightarrow \lim_{T\to\infty} \frac{R(T)}{T} = 0$).

   (a) Play all arms exactly once. For each arm i, initialize s_i to be its observed reward and $n_i := 1$. At each time $t \leq T$, play $I_t := \arg\max_i s_i/n_i$ (break ties in any fixed manner), get (stochastic) reward R_t and update $s_{I_t} \leftarrow s_{I_t} + R_t$, $n_{I_t} \leftarrow n_{I_t} + 1$.

   (b) Play all arms exactly once. For each arm i, initialize s_i to be its observed reward and $n_i := 1$. At each time $t \leq T$, toss an independent coin with probability of heads $p := 1/\sqrt{T}$. Play $I_t := \arg\max_i s_i/n_i$ (break ties in any fixed manner) if the coin lands heads, else play a uniformly random arm, get (stochastic) reward R_t and update $s_{I_t} \leftarrow s_{I_t} + R_t$, $n_{I_t} \leftarrow n_{I_t} + 1$.

   (c) Same as the previous part but with $p := 1/T$.

   (d) Same as the previous part but with $p := 1/K$.

   (e) For each arm $i \in [N]$, initialize $u_i = 1, v_i = 1$. At each time $t \leq T$, sample independent random variables $\theta_i(t) \sim \text{Beta}(u_i, v_i)$, and play $I_t := \arg\max_i \theta_i(t)$ (break ties in any fixed manner). Get (stochastic) reward R_t and update $u_{I_t} \leftarrow u_{I_t} + R_t$, $v_{I_t} \leftarrow v_{I_t} + (1 - R_t)$.

   (f) Play all arms exactly once. For each arm i, initialize s_i to be its observed reward and $n_i := 1$. At each time $t \leq T$, let $A_t := \arg\max_i s_i/n_i$ and $B_t := \arg\max_{i \neq A_t} s_i/n_i$ denote the best and second-best arms in terms of sample mean, respectively. Play $I_t \in \{A_t, B_t\}$ chosen uniformly at random, get (stochastic) reward R_t and update $s_{I_t} \leftarrow s_{I_t} + R_t$, $n_{I_t} \leftarrow n_{I_t} + 1$.

3. *Experts game with stochastic observations*
   Consider a stochastic online learning problem with 2 actions or arms $\{1,2\}$ with Bernoulli reward distributions. Moreover, suppose you know that the arms' Bernoulli parameters (μ_1, μ_2) can be either (μ_-, μ_+) or (μ_+, μ_-), where $\mu_- := \frac{1-\varepsilon}{2}$ and $\mu_+ := \frac{1+\varepsilon}{2}$, for an unknown $\varepsilon \in (0, \frac{1}{2})$.

   At each round $1 \leq t \leq T$, a learner plays a single action $I_t \in \{1,2\}$ and gets observations as described below. Recall that the (pseudo) regret of the learner after T rounds is $\varepsilon \cdot \mathbb{E}[\text{number of times arm with mean } \mu_- \text{ is played in } T \text{ rounds}]$.

[1] independent and identically distributed

(a) Suppose that after each play, the learner only observes an independent reward sample from the action which it plays. Describe an algorithm for playing arms and a non-trivial (sub-linear in T) regret bound for it.

(b) Suppose now that after each play I_t, the learner observes independent reward samples from *both* the actions' reward distributions, i.e., it observes $X_1(t) \sim \text{Ber}(\mu_1)$ *and* $X_2(t) \sim \text{Ber}(\mu_2)$ (note that the reward earned by the learner is the same, but the other, unplayed arm's reward is also observed). Design an algorithm with as small regret in T rounds as possible. (A concrete regret bound is expected, but without needing to be precise about constants.)
(Hint: You can achieve much better regret than before, with a simpler strategy.)

4. *Exponential Weights as active Online Mirror Descent*

(a) Prove the following result. Suppose (active) OMD is run on the convex decision set $\mathcal{K}$ with a Legendre function R, where R is α-strongly convex with respect to some norm $||\cdot||$ on $\mathcal{K}$, $R(x) - R(w_1) \leq B^2 \; \forall x \in \mathcal{K}$, and the gradients of the loss functions are at most G in the dual[2] norm $||\cdot||_*$. Then, with a step size $\eta := \frac{B}{G}\sqrt{\frac{2}{T}}$, the T-round regret of OMD is at most $BG\sqrt{\frac{2T}{\alpha}}$.
Hint: In the regret bound for active OMD in class, find an upper bound for the term $D_R(w_t, w'_{t+1}) - D_R(w_{t+1}, w'_{t+1})$.

(b) Using this and the previous exercises, argue an appropriate regret bound for the Exponential weights algorithm run on the simplex Δ_d, and with linear loss functions having weights in $[0,1]$.

5. *Worst case regret for Explore-Then-Commit*
Consider the Explore-Then-Commit bandit algorithm[3], 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 $\varepsilon \in [0,1]$. Let $\Delta = \mu_1 - \mu_2 > 0$.

(a) 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.[4]

(b) 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.

[2] Recall that for a norm $||\cdot||$ in $\mathbb{R}^d$, its dual norm is defined by $||y||_* := \max_{x:||x||=1} x^T y$.

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

[4] 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.