E1 245 – Online Prediction and Learning, Fall 2022
Homework #3

1. *Approximation to the exponential*
   Prove that $\forall x \geq 0 : e^{-x} \leq 1 - x + \frac{x^2}{2}$, used in the regret bound for the EXP3 bandit algorithm.

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. *Lower Confidence Bound for stochastic 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[2] (in terms of the time horizon T) does this algorithm get and why? (Argue as explicitly as you can.)

[1] independent and identically distributed
[2] expected pseudo-regret, as usual

4. *Conjugate priors*
If the posterior distributions $\mathbb{P}\left[\theta \mid X\right]$ are in the same probability distribution family as the prior probability distribution $\mathbb{P}[\theta]$ upon observing $X \sim \mathbb{P}_\theta$ (the sample distribution), the prior is called a conjugate prior for the likelihood (sample distribution). We have seen that a Beta prior is a conjugate prior for a Bernoulli likelihood. Show explicitly the following conjugate priors for various likelihoods[3] (sample distributions):

   (a) Beta is a conjugate prior for Geometric.

   (b) Gamma is a conjugate prior for Poisson.

   (c) Normal is conjugate prior for Normal (with variance 1).

5. Three point equality for Bregman divergences
Show the following ('law of cosines') for the Bregman divergence $D_R(x,y)$ induced by a differentiable convex function $R : \mathbb{R}^d \to \mathbb{R}$:

$$\forall u, v, w \in \mathbb{R}^d : \quad D_R(u,v) + D_R(v,w) = D_R(u,w) + \langle u - v, \nabla R(w) - \nabla R(v) \rangle.$$

6. *Programming exercise*
Implement the following algorithms for a 10-armed Bernoulli bandit with the arms' means equally spaced in $(0,1)$: (a) ε-Greedy[4] with $\varepsilon = 1$ (i.e., just uniform sampling), (b) ε-Greedy, $\varepsilon = 0.1$, (c) UCB, (d) EXP3, (e) Thompson Sampling with a uniform prior.

   For each of the algorithms, plot the average cumulative regret vs. # rounds (averaged over suitably many independent trials), along with its standard deviation, for as long a time horizon T as you can. Summarize your findings.

[3] Look up the definitions of probability distributions on Wikipedia.

[4] Explores in each round independently with probability ε. If exploiting, plays the best arm w.r.t empirical mean from all past exploration rounds.