E1 245 – Online Prediction and Learning, Aug-Dec 2024
Homework #1

1. *Conjugate priors*

   If the posterior distributions $\mathbb{P}\left[\theta \,\middle|\, X\right]$ 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 (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).

   (Look up the definitions of these probability distributions on Wikipedia.)

2. *Boltzmann exploration*

   Consider the following algorithm for playing actions in a *2-armed bandit*:

Algorithm 1 Boltzmann Exploration

Require: Time horizon $n \geq 0$, Step sizes $\beta_1, \ldots, \beta_n$
  Play each arm $i \in \{1,2\}$ once and initialize its sample mean reward
  for $t = 3, 4, \ldots, n$ **do**
    Play an arm $i \in \{1,2\}$ with probability proportional to $e^{\beta_t \hat{\mu}_t(i)}$ (Note: $\hat{\mu}_t(i)$ denotes the sample mean reward from all plays of arm i up to and including time $t-1$)
  end for

   (a) Suppose all the step sizes are *constant* over time: $\beta_1 = \cdots = \beta_n = \beta > 0$. Moreover, assume that both arms yield Bernoulli-distributed rewards with parameters μ_1 and μ_2, where $0 < \mu_1 < \mu_2 < 1$. Does the algorithm obtain sub-linear[1] regret? Why? (Hint: Think about the probability of playing the worse arm.)

   (b) Now suppose the arms yield *deterministic* rewards equal to their mean values μ_1, μ_2. Suggest a suitable increasing step size schedule (i.e., how β_t should depend on t and Δ and increase with t) so that the expected number of times that the sub-optimal arm is played (i.e., $\mathbb{E}[N_n(1)]$) approximately meets the Lai-Robbins lower bound of $\frac{\log(n)}{\Delta^2}$. (Note: Assume n to be large, and feel free to ignore universal constants; the order of the answer is what is important.)

3. *Worst case (gap-independent) regret for Explore-Then-Commit*

   Consider the Explore-Then-Commit bandit algorithm[2], 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 n rounds, and an initial exploration phase of $m \leq n$ rounds. Let $\Delta = \mu_1 - \mu_2 > 0$.

[1] Sub-linear regret is when the regret in n rounds is a vanishing fraction of n.

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

(a) Write down[3] an upper bound $R(n,m,\Delta)$ for the regret of the algorithm as a function of m, n and Δ. (Hint: This has been done in class.)

(b) Suppose the exploration phase length is chosen to be *larger* than $n^{\frac{2}{3}}$: $m = n^{\left(\frac{2}{3}+\delta\right)}$ where $\delta > 0$. Find a 'bad' value for the gap Δ (depending in general on n) so that your regret bound $R(n,m,\Delta)$ becomes *at least* $n^{\frac{2}{3}}$ (order-wise).

(c) Now, on the other hand, suppose the exploration phase length is chosen to be *smaller* than $n^{\frac{2}{3}}$: $m = n^{\left(\frac{2}{3}-\delta\right)}$ where $\delta > 0$. Find a 'bad' value for the gap Δ (depending in general on n) so that your regret bound $R(n,m,\Delta)$ becomes of order $n^{\frac{2}{3}}$ (exact constants don't matter).

(d) Based on your answers above, what can you conclude about the quantity

$$\min_{1 \le m \le n} \max_{0 \le \Delta \le 1} R(n,m,\Delta),$$

as a function of n (order-wise)?

4. *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) 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] No need to derive explicitly.

[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.