Principles of Online Decision-Making (CS-303)
Problem Set 2

Problem 1

We saw in class that UCB never explicitly commits to a winning arm, but instead transitions gradually from exploring to exploiting. This might seem to suggest that a "permanently bad" sample path (like in ETC), where the optimal arm is ignored all the way to the horizon n, is not possible in UCB. This is, alas, not the case.

Think about a scenario (no math required, just reason about the operation of the UCB algorithm) in which such a sample path would arise.

Think of a fix to the preceding problem.

Problem 2

We had shown in class that the regret of ETC (with $k = 2$) satisfies

$$R_n \leq m\Delta + (n - 2m)\Delta \exp\left(-\frac{m\Delta^2}{4}\right). \tag{1}$$

We further showed that the optimal m given Δ is

$$m = \max\left\{1, \left\lceil \frac{4}{\Delta^2} \ln\left(\frac{n\Delta^2}{4}\right) \right\rceil\right\}. \tag{2}$$

(This expression accounts for the need to round to an integer, which we ignored in class.)

Show that for this optimal m, the regret is indeed bounded as $R_n \leq \Delta + C\sqrt{n}$ (with C an universal constant, i.e., that does not depend on any other parameters).

Hint: treat the cases of $\Delta \leq 1/\sqrt{n}$ and $\Delta > 1/\sqrt{n}$ separately.

Problem 3

In addition to the code of Homework 1, we provide the implementation of three new algorithms which we've seen in the last two classes: Adaptive Explore-Then-Commit (AETC), Sequential elimination (SE) and Upper Confidence Bound (UCB). All three rely on the concept of confidence intervals. We shortly remind the principle of these algorithms:

- AETC is a variation of ETC which decides to stop the exploration phase (*i.e.* commit) when the confidence interval of one of the arms is stricly superior to the confidence intervals of the others.
- SE is a smoother variation of AETC, which gradually eliminates the arms as soon as their confidence interval is stricly lower than the confidence interval of at least one other arm.
- UCB is an optimistic algorithm which always plays the arm with the highest upper confidence bound.

In addition to these three algorithms, we give you three new bandit environments: Uniform, Gaussian and Student's t distribution bandits.

(a) Complete the code of classes GaussianBandit, UniformBandit and SequentialElimination (the lines to complete are marked by `# PODMexercise`).

(b1) Compare the cumulative regrets of AETC and SE for a Gaussian bandit with three arms, having means $(1, 0.5, 0.5)$ and variance 1. Does one algorithm perform better than the other on this environment? Reason why that is the case.

(b2) Compare the cumulative regrets of AETC and SE for a Gaussian bandit with three arms, having means $(1, 1, 0.5)$ and variance 1. Does one algorithm perform better than the other on this environment? Reason why that is the case.

(c) Fix `error_prob_bound` (in lecture and in L&S, this is denoted as δ) to 0.01. Plot the time evolution of average cumulative regrets (average is taken across different samples) with time horizon 10^4. You should observe that the cumulative regret grows as almost linear to the time. Reason why this happens.

(d) Generate 10^8 samples from the standard Gaussian distribution ($\mu = 0$, $\sigma^2 = 1$). Independently, generate 10^8 samples from a Student's t-distribution with mean 0, variance 1, and degrees of freedom 3.

1. **Sub-Gaussian tail (single variable).**
   Empirically verify whether the following bound holds for both distributions:
   $$\mathbb{P}\left(|X_i| \geq \sqrt{2 \ln \frac{2}{\delta}}\right) \leq \delta$$
   for $\delta \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$.

2. **Sub-Gaussian tail (empirical mean).**
   Empirically test the tail inequality for the sample average of independent sub-Gaussian random variables:
   $$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq \sqrt{\frac{2\ln(2/\delta)}{n}}\right) \leq \delta,$$
   for $\delta \in \{10^{-3}, 10^{-2}, 10^{-1}\}$. In particular, take $n = 100$, compute 10^6 empirical means (each based on 100 samples), and compute the corresponding empirical probabilities.

(e1) Consider the gaussian bandit with three arms with means $(0.5, 0.5, 1)$ and variances $(1, 100, 100)$. Run UCB on this gaussian bandit environment several times. What do you observe? Can you explain why?

(e2) Again consider the gaussian bandit with three arms with means $(0.5, 0.5, 1)$ and variances $(1, 100, 100)$. The UCB algorithm does not properly take into account the different variances of the arms. Fill in the code of UCB_Var (in `bandit_algorithm.py`) which is a variant of UCB that handles the different variance of the arms. Run UCB_Var on this gaussian bandit. Can you explain the effect of having different variances on the choices of arms to this algorithm? (see exercise 7.2 in LS)

(f) Try Uniform, Gaussian, and Student's t distributions with the means $(1, 1, 2)$ and variance 1 for the arms (and for Student's t distribution with `df` $= 3$). What parameters should you use for UCB_Var?