# Principles of Online Decision-Making (CS-303)

## Problem Set 2 - Solutions

## 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.

---

This can happen if at some point (probably early in the execution of the algorithm), the $UCB(t-1,\delta)$ of the optimal arm drops below its mean $\mu^*$. Of course, this happens with small probability $\leq \delta$, but it is not impossible. Then it is further possible that the UCB value of another arm $i$ never drops below the "stuck" UCB value of $a^*$ (even though of course $\mu_i < \mu^*$), and therefore $a^*$ never gets played again.

---

Think of a fix to the preceding problem.

---

One principled approach would be to make $\delta$ depend on $t$, and slowly drive it to 0. This would imply that for any "stuck" arm that is not being played any longer, its UCB value would grow slowly over time. This in turn would guarantee that the arm will be played again eventually (assuming large $n$). The new samples will then eventually bring UCB back above $\mu^*$.

There are of course other ways to achieve a similar effect, e.g., by explicitly forcing random sampling of arms with some decreasing probability $\epsilon_n \to 0$.

## 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 $\Delta$ 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.

---

Case $\Delta \leq 1/\sqrt{n}$: note that the worst case where we only played the bad arm provides $R_n \leq n\Delta$, which gives $R_n \leq \sqrt{n}$, and the claim is satisfied for $C > 1$.

Case $\Delta > 1/\sqrt{n}$:

$$\begin{aligned} R_n &\leq m\Delta + (n-2m)\Delta \exp\left(-\frac{m\Delta^2}{4}\right) && (3)\\ &\leq m\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right) && (4)\end{aligned}$$

Then note that

$$\begin{aligned} m\Delta &= \Delta \cdot \max\left\{1, \left\lceil \frac{4}{\Delta^2}\ln\left(\frac{n\Delta^2}{4}\right)\right\rceil\right\} && (5)\\ &\leq \max\left\{\Delta, \Delta\left(1 + \frac{4}{\Delta^2}\ln\left(\frac{n\Delta^2}{4}\right)\right)\right\} && (6)\\ &\leq \max\left\{\Delta, \Delta + \frac{4}{\Delta}\ln\left(\frac{n\Delta^2}{4}\right)\right\} && (7)\\ &= \Delta + \frac{4}{\Delta}\max\left\{0, \ln\left(\frac{n\Delta^2}{4}\right)\right\} && (8)\end{aligned}$$

Note that the first term dominates the max when the ln is negative, which is possible. For $n$ large enough, the second term dominates the max.

To upper-bound $\exp\left(-m\Delta^2/4\right)$, we need to lower-bound $m\Delta^2$:

$$m\Delta^2 \geq \max\left\{\Delta^2, 4\ln\left(\frac{n\Delta^2}{4}\right)\right\}. \quad (9)$$

Therefore,

$$n\Delta\exp\left(-m\Delta^2/4\right) \leq n\Delta\min\left\{\exp\left(-\Delta^2/4\right), \frac{4}{n\Delta^2}\right\} \leq \frac{4}{\Delta}, \quad (10)$$

Inserting this into the bound on $R_n$ above,

$$R_n \leq \Delta + \frac{4}{\Delta}\left(1 + \max\left\{0, \ln\left(\frac{n\Delta^2}{4}\right)\right\}\right) \quad (11)$$

We still need to remove the dependence on $\Delta$ for the second term. For this, note that the last term $\frac{4}{\Delta}\max\left\{0, \ln\left(\frac{n\Delta^2}{4}\right)\right\}$ is increasing for $\Delta < \frac{2e}{\sqrt{n}}$ and decreasing for $\Delta > \frac{2e}{\sqrt{n}}$ (we find this by studying the derivative), so its maximum is reached at $\Delta = \frac{2e}{\sqrt{n}}$, for a value of $\frac{4\sqrt{n}}{e}$. So

$$R_n \leq \Delta + 4\sqrt{n} + \max_{x>0}\frac{4}{x}\max\left\{0, \ln\left(\frac{nx^2}{4}\right)\right\} = \Delta + 4\sqrt{n} + \frac{4\sqrt{n}}{e} \quad (12)$$

This gives $R_n \leq \Delta + (4 + 4e^{-1})\sqrt{n}$.

We have shown the bound, and the universal constant is $C > 4 + 4e^{-1}$.

## 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.

---

AETC and SE perform almost the same because SE cannot eliminate either of the suboptimal arms until nearly the end of the exploration phase, as both have the same mean and thus are equally suboptimal. As a result, both algorithms spend almost equivalent amount of time exploring the two suboptimal arms before committing to the optimal one, leading to similar cumulative regrets. However, SE can eliminate one of the two suboptimal arms near the end of the exploration period, which can give SE just a slight advantage over AETC in terms of cumulative regret.

---

(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.

---

SE performs better than AETC in this setting. Since both arms 0 and 1 have the highest mean (1), SE can quickly eliminate arm 2 (mean 0.5) as soon as its confidence interval is strictly lower than those of arms 0 and 1. After eliminating arm 2, SE focuses its exploration and exploitation on arms 0 and 1, both of which are optimal. In contrast, AETC continues to explore all arms until it is confident that one arm is better than the others, which takes significantly longer if not never because arms 0 and 1 are indistinguishable in terms of mean. As a result, SE incurs less cumulative regret than AETC, since it spends significantly less time on the suboptimal arm (arm 2).

---

(c) Fix `error_prob_bound` (in lecture and in L&S, this is denoted as $\delta$) 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.

---

The cumulative regret grows almost linearly with time because, with a fixed `error_prob_bound`. This means that the probability of incorrectly identifying the optimal arm remains roughly constant (roughly proportinal to `error_prob_bound`) per exploration phase, and the algorithm may continue to exploit a suboptimal arm for a significant portion of the time horizon. (Specifically, in the worst case scenario, the top of confidence interval of the optimal arm can be lower than the true mean of

the suboptimal arm. When this happens, as optimal arm is never triggered thereafter, the top of confidence interval of optimal arm stays the same. This tail event happens roughly proportinal to `error_prob_bound`.) As a result, the expected cumulative regret is approximately proportional to both the time horizon and the error probability bound.

---

(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.

---

The reward estimate by ETC for the gaussian bandit converges to the true mean (1) much faster than that for the Student's t distribution bandit. This is because the gaussian distribution has lighter tails compared to the Student's t distribution with low degrees of freedom (df=1.4). As a result, samples from the gaussian distribution are less variable and more concentrated around the mean, leading to quicker convergence of the reward estimate. In contrast, the Student's t distribution with df=1.4 has heavier tails, which means that samples can be more extreme and variable, causing slower convergence of the reward estimate to the true mean. This is illustrated by several spikes in the reward estimate for the Student's t distribution bandit, which are less frequent in the gaussian bandit.

---

(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?

---

We observe that UCB frequently fails to identify the optimal arm (arm 3 with mean 1) and instead frequently selects the suboptimal arms (arms 0 and 1 with mean 0.5). This happens because UCB assumes that all arms are subgaussian with the same variance proxy 1, the confidence intervals it constructs for each arm are based on this assumption.

As the variance for arm 2 is quite high, the reward sampled from this arm can be much lower than 0.5 (the means of arms 0 and 1), so can be estimated reward of arm 2. Since UCB computes the confidence interval based on the assumption of variance proxy being 1, the confidence interval is far underestimated. This leads to a situation where the top of confidence bounds of arm 0 can be frequently higher than that of arm 2, causing UCB to select arm 0 over arm 2.

---

(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)

Because of the high variances, arms 1 and 2 need many samples to determine which arm is better, while arm 0 is significanly less explored.

---

(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?

---

What we need to do is the find the right variance proxy for each bandit arm.

For Uniform distribution, first we need to compute the proper interval $[a, b]$ that leads mean 1 and variance 1. Notice that pdf is $p(x) = 1/x$ for $x \in [a, b]$, and the conditions can be rewritten as

- $\int_a^b xp(x)dx = 1$
- $\int_a^b x^2p(x)dx = 1$.

By solving these equations, we get $[0, 2]$ as the support of the uniform distribution. By Hoeffding's lemma, this is subgaussian with variance proxy 1. (variance $\leq$ varaiance proxy holds but not necessarily variance = variance proxy. )

The subgaussian's variance proxy for gaussian distribution $N(1, 1)$ is 1.

Since the t-distribution is not subgaussian, we cannot find a proper parameter for UCB_Var.