
Problem Set 3 Bandits

For the Exercise Sessions on Sep 29 and Oct 6 — **Due: Tue, October 13, 10am, on Moodle**

1 The Homework

The following three problems are the core of this homework. You work out solutions and turn them in. Problem 1 is suitable for Sept 29. Problems 2 and 3 are suitable for Oct 6.

In class, we studied bandit algorithms with certain requirements on what the learner knows and how it makes decisions. The graded problems lift one of these requirements at a time:

- *Problem 1: Choose an exploration length without knowing the gap between the arms, then study a strategy that does not need to know the time horizon in advance.*
- *Problem 2: Adapt UCB to Gaussian rewards without assuming that their variances are known.*
- *Problem 3: Modify UCB so that it does not need to recompute its choice of arm at every round.*

Problem 1: Epsilon-Greedy Algorithm

Recall our original *explore-then-exploit* strategy. We had a fixed time horizon n . For some m , a function of n and the gaps $\{\Delta_k\}$, we explore each of the K arms m times initially. Then we pick the best arm according to their empirical gains and play this arm until we reach round n . We have seen that this strategy achieves an asymptotic regret of order $\ln(n)$ if the environment is fixed and we think of n tending to infinity but a worst-case regret of order $\sqrt{n}$. To get these results, we used the gaps Δ_k in order to select m . What if we cannot use the gaps?

In the sequel, to keep things simple, let us restrict attention to the case of two arms. (But you are invited to extend the analysis to the general case of K arms.) Moreover, let us restrict to the case where the rewards are in $[0, 1]$. Note that this implies that $0 \leq \Delta \leq 1$.

- (a) Recall the regret upper bound from lecture. For the two-arms case: $R_n \leq m\Delta + n\Delta \exp(-m\Delta^2/4)$. Let us select the exploration phase to be of duration $m = n^{\frac{2}{3}}$. Note that this choice of m does not require us to know that gap Δ . Show that the worst-case regret (worst case over Δ) is of order no larger than $n^{\frac{2}{3}}$.

Here is a slightly different algorithm. Let $\epsilon_t = t^{-\frac{1}{3}}$. For each round $t = 1, \dots$, toss a coin with success probability ϵ_t . If success, then explore arms uniformly at random. If not success, then pick in this round the arm that currently has the highest empirical average.

- (b) Let $T_k(i-1)$ denote the number of times we chose arm k up to and including time $i-1$. Show that with this coin tossing approach, at any given time i , we have

$$\mathbb{P}\left[T_k(i-1) < (i-1)^{2/3}/8\right] \leq \frac{16}{(i-1)^{2/3}},$$

for $k = 1, 2$, that is, for both arms. *Hint:* Define Bernoulli random variables $B_k(t)$ such that $B_k(t) = 1$ represents the event that at time t , the coin toss result is 1 and arm k is picked. Show that $\text{Var}\left(\sum_{t=1}^{i-1} B_k(t)\right) \leq \frac{3}{2}(i-1)^{2/3}$, for both arms, $k = 1, 2$, and use the Chebyshev inequality.

- (c) Finally, we want to analyze the regret of the full algorithm. Namely, show that for this algorithm the expected regret at *any* time t is upper bounded by $t^{\frac{2}{3}}$ (orderwise). This is similar to the worst-case of the explore-then-exploit strategy but here we do not need to know the horizon a priori.

Hint: First, condition on the event where $T_1(t) \geq t^{2/3}/2$ and $T_2(t) \geq t^{2/3}/2$. Bound the regret in this case.

Solution 1. We take up the parts.

- (a) For starters, note that $xe^{-\gamma x^2} \leq \frac{1}{\sqrt{2\gamma}}$. Hence, we have

$$R_n \leq m\Delta + n\Delta \exp(-m\Delta^2/4) \quad (1)$$

$$\leq m\Delta + n \frac{1}{\sqrt{m/4}} \quad (2)$$

Plugging in $m = n^{\frac{2}{3}}$, and observing that $\Delta \leq 1$,

$$R_n \leq n^{\frac{2}{3}} + 2n^{\frac{2}{3}}. \quad (3)$$

- (b) Let us start by considering the random variables $B_k(t)$. First, note that $\mathbb{P}(B_k(t) = 1) = \epsilon_t/K$ (and otherwise, $B_k(t) = 0$). Therefore, we can easily find the expected value:

$$\nu := \mathbb{E}\left[\sum_{t=1}^{i-1} B_k(t)\right] = \sum_{t=1}^{i-1} \mathbb{E}[B_k(t)] = \frac{1}{K} \sum_{t=1}^{i-1} \epsilon_t = \frac{1}{K} \sum_{t=1}^{i-1} t^{-1/3}. \quad (4)$$

We need to upper and lower bound this. Namely,

$$\begin{aligned} \sum_{t=1}^{i-1} t^{-\frac{1}{3}} &\leq 1 + \int_1^{i-1} x^{-\frac{1}{3}} dx \leq \frac{3}{2}(i-1)^{\frac{2}{3}}, \\ \sum_{t=1}^{i-1} t^{-\frac{1}{3}} &\geq \int_1^i x^{-\frac{1}{3}} dx \geq \frac{3}{2}(i^{\frac{2}{3}} - 1), \end{aligned}$$

and thus, we have the following upper and lower bounds on the mean ν :

$$\frac{3}{2K}(i^{\frac{2}{3}} - 1) \leq \nu \leq \frac{3}{2K}(i-1)^{\frac{2}{3}}. \quad (5)$$

Moreover, since the $B_k(t)$ are independent random variables, we have (see Homework 1, Problem 4)

$$\text{Var}\left(\sum_{t=1}^{i-1} B_k(t)\right) = \sum_{t=1}^{i-1} \text{Var}(B_k(t)). \quad (6)$$

Of course, for a single Bernoulli random variable, we have

$$\text{Var}(B_k(t)) = \frac{\epsilon_t}{K}(1 - \frac{\epsilon_t}{K}) \leq \frac{\epsilon_t}{K}. \quad (7)$$

Hence,

$$\text{Var} \left(\sum_{t=1}^{i-1} B_k(t) \right) \leq \frac{1}{K} \sum_{t=1}^{i-1} \epsilon_t = \frac{1}{K} \sum_{t=1}^{i-1} t^{-1/3} = \nu. \quad (8)$$

Now, how to use this? First of all, observe that $T_k(i) \geq \sum_{t=1}^{i-1} B_k(t)$. Therefore,

$$\mathbb{P} \left[T_k(i-1) < (i-1)^{2/3}/(4K) \right] \leq \mathbb{P} \left[\sum_{t=1}^{i-1} B_k(t) < (i-1)^{2/3}/(4K) \right].$$

Next, using ν as above, we can write

$$\mathbb{P} \left[\sum_{t=1}^{i-1} B_k(t) < (i-1)^{2/3}/(4K) \right] = \mathbb{P} \left[\sum_{t=1}^{i-1} B_k(t) - \nu < (i-1)^{2/3}/(4K) - \nu \right].$$

Now, it seems that we cannot use Chebyshev... well, here is a fun trick:

$$\mathbb{P} \left[\sum_{t=1}^{i-1} B_k(t) - \nu < (i-1)^{2/3}/(4K) - \nu \right] = \mathbb{P} \left[\nu - \sum_{t=1}^{i-1} B_k(t) > \nu - (i-1)^{2/3}/(4K) \right].$$

Moreover, clearly, taking absolute values inside the expression,

$$\mathbb{P} \left[\nu - \sum_{t=1}^{i-1} B_k(t) > \nu - (i-1)^{2/3}/(4K) \right] \leq \mathbb{P} \left[\left| \nu - \sum_{t=1}^{i-1} B_k(t) \right| > \nu - (i-1)^{2/3}/(4K) \right].$$

From our lower bound on ν , we know that for $i \geq 2$, we have $\nu - (i-1)^{2/3}/(4K) > 0$. Thus, we can now use Chebyshev to conclude:

$$\begin{aligned} \mathbb{P} \left[\left| \nu - \sum_{t=1}^{i-1} B_k(t) \right| > \nu - (i-1)^{2/3}/(4K) \right] &\leq \frac{\mathbb{E}[(\nu - \sum_{t=1}^{i-1} B_k(t))^2]}{(\nu - (i-1)^{2/3}/(4K))^2} \\ &= \frac{\text{Var}(\sum_{t=1}^{i-1} B_k(t))}{(\nu - (i-1)^{2/3}/(4K))^2}. \end{aligned}$$

But as we have shown, $\text{Var}(\sum_{t=1}^{i-1} B_k(t)) \leq \nu$, hence, combining, we can conclude

$$\mathbb{P} \left[T_k(i-1) < (i-1)^{2/3}/(4K) \right] \leq \frac{\nu}{(\nu - (i-1)^{2/3}/(4K))^2} \leq \frac{8K}{(i-1)^{2/3}},$$

where for the last step, we have plugged in our lower bound on ν (observing that the middle expression is a decreasing function of ν).

The expected regret has two components. The first component is due to the fact if the coin toss results in success then explore. In this case we can get a regret of at most 1. Therefore, this contribution can be upper bounded by

$$\sum_{i=1}^t i^{-\frac{1}{3}} \leq 1 + \int_1^t x^{-\frac{1}{3}} dx \leq \frac{3}{2} t^{\frac{2}{3}}.$$

The second contribution comes from the exploitation phase.

Let $B_t \in \{0, 1\}$ denote the result of coin-toss at round t with $\mathbb{P}\{B_t = 1\} = \epsilon_t$ (success) and $\mathbb{P}\{B_t = 0\} = 1 - \epsilon_t$ (fail). Then the average regret at time t with X_t as the reward for the round t is

$$\begin{aligned}
R_t &= t\mu^* - \mathbb{E}\left[\sum_{i=1}^t X_i\right] \\
&= t\mu^* - \sum_{i=1}^t \mathbb{E}[\mathbb{E}[X_i|B_i]] \\
&= t\mu^* - \sum_{i=1}^t (\mathbb{E}[X_i|B_i = 1]\mathbb{P}\{B_i = 1\} + \mathbb{E}[X_i|B_i = 0]\mathbb{P}\{B_i = 0\}) \\
&= t\mu^* - \sum_{i=1}^t \epsilon_i \mathbb{E}[X_i|B_i = 1] - \sum_{i=1}^t (1 - \epsilon_i) \mathbb{E}[X_i|B_i = 0]
\end{aligned}$$

Note that given $B_i = 1$, X_i is the reward that we uniformly pick an arm. Hence

$$E[X_i|B_i = 1] = \frac{1}{K} \sum_{k=1}^K \mu_k$$

Note that given $B_i = 0$, X_i is the reward that we pick the arm that has highest empirical average. Hence

$$E[X_i|B_i = 0] = \sum_{k=1}^K \mu_k \mathbb{P}\{k = \arg \max_j \hat{\mu}_j(i-1)\}$$

where $\hat{\mu}_j(i-1)$ is the empirical estimator of μ_j arm j until round $i-1$. We assume that arm 1 has the largest expected reward, $\mu^* = \mu_1$. Since the probability that the empirical mean converges to the real mean grows with the number of samples, and consequently so does the probability that we choose the arm with the largest mean. Then:

$$\begin{aligned}
&\mathbb{P}\{k = \arg \max_j \hat{\mu}_j(i-1)\} \\
&\leq \mathbb{P}\{\hat{\mu}_1(i-1) - \hat{\mu}_k(i-1) \leq 0\} \\
&= \mathbb{P}\left\{\frac{1}{T_1(i-1)} \sum_{j=1}^{T_1(i-1)} X_j^{(1)} - \frac{1}{T_k(i-1)} \sum_{j=1}^{T_k(i-1)} X_j^{(k)} \leq 0\right\} \\
&= \mathbb{E}\left[1 \left\{\frac{1}{T_1(i-1)} \sum_{j=1}^{T_1(i-1)} X_j^{(1)} - \frac{1}{T_k(i-1)} \sum_{j=1}^{T_k(i-1)} X_j^{(k)} \leq 0\right\}\right] \\
&\stackrel{(i)}{=} \mathbb{E}\left[\mathbb{E}\left[1 \left(\frac{1}{T_1(i-1)} \sum_{j=1}^{T_1(i-1)} (X_j^{(1)} - \mu_1) - \frac{1}{T_k(i-1)} \sum_{j=1}^{T_k(i-1)} (X_j^{(k)} - \mu_k) \leq \mu_1 - \mu_k\right) \middle| T_1(i-1), T_k(i-1)\right]\right] \\
&\stackrel{(ii)}{\leq} \mathbb{E}\left[e^{-\frac{1}{2} \frac{\Delta_k^2}{4T_1(i-1) + 4T_k(i-1)}}\right] \\
&\leq \mathbb{P}\left[T_1(i-1) < (i-1)^{2/3}/K \vee T_k(i-1) < (i-1)^{2/3}/K\right] \times 1 \\
&\quad + \mathbb{P}\left[T_1(i-1) \geq (i-1)^{2/3}/K \wedge T_k(i-1) \geq (i-1)^{2/3}/K\right] e^{-\frac{1}{2} \frac{\Delta_k^2}{\frac{K}{4(i-1)^{2/3}} + \frac{K}{4(i-1)^{2/3}}}} \\
&\leq \mathbb{P}\left[T_1(i-1) < (i-1)^{2/3}/K \vee T_k(i-1) < (i-1)^{2/3}/K\right] + e^{-\frac{(i-1)^{2/3} \Delta_k^2}{K}}
\end{aligned}$$

Where (i) comes from the law of total Expectation and (ii) from the σ -sub Gaussianity assumption.

Now, from Part (b) of the problem, we know that the probability that $T_k(i-1) < c(i-1)^{2/3}$ is low for some constant c .

Unfortunately, the original solution to the problem, which we reproduce now, has a slight issue in this very calculation:

At every step t there is a probability $t^{-1/3}/K$ that we increase it. Let $B_k(t)$ be a Bernoulli random variable that represent the event that at step t we choose arm k with probability $t^{-1/3}/K$. We have that $T_k(i-1) \geq \sum_{t=1}^{i-1} B_k(t)$. The variance of $B_k(t)$ is $t^{-1/3}/K(1 - t^{-1/3}/K) \leq t^{-1/3}/K$ and so the variance of $\sum_{t=1}^{i-1} B_k(t)$ is upperbounded by $\frac{1}{K} \sum_{t=1}^{i-1} t^{-1/3} \leq \frac{3(i-1)^{2/3}}{K}$.

$$\begin{aligned}
\mathbb{P}\left[T_k(i-1) < (i-1)^{2/3}/K\right] &\leq \mathbb{P}\left[\sum_{t=1}^{i-1} B_k(t) - \nu < (i-1)^{2/3}/K - \nu\right] \\
&\leq \mathbb{P}\left[\sum_{t=1}^{i-1} B_k(t) - \nu < -\frac{(i-1)^{2/3}}{2K} + 1\right] \\
&\stackrel{(\times)}{\leq} \mathbb{P}\left[\sum_{t=1}^{i-1} B_k(t) - \nu < -\frac{(i-1)^{2/3}}{3K}\right] \\
&\stackrel{(\star)}{\leq} \frac{27K}{4(i-1)^{2/3}}
\end{aligned}$$

where $(\times)$ is not exactly correct (this is precisely the issue mentioned above), but it does hold for $i > 1 + (6K)^{3/2}$ and where $(\star)$ follows by Chebychev inequality with the function $f(x) = x^2$, i.e. $\mathbb{P}[X - E[X] < -a] \leq \text{var}(X^2)/a^2$ and ν denotes the mean of $\sum_t B_k(t)$. Thus:

$$\begin{aligned}
\mathbb{P}\{k = \arg \max_j \hat{\mu}_j(i-1)\} &\leq \mathbb{P}\left[T_1(i-1) < (i-1)^{2/3}/K \vee T_k(i-1) < (i-1)^{2/3}/K\right] + e^{-\frac{(i-1)^{2/3}\Delta_k^2}{K}} \\
&\leq \mathbb{P}\left[T_1(i-1) < (i-1)^{2/3}/K\right] + \mathbb{P}\left[T_k(i-1) < (i-1)^{2/3}/K\right] + e^{-\frac{(i-1)^{2/3}\Delta_k^2}{K}} \\
&\leq \frac{3K}{2(i-1)^{2/3}} + e^{-\frac{(i-1)^{2/3}\Delta_k^2}{K}}
\end{aligned}$$

Hence, the average regret at round t is given by

$$\begin{aligned}
R_t &= t\mu^* - \sum_{i=1}^t \epsilon_i \frac{1}{K} \sum_{k=1}^K \mu_k - \sum_{i=1}^t (1 - \epsilon_i) \sum_{k=1}^K \mu_k \mathbb{P}\{k = \arg \max_j \hat{\mu}_j(i-1)\} \\
&= \sum_{i=1}^t \epsilon_i \left(\mu^* - \frac{1}{K} \sum_{k=1}^K \mu_k\right) + \sum_{i=1}^t (1 - \epsilon_i) \left(\mu^* - \sum_{k=1}^K \mu_k \mathbb{P}\{k = \arg \max_j \hat{\mu}_j(i-1)\}\right) \\
&= \mathbb{E}[\Delta_k] \sum_{i=1}^t \epsilon_i + \sum_{i=2}^t (1 - \epsilon_i) \sum_{k=1}^K \Delta_k \mathbb{P}\{k = \arg \max_j \hat{\mu}_j(i-1)\} \\
&\leq \frac{3}{2} t^{\frac{2}{3}} + \sum_{i=2}^t (1 - \epsilon_i) K \left(\frac{3K}{2(i-1)^{2/3}} + e^{-\frac{(i-1)^{2/3}\Delta_k^2}{K}} \right) \\
&< \frac{3}{2} t^{\frac{2}{3}} + \frac{9K^2(t-1)^{1/3}}{2} + (t - t^{2/3}) K e^{-\frac{(t-1)^{2/3}\Delta_k^2}{K}}
\end{aligned}$$

where $\Delta_* = \min_{j \in \{2, \dots, K\}} \Delta_j$. Note that $\mathbb{E}[\Delta_k] \leq 1$ due to $0 \leq \Delta_k \leq 1$, and $\sum_{i=1}^t \epsilon_i = \sum_{i=1}^t i^{-1/3} \leq \int_1^t x^{-1/3} dx \leq \frac{3}{2} t^{2/3}$ and $\sum_{i=2}^t (i-1)^{-2/3} \leq \int_1^{t-1} x^{-2/3} dx \leq 3(t-1)^{1/3}$.

Problem 2: Gaussian bandits with unknown mean and variance

Consider the standard setup of the bandit problem we discussed in the course. We assumed that the K arms have unknown means and are all 1-subgaussian. For the *upper confidence bound* (UCB) algorithm, after sampling each arm once in the beginning, in each subsequent round we decided on the arm according to

$$\operatorname{argmax}_k \underbrace{\hat{\mu}_{k,t-1} + \sqrt{\frac{2}{T_k(t-1)} \ln \frac{1}{\delta_t}}}_{\text{decision metric}}. \quad (9)$$

Here, $T_k(t-1)$ denotes the number of times we chose arm k up to and including time $t-1$, $\hat{\mu}_{k,t-1}$ is the empirical mean of arm k at time $t-1$, using the relevant $T_k(t-1)$ samples, and δ_t denotes the confidence we want to have at time t .

a) Write down the decision metric if arm k , $1 \leq k \leq K$, is σ_k^2 -subgaussian and the parameters $\{\sigma_k^2\}$ are known?

In general the parameters $\{\sigma_k^2\}_{k=1}^K$ are unknown. However, we can estimate their values from the samples. More precisely, if we are given iid samples $W_1, \dots, W_n$ let $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n W_i$ and let $\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (W_i - \hat{\mu}_n)^2$ be the sample mean and the sample variance, respectively. It is then tempting to conjecture that we can construct an UCB algorithm by replacing the variance terms σ_k^2 in point (a) above with their empirical values $\hat{\sigma}_{k,t-1}^2$ and possibly change the involved constants.

We will now confirm this conjecture for the concrete case where all arms have a Gaussian distribution. More precisely, we assume that arm k , $1 \leq k \leq K$, is distributed according to $\mathcal{N}(\mu_k, \sigma_k^2)$. The parameters $\{(\mu_k, \sigma_k^2)\}_{k=1}^K$ are unknown.

b) Write down the decision metric for this case.

Hint: If we are given i.i.d. samples $W_1, \dots, W_n$ from a Gaussian distribution with parameters (μ, σ^2) , and $(\hat{\mu}_n, \hat{\sigma}_n^2)$ are the respective empirical quantities, then

$$S_n = \frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n / \sqrt{n}}$$

is distributed according to the so-called student t -distribution with $n-1$ degrees of freedom. Note that the distribution of S_n does not depend on (μ, σ^2) and is symmetric around 0. Further, the confidence values for this distribution can be looked up in tables or can be computed. I.e., you can assume that for any $\delta \in [0, 1]$ the real numbers $\alpha_{\delta, n}$, so that

$$\mathbb{P}\{S_n > \alpha_{\delta, n}\} = \delta$$

are known.

c) Show that the distribution of S_n indeed does not depend on (μ, σ^2) .

Hint: The distribution of $(\hat{\mu}_n - \mu)$ is Gaussian with mean zero and variance σ^2/n . Further, $(n-1)\hat{\sigma}_n^2/\sigma^2$ follows the so-called χ^2 distribution with $n-1$ degrees of freedom. Note that the χ^2 distribution does not depend on (μ, σ^2) . Further, $(\hat{\mu}_n - \mu)$ and $(n-1)\hat{\sigma}_n^2/\sigma^2$ are independent.

Solution 2.

a) In this case we would decide on the next arm according to

$$\operatorname{argmax}_k \underbrace{\hat{\mu}_{k,t-1} + \sigma_k \sqrt{\frac{2}{T_k(t-1)} \ln \frac{1}{\delta_t}}}_{\text{decision metric}}.$$

b) For a given value $\delta > 0$ (the confidence value) and $n \in \mathbb{N}$ (the degrees of freedom), let $\alpha_{\delta,n}$ be the value so that $P\{S_n > \alpha_{\delta,n}\} = \delta$ according to the student t -distribution with n degrees of freedom. This value can be looked up in a table or computed numerically.

According to the hint, $\frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n/\sqrt{n}}$ has a student t -distribution with $n - 1$ degrees of freedom. Hence

$$P\left\{\frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n/\sqrt{n}} > \alpha_{\delta,n-1}\right\} = \delta.$$

By the symmetry of this distribution we have

$$P\left\{\frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n/\sqrt{n}} \geq -\alpha_{\delta,n-1}\right\} = 1 - \delta.$$

This can be rewritten as

$$P\left\{\mu \leq \hat{\mu}_n + \frac{\alpha_{\delta,n-1}\hat{\sigma}_n}{\sqrt{n}}\right\} = 1 - \delta.$$

In turn this is equivalent to

$$P\left\{\mu > \hat{\mu}_n + \frac{\alpha_{\delta,n-1}\hat{\sigma}_n}{\sqrt{n}}\right\} = \delta.$$

Hence we see that the decision should be taken according to

$$\operatorname{argmax}_k \hat{\mu}_{k,t-1} + \frac{\alpha_{\delta,T_k(t-1)-1}\hat{\sigma}_{k,t-1}}{\sqrt{T_k(t-1)}}.$$

This is of the same form as our original decision metric, just with a different constant and we swapped out the true standard deviation for the empirical standard deviation.

c) We have

$$S_n = \frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n/\sqrt{n}} = \frac{(\hat{\mu}_n - \mu)/\sigma}{\hat{\sigma}_n/(\sigma\sqrt{n})}.$$

Now note that by the hint, the numerator and the denominator are independent random variables and that the distributions of those random variables do not depend on (μ, σ^2) . Hence, S_n has a distribution that does not depend on (μ, σ^2) .

Problem 3: UCB With Geometric Intervals

Consider the following slight variant of the UCB algorithm. We have K arms. As in the lecture notes, assume that each of these K arms corresponds to a random variable which is 1-subgaussian. For the first K steps we sample each of these arms once. After these K first steps we have an interval of length 1, then an interval of length 2, then one of length 4, and so on. At the beginning of each such interval we choose the arm in the same manner as the UCB algorithm. More precisely, if t marks the beginning of a new interval then

$$A_t = \operatorname{argmax}_k \hat{\mu}_k(t-1) + \sqrt{\frac{2 \ln f(t)}{T_k(t-1)}},$$

where $f(g) = 1 + t \ln^2(t)$ as for the case we discussed in the course and where $T_k(t-1)$ denotes the number of times we have chosen arm k in the last $t-1$ steps. But unlike the standard UCB algorithm, for all other steps in this interval we keep the same arm. Why might we be interested in such an algorithm? One motivation is complexity. Computing which arm is best takes some effort. In this way we only have to compute the best arm a logarithmic (in the time horizon) number of times.

Recall that in the analysis of the original algorithm the key to the analysis was to find a good upper bound on $T_k(n)$ for $k > 1$, assuming that arm 1 is the optimum arm. In turn, we upper bounded the probability that we choose arm k at a particular point in time t by the probability that arm 1 had an empirical mean at least an ϵ below its true mean μ_1 and that the empirical mean of arm k was above $\mu_1 - \epsilon$. In formulae we had

$$T_k(n) = \sum_{t=1}^n 1_{\{A_t=k\}} \leq \sum_{t=1}^n 1_{\{\hat{\mu}_1(t-1) + \sqrt{\frac{2 \ln f(t)}{T_1(t-1)}} \leq \mu_1 - \epsilon\}} + \sum_{t=1}^n 1_{\{\hat{\mu}_k(t-1) + \sqrt{\frac{2 \ln f(t)}{T_k(t-1)}} \geq \mu_1 - \epsilon \wedge A_t=k\}} \quad (10)$$

Let us proceed in the same fashion. Let $n = K + 2^L - 1$. In words, we are at the end of the L -th interval, where $L \in \mathbb{N}$.

(a) What is the expression equivalent to (10) for our case?

Solution: Let $t(l) = K + 2^{l-1}$. This is the time at the beginning of the l -th interval. In the sequel, to lighten the notation, we will sometimes just write t . Then we have

$$T_k(n) = \sum_{l=1}^L 1_{\{A_{t(l)}=k\}} 2^{l-1} \leq \sum_{l=1}^L 2^{l-1} 1_{\{\hat{\mu}_1(t(l)-1) + \sqrt{\frac{2 \ln f(t(l))}{T_1(t(l)-1)}} \leq \mu_1 - \epsilon\}} + \sum_{l=1}^L 2^{l-1} 1_{\{\hat{\mu}_k(t(l)-1) + \sqrt{\frac{2 \ln f(t(l))}{T_k(t(l)-1)}} \geq \mu_1 - \epsilon \wedge A_{t(l)}=k\}} \quad (11)$$

(b) Look at the first of the two terms on the right of (10) in your equivalent expression. Derive a suitable upper bound for this first term. If you do not have time for the whole derivation just write down the first few steps. These are the most crucial ones.

Solution:

For the first term the derivation is almost the same as what we saw in the course. to ease the

notation burden, we will write $t(l)$ as t in the following (in)equalities. We have

$$\begin{aligned}
\mathbb{E}\left[\sum_{l=1}^L 2^{l-1} 1_{\{\hat{\mu}_1(t-1) + \sqrt{\frac{2 \ln f(t)}{T_1(t-1)}} \leq \mu_1 - \epsilon\}}\right] &= \sum_{l=1}^L 2^{l-1} \mathbb{P}\left(\hat{\mu}_1(t-1) + \sqrt{\frac{2 \ln f(t)}{T_1(t-1)}} \leq \mu_1 - \epsilon\right) \\
&\leq \sum_{l=1}^L 2^{l-1} \sum_{s=1}^t \mathbb{P}\left(\hat{\mu}_{1,s} + \sqrt{\frac{2 \ln f(t)}{s}} \leq \mu_1 - \epsilon\right) \\
&\quad \text{(condition over } T_1 \text{ and use average is less than the sum)} \\
&\leq \sum_{l=1}^L 2^{l-1} \sum_{s=1}^t e^{-\frac{s}{2}(\sqrt{\frac{2 \ln f(t)}{s}} + \epsilon)^2} \\
&= \sum_{l=1}^L 2^{l-1} \sum_{s=1}^t e^{-\ln(f(t)) - \sqrt{2s \ln f(t)} - \frac{s}{2}\epsilon^2} \\
&\leq \sum_{l=1}^L 2^{l-1} \frac{1}{f(t)} \sum_{s=1}^t e^{-\frac{s}{2}\epsilon^2} \\
&= \sum_{l=1}^L 2^{l-1} \frac{1}{f(t)} \frac{e^{-\frac{\epsilon^2}{2}}}{1 - e^{-\frac{\epsilon^2}{2}}} \\
&= \sum_{l=1}^L 2^{l-1} \frac{1}{f(t)} \underbrace{\frac{1}{e^{\frac{\epsilon^2}{2}} - 1}}_{\substack{\text{take Taylor series} \\ \text{all terms are positive; keep only first two}}} \\
&\leq \sum_{l=1}^L \frac{1}{f(t)} 2^{l-1} \frac{2}{\epsilon^2} \\
&\leq \frac{2}{\epsilon^2} \sum_{l=1}^L 2^{l-1} \frac{1}{1 + t \ln(t)^2} \\
&\leq \frac{2}{\epsilon^2} \left(1 + \sum_{l=2}^L \frac{1}{(l-1)^2}\right) \quad (t \ln(t)^2 \geq 2^{l-1}(l-1)^2) \\
&\leq \frac{2}{\epsilon^2} \left(2 + \int_1^{L-1} \frac{1}{x^2} dx\right) \quad \text{(upper bound discrete sum with integral)} \\
&\leq \frac{6}{\epsilon^2}
\end{aligned}$$

2 Additional Problems

For the next six problems, you do not need to turn in solutions. For the exam, we do expect you know how to solve these problems. So, you may keep these problems around and tackle them in preparation for the Midterm and/or Final exam. Or you may solve them now.

Problem 4: Finding the Fair Coin

Your colleague challenges you to a game. It goes as follows: You are given three coins, one being fair, two being weighted. Coin one has a probability of flipping heads $1/2$, coin two has probability of flipping heads $1/2 - p$, and coin three has probability of flipping heads $1/2 + p$. You can assume that $p \in (0, 1/3]$. You are allowed to flip every coin m times. If thereafter, you manage to correctly identify the fair coin, you win. Otherwise you lose.

- a) Describe a simple strategy that is winning for $m \rightarrow \infty$.
- b) Write down the events that make this strategy fail (assume finite m).
- c) Give a simple upper bound on the total failure probability in terms of m and p .
- d) Your colleague makes the following proposal: you can flip each coin $m = 20$ times, and he ensures you that $p = 1/3$ (he is a good friend, so you trust him). If you can identify the fair coin, you win 3 CHF, otherwise you lose 2 CHF. Do you accept?

In general, for $p \in (0, 1/3]$ fixed, for which values of m , α would you agree to play the game, where α is defined as the ratio between the potential gain and potential loss in CHF? You can assume that you know p . It is sufficient to state bounds that are tight up to multiplicative constants.

Solution 3. a) Associate heads with 0s and tails with 1s. Compute the empirical means $\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3$ of each of the three coins. Select the coin corresponding to the second largest empirical mean. This strategy is winning for $m \rightarrow \infty$ by the law of large numbers.

There is a small subtlety here: What if two (or more) of the empirical means are exactly equal? Let us agree that in this case, our algorithm gives up and declares “failure.”

- b) $E_1 = \{\hat{\mu}_2 \leq \hat{\mu}_3 \leq \hat{\mu}_1\}$, $E_2 = \{\hat{\mu}_1 \leq \hat{\mu}_2 \leq \hat{\mu}_3\}$, $E_3 = \{\hat{\mu}_3 \leq \hat{\mu}_2 \leq \hat{\mu}_1\}$, $E_4 = \{\hat{\mu}_1 \leq \hat{\mu}_3 \leq \hat{\mu}_2\}$
- c) Denote by $X_{j,k}$ the random variable associated with flipping the k -th coin for the j -th time. Notice that

$$\{\text{failure}\} = \bigcup_i E_i \subseteq \{\hat{\mu}_1 \leq \hat{\mu}_2\} \cup \{\hat{\mu}_3 \leq \hat{\mu}_1\},$$

so that

$$\mathbb{P}(\text{failure}) \leq \mathbb{P}(\{\hat{\mu}_1 \leq \hat{\mu}_2\} \cup \{\hat{\mu}_3 \leq \hat{\mu}_1\}) \quad (12)$$

$$\leq \mathbb{P}(\hat{\mu}_1 \leq \hat{\mu}_2) + \mathbb{P}(\{\hat{\mu}_3 \leq \hat{\mu}_2\}) \quad (13)$$

$$= 2 \cdot \mathbb{P}(\hat{\mu}_3 \leq \hat{\mu}_1) \quad (14)$$

$$= 2 \cdot \mathbb{P}\left(\frac{1}{m} \sum_j X_{j,3} - X_{j,1} \leq 0\right). \quad (15)$$

Letting $Y_j = X_{j,3} - X_{j,1}$, we have $\mathbb{E}[Y_j] = p$ and $Y_j \in [-1, 1]$, so that

$$\mathbb{P}\left(\frac{1}{m} \sum_j X_{j,3} - X_{j,1} \leq 0\right) = \mathbb{P}\left(\sum_j (Y_j - \mathbb{E}[Y_j]) \leq -pm\right) \leq \exp\left(-\frac{2p^2m^2}{\sum_j (1 - (-1))^2}\right) = \exp\left(-\frac{mp^2}{2}\right),$$

where the inequality is Hoeffding's bound for centered, bounded random variables. Consequently,

$$\mathbb{P}(\text{failure}) \leq 2 \exp\left(-\frac{mp^2}{2}\right).$$

- d) Plugging in $m = 20$, $p = 1/3$ into the above bound we obtain that $\mathbb{P}(\{\text{failure}\}) \leq 2e^{-10/9} \approx 0.66$. To have a positive expected payoff, we require:

$$\mathbb{E}[\text{gain}] = 3 \cdot \mathbb{P}(\text{success}) - 2 \cdot \mathbb{P}(\text{failure}) > 0 \iff \mathbb{P}(\text{failure}) < \frac{3}{3+2} = 0.60.$$

Although our bound is non-vacuous, $0.66 > 0.60$, so we have no guarantee of a positive gain with the above strategy and should not accept.

In general, we should accept whenever $\mathbb{P}(\{\text{failure}\})/\mathbb{P}(\{\text{success}\}) < \alpha$, i.e. when $\frac{2e^{-mp^2/2}}{1-2e^{-mp^2/2}} < \alpha \iff m > \frac{2}{p^2} \ln(2/\alpha + 2)$. Note that sharper results can be obtained by bounding $\mathbb{P}(\{\text{failure}\})$ more carefully.

Problem 5: Bandits with Infinitely Many Arms

In the course we considered bandits with a finite number of K arms. In this problem we will see that the same ideas apply if we have infinitely many arms as long as there is some additional structure.

Assume that there is an unknown unit-norm vector $\theta \in \mathbb{R}^d$. For every unit-norm vector $u \in \mathbb{R}^d$, there is a bandit. It gives the reward $X_u = \langle u, \theta \rangle + Z_u$, where Z_u is a zero-mean unit-variance Gaussian that is independent over time and independent with respect to different bandits. The nature of the reward is known to the player.

Find a policy, i.e., a strategy of what bandit to probe at any given point in time given a specific history, that has a sublinear regret as time tends to infinity. You can assume that you know the horizon, i.e., we are looking for fixed-horizon policies.

Solution 4. For a simple fixed-horizon scheme consider the following. Take the d orthonormal unit vectors e_i , $1 \leq i \leq d$. Each of those corresponds to a bandit. Dedicate an ϵ fraction of the time, i.e., $n\epsilon$ steps to exploring. In these first $n\epsilon$ steps probe each of those d bandits $m = n\epsilon/d$ times.

Note that the unknown vector θ can be written as

$$\theta = \sum_{i=1}^d \langle e_i, \theta \rangle e_i = \sum_{i=1}^d \theta_i e_i,$$

where by some abuse of notation we introduced the scalars $\theta_i = \langle e_i, \theta \rangle$. From our discussion in class we get for each such constant θ_i , m noisy estimates $\theta_i + Z$, where Z is 1-sub-Gaussian. Therefore, $\hat{\theta}_i$ has the form $\hat{\theta}_i = \theta_i + \frac{1}{m} \sum_{k=1}^m Z_m$. Hence,

$$\text{Prob}\{|\hat{\theta}_i - \theta_i| \geq \delta\} = \text{Prob}\left\{\left|\frac{1}{m} \sum_{k=1}^m Z_m\right| \geq \delta\right\} \leq 2e^{-m\delta^2/2} = 2e^{-\frac{n\epsilon\delta^2}{2d}}.$$

The expected regret for n steps can therefore be upper bounded by

$$\begin{aligned} R_n &\leq n\epsilon + n(1 - \epsilon) \left[\sum_{i=1}^d |\theta_i \cdot \theta_i - \theta_i \cdot \hat{\theta}_i| + 4de^{-\frac{n\epsilon\delta^2}{2d}} \right] \\ &\leq n\epsilon + n(1 - \epsilon) \left[\delta \sum_{i=1}^d |\theta_i| + 4de^{-\frac{n\epsilon\delta^2}{2d}} \right] \\ &\leq n\epsilon + n(1 - \epsilon) [\delta\sqrt{d} + 4de^{-\frac{n\epsilon\delta^2}{2d}}] \\ &\leq n(\epsilon + \delta\sqrt{d} + 4de^{-\frac{n\epsilon\delta^2}{2d}}). \end{aligned}$$

The explanation is as follows: In the first $n\epsilon$ steps we have a regret of at most 1 per step. In the remaining $n(1 - \epsilon)$ steps: With high probability we have estimated each component with an error of at most δ . With the small probability $2de^{-\frac{n\epsilon\delta^2}{2d}}$ we have a larger estimation error and in this case our regret is again upper bounded by a constant, namely 2, per step. Note that in the first step we first took the absolute value of the sum to obtain a simple upper bound and then used the triangle inequality. In the third step we used the fact that $\|\theta\|_1 \leq d\|\theta\|_2 = d$ for any θ .

If we pick e.g., $\epsilon = d \log(n) n^{-\frac{1}{3}}$ and $\delta = n^{-\frac{1}{3}}$ then we see that the expected regret is of the order $O(dn^{\frac{2}{3}} \log(n) + n^{\frac{2}{3}} \sqrt{d} + d\sqrt{n}) = O(n^{\frac{2}{3}} \log(n))$. This is indeed sublinear in n .

Problem 6: Upper Confidence Bound Algorithm

In the course we analyzed the Upper Confidence Bound algorithm. As was suggested in the course,

we should get something similar if instead we use the Lower Confidence Bound algorithm. It is formally defined as follows.

$$A_t = \begin{cases} t, & t \leq K, \\ \arg \max_k \hat{\mu}_k(t-1) - \sqrt{\frac{2 \ln f(t)}{T_k(t-1)}}, & t > K. \end{cases}$$

Analyze the performance of this algorithm in the same way as we did this in the course for the UCB algorithm.

Hint: Is this algorithm well designed?

Solution 5. Recall the lower bound

$$\Pr\{\hat{\mu}(X_1, \dots, X_m) \leq \mu - \epsilon\} \leq \exp(-m\epsilon^2/2)$$

If we set the right-hand side to $\delta > 0$ and solve for δ we get

$$\Pr\{\hat{\mu}(X_1, \dots, X_m) - \mu \leq \sqrt{\frac{2}{m} \ln(\frac{1}{\delta})}\} \leq \delta$$

If we consider δ as small then this suggests that, at time $t-1$, it is unlikely that our empirical estimator $\hat{\mu}_k(t-1)$ of the k -th bandit arm underestimates its mean by more than $\frac{2}{T_k(t-1)} \ln(\frac{1}{\delta})$, where $T_k(t-1)$ denotes the number of times we have chosen arm k in the first $t-1$ steps. We choose the *confidence* level δ_t as

$$\delta_t = \frac{1}{f(t)} = \frac{1}{1 + t \ln^2(t)}$$

We have the algorithm A_t shown as the problem statement.

Actually, this Lower Confidence Bound algorithm is not well designed. Consider the time $t \geq K+1$, for the k -th arm, define the

$$B_k(t) = \hat{\mu}_k(t-1) - \sqrt{\frac{2 \ln f(t)}{T_k(t-1)}}$$

Suppose that

$$k^* = \arg \max_k B_k(K+1) \tag{16}$$

Then the k^* -th arm is chosen at the $K+1$ round.

For the next round $t = K+2$, for the k^* -th arm, we have

$$\begin{aligned} B_{k^*}(K+2) &= \hat{\mu}_{k^*}(K+1) - \sqrt{\frac{2 \ln f(K+2)}{T_{k^*}(K+1)}} \\ &= \hat{\mu}_{k^*}(K+1) - \sqrt{\frac{2 \ln f(K+2)}{2}} \end{aligned}$$

Since the sample from the k^* -th arm at $K+1$ round can be large or small, we don't know which of $\hat{\mu}_{k^*}(K+1)$ and $\hat{\mu}_{(k^*)}(K)$ is larger. Thus, $B_{k^*}(K+2)$ may be larger than $B_{k^*}(K+1)$.

For the other arms other than k^* , we have

$$\begin{aligned}
B_k(K+2) &= \hat{\mu}_k(K+1) - \sqrt{\frac{2 \ln f(K+2)}{T_k(K+1)}} \\
&= \hat{\mu}_k(K) - \sqrt{\frac{2 \ln f(K+2)}{T_k(K)}} \\
&< \hat{\mu}_k(K) - \sqrt{\frac{2 \ln f(K+1)}{T_k(K)}} \\
&= B_k(K+1)
\end{aligned}$$

as $\hat{\mu}_k(K+1) = \hat{\mu}_k(K)$ and $T_k(K+1) = T_k(K) = 1$, since the k -th arm was not selected in last round.

This means that choosing the k^* -th arm at $t = K+1$ decreases the “confidence” $B_k(t)$ of other arms, while the confidence of the chosen arm k^* is not necessarily decreases. And if at $t = K+1$, we unluckily choose a suboptimal arm, we may get stuck at the suboptimal arm.

If you compare this Lower Confidence Bound algorithm with the Upper Confidence Bound algorithm in the lecture notes, you can find that in the UCB algorithm, choosing one arm in current round will reduce the increase rate of confidence of such arm in the next round compared with other unchosen arms. Thus, in the next round, it is more likely to choose other arms. As a result, every arm should be sampled enough instead of being trapped in one arm.

Problem 7: Thompson Sampling with Bernoulli Losses

This problem deals with a Bayesian approach to multi-arm bandits. Although we will not pursue this facet in the current problem, the Bayesian approach is useful since within this framework it is relatively easy to incorporate prior information into the algorithm.

Assume that we have K bandits, and that bandit k outputs a $\{0, 1\}$ -valued Bernoulli random variable with parameter $\theta_k \in [0, 1]$. Let π be the uniform prior on $[0, 1]^K$, i.e., the uniform prior on the set of all parameters $\theta = (\theta_1, \dots, \theta_K)$. Let

$$\begin{aligned}
T_k^1(t) &= |\{\tau \leq t : A_\tau = k; Y_\tau = 1\}|, \\
T_k^0(t) &= |\{\tau \leq t : A_\tau = k; Y_\tau = 0\}|.
\end{aligned}$$

In words, $T_k^1(t)$ is the number of times up to and including time t that we have chosen action k and the output of arm k was 1 and similarly $T_k^0(t)$ is the number of times up to and including time t that we have chosen action k and the output of the arm k was 0.

The goal is to find the arm with the highest parameter, i.e., the goal is to determine

$$k^* = \operatorname{argmax}_k \theta_k.$$

In the Bayesian approach we proceed as follows. At time t :

1. Compute for each arm k the distribution $p(\theta_k(t) | T_k^1(t-1), T_k^0(t-1))$.
2. Generate samples of these parameters according to their distributions.
3. Pick the arm j with the largest sample.
4. Observe the output of the j -th arm, call it $Y_j(t)$, and update the counters T_j^1 and T_j^0 accordingly.

Show that this algorithm “works” in the sense that eventually it will pick the best arm. More precisely, show the following two claims.

1. Show that $p(\theta_k(t)|T_k^1(t-1), T_k^0(t-1))$ is a Beta distributed and determine α and β .
2. Show that as t tends to infinity the probability that we choose the correct arm tends to 1. [HINT: To simplify your life, you can assume that for every arm k , $T_k^1(t-1) + T_k^0(t-1) \xrightarrow{t \rightarrow \infty} \infty$.]

NOTE: Recall that the density of the Beta distribution on $[0, 1]$ with parameters α and β is equal to

$$f(x; \alpha, \beta) = \text{constant } x^{\alpha-1} (1-x)^{\beta-1}.$$

Further, the expected value of $f(x; \alpha, \beta)$ is $\frac{\alpha}{\alpha+\beta}$ and its variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Solution 6. 1. A quick calculation shows that $p(\theta_k(t)|T_k^1(t-1), T_k^0(t-1)) = f(x; 1+T_k^1(t-1), 1+T_k^0(t-1))$. Note that this is the same calculation that we did when we showed that the Beta distribution is the conjugate prior to the Binomial distribution. Explicitly, and dropping the time index as well as the index indicating the arm, we have

$$\begin{aligned} p(\theta | T^1, T^0) &\sim p(\theta)p(T^1, T^0 | \theta) \\ &\sim \theta^{T^1} (1-\theta)^{T^0} \\ &= f(\theta; 1+T^1, 1+T^0). \end{aligned}$$

2. According to the hint and our computation above, the expected value at time t is equal to

$$\frac{1 + T_k^1(t-1)}{2 + T_k^1(t-1) + T_k^0(t-1)}.$$

By assumption $T_k^1(t-1) + T_k^0(t-1) \xrightarrow{t \rightarrow \infty} \infty$ and by the law of larger numbers $T_k^1(t-1)/(T_k^1(t-1) + T_k^0(t-1))$ and hence also $(1 + T_k^1(t-1))/(2 + T_k^1(t-1) + T_k^0(t-1))$, converges to θ_k almost surely. Therefore, our estimates for all means converge to the correct values almost surely. Further, all variances tend to 0 and hence the probability that we choose the correct arm will tend to 1 as t tends to infinity.

Problem 8: Time-Varying Bandits

Let ν be a time varying environment with K arms where all the arms except arm i has distribution $\mathcal{N}(0, 1)$ and the i -th arm has distribution $\mathcal{N}(\Delta_t, 1)$. Note that the distributions changes with time – hence the name “time-varying bandits.” Let π be our policy, where we assume that the policy does not depend on time. Our time horizon is T .

- (a) Let $\Delta_t = \frac{1}{t^p}$, where $p \in (0, 1)$. Show that for every policy the regret is upper bounded by cT^{1-p} , where c is a constant. (*Hint*: Integrals are simpler than sums.)

Solution:

$$\sum_{t=1}^T \frac{1}{t^p} \propto T^{1-p}.$$

- (b) Consider any policy whose regret scales as $o(T^{1-p})$. I.e.,

$$\lim_{T \rightarrow \infty} \frac{R_T(\nu, \pi)}{T^{1-p}} = 0$$

Show that for such a policy we must have

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T P_\nu(A_t \neq i) = 0$$

Solution:

$$R_T(\nu, \pi) = \sum_{t=1}^T P_\nu(A_t \neq i) \Delta_t$$

Suppose that

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T P_\nu(A_t \neq i) \neq 0.$$

Then

$$\begin{aligned} \exists \epsilon > 0 \ \forall T_0 \in \mathbb{N} \ \exists T \geq T_0 \ \frac{1}{T} \sum_{t=1}^T P_\nu(A_t \neq i) > \epsilon \\ \exists \epsilon > 0 \ \forall T_0 \in \mathbb{N} \ \exists T \geq T_0 \ \sum_{t=1}^T P_\nu(A_t \neq i) > T\epsilon \end{aligned}$$

Also note that since $P_\nu(A_t \neq i) \in [0, 1]$ and Δ_t is decreasing,

$$\begin{aligned} R_T(\nu, \pi) &= \sum_{t=1}^T P_\nu(A_t \neq i) \Delta_t \geq \sum_{t=T-\sum_{t=1}^T P_\nu(A_t \neq i)}^T \Delta_t \\ &\propto \int_{t=T-\sum_{t=1}^T P_\nu(A_t \neq i)}^T \Delta_t \end{aligned}$$

Therefore, for some ϵ ,

$$\begin{aligned} \forall T_0 \in \mathbb{N} \ \exists T \geq T_0 \ R_T(\nu, \pi) &\geq \int_{t=T(1-\epsilon)}^T \Delta_t \\ &= \int_{t=T(1-\epsilon)}^T \frac{1}{t^p} \\ &= \frac{1}{1-p} (1 - (1-\epsilon)^{1-p}) T^{1-p} \\ &= c T^{1-p}, \ c > 0 \end{aligned}$$

which contradicts with the statement that,

$$\lim_{T \rightarrow \infty} \frac{R_T(\nu, \pi)}{T^{1-p}} = 0$$

- (c) Now suppose that $p \in (\frac{1}{2}, 1)$. Let ν' be an environment which has the same distributions except at arm $i' \neq i$. At arm i' , ν' gives reward distributed from $\mathcal{N}(2\Delta_t, 1)$. Show that

$$\sup_{T \in \mathbb{N}} D(P_\nu(A_1, X_1, \dots, A_{T-1}, X_T) || P_{\nu'}(A_1, X_1, \dots, A_{T-1}, X_T)) < \infty$$

Solution:

Note that ν and ν' differ only in arm i' , for environment ν , arm i is $(0, 1)$ and for environment

ν' , arm i is $(2\Delta_t, 1)$. Therefore $\forall T$,

$$\begin{aligned}
& D(P_\nu(A_1, X_1, \dots, A_{T-1}, X_T, A_T) \| P_{\nu'}(A_1, X_1, \dots, A_{T-1}, X_T, A_T)) \\
&= \sum_{t=1}^T \sum_k P_\nu(A_t = k) D((0, 1) \| (2\Delta_t, 1)) \\
&= \sum_{t=1}^T P_\nu(A_t = k) 2\Delta_t^2 \\
&\leq \sum_{t=1}^T 2\Delta_t^2 \\
&\leq \sum_{t=1}^{\infty} 2\Delta_t^2 \\
&< \infty
\end{aligned}$$

(d) Show that, if,

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T P_\nu(A_t = i') = 0$$

then,

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T P_{\nu'}(A_t = i') < 1$$

(Hint: If $d_2(0 \| p)$ is a finite number, can p be very close to 1?)

Solution:

Note that by convexity of KL divergence and data processing inequality,

$$\begin{aligned}
D\left(\frac{1}{T} \sum_{t=1}^T P_\nu(A_t = i) \left\| \frac{1}{T} \sum_{t=1}^T P_{\nu'}(A_t = i)\right.\right) &\leq \frac{1}{T} \sum_{t=1}^T D(P_\nu(A_t = i) \| P_{\nu'}(A_t = i)) \\
&\leq \frac{1}{T} \sum_{t=1}^T D(P_\nu(A_1, X_1, \dots, X_T, A_T) \| P_{\nu'}(A_1, X_1, \dots, X_T, A_T)) \\
&< \infty
\end{aligned}$$

Since $\frac{1}{T} \sum_{t=1}^T P_\nu(A_t = i)$ goes to 1 by assumption, $\frac{1}{T} \sum_{t=1}^T P_{\nu'}(A_t = i)$ cannot go to 0, because otherwise the divergence would not be bounded. Therefore, $\frac{1}{T} \sum_{t=1}^T P_{\nu'}(A_t \neq i')$ cannot get arbitrarily close to 0.

(e) Show that, if for $p \in (1/2, 1)$, $R(\nu, \pi) = o(T^{1-p})$, then

$$R_T(\nu', \pi) \neq o(T^{1-p})$$

Solution:

Note that in the previous parts we have shown that if $R(\nu, \pi) = o(T^{1-p})$ and $p \in (1/2, 1)$ then

$$\begin{aligned}
& \liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T P_{\nu'}(A_t \neq i') > 0 \\
\implies & \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T P_{\nu'}(A_t \neq i') \neq 0
\end{aligned}$$

The result part 2 holds also for the environment ν' , therefore, by contrapositive of part 2,

$$\lim_{T \rightarrow \infty} \frac{R_T(\nu',)}{T^{1-p}} \neq 0$$

which is by definition, $R_T(\nu', pi) \neq o(T^{1-p})$.

- (f) Conclude that, if we are in an adversarial setting where the adversary is allowed to choose a time varying environment, no matter how long we play, $\sqrt{T}$ is the best regret we can ever hope to achieve. That is, show that, for any policy π , for any $\alpha < 1/2$ there exists a time varying environment ν s.t.

$$R_T(\nu, \pi) \neq o(T^\alpha)$$

Solution:

Choose $p = 1/2 + \epsilon$ such that $\epsilon > 0$ and $1 - p > \alpha$, either $R_T(\nu, \pi) \neq o(T^{1-p})$ or $R_T(\nu', \pi) \neq o(T^{1-p})$.

Problem 9: Thompson Sampling for Two-Armed Bernoulli Bandits

Consider the following two-arm setting:

- Let μ_1 and μ_2 be two independent samples from the uniform distribution on the interval $[0, 1]$. As in class, denote $\Delta = |\mu_1 - \mu_2|$. But note that we do not know which arm is better.
- Now, pulling arm $a \in \{1, 2\}$ at round t yields a reward
$$X_t \sim \text{Bernoulli}(\mu_a).$$
- Let $S_a(t-1)$ and $F_a(t-1)$, respectively, denote the number of observed successes (reward=1) and failures (reward=0) of arm a up to round $t-1$ (for $a=1$ and $a=2$).

We use the following algorithm (*Thompson Sampling*):

At round $t = 1, 2, \dots$:

1. Sample independently

$$\begin{aligned}\tilde{\mu}_1^{(t)} &\sim \text{Beta}(1 + S_1(t-1), 1 + F_1(t-1)) \\ \tilde{\mu}_2^{(t)} &\sim \text{Beta}(1 + S_2(t-1), 1 + F_2(t-1))\end{aligned}$$

2. Pull the arm with the larger sample:

$$A_t = \arg \max_{a \in \{1, 2\}} \tilde{\mu}_a^{(t)}.$$

3. Observe reward $X_t \in \{0, 1\}$ from arm A_t .

Let $T_{\text{bad}}(n)$ denote the number of times the bad arm (the arm with the smaller value μ_a) is pulled up to horizon n . As in class, the regret can then be written as $R_n = \Delta \mathbb{E}[T_{\text{bad}}(n)]$.

Answer the following:

1. **(Formulation)** Recall that the means μ_a were drawn from a uniform prior distribution. Suppose by time t , arm a has been pulled $T_a(t)$ times with $S_a(t)$ successes and $F_a(t)$ failures. Give an expression for the posterior distribution $p(\mu_a | S_a(t), F_a(t))$. What type of a distribution is it?

2. **(Partition of the time axis)** Show that for $m > 0$

$$\mathbb{E}[T_a(n)] \leq m + \mathbb{E}[\# \text{ of pulls of arm } a \text{ after it has already been pulled } m \text{ times}].$$

(Hint: decompose according to whether $T_a(t) < m$ or $T_a(t) \geq m$, where $T_a(t)$ is the number of pulls of arm a before round t .)

3. **(Concentration of the Beta posterior)** This part is a little more tedious. You may want to save it for last and first tackle parts (d) and (e). For a constant $c > 0$ to be chosen later, define

$$m := \left\lceil \frac{c \log n}{\Delta^2} \right\rceil.$$

Argue that once arm a has been pulled m times, the posterior distribution you found in Part (a) is concentrated around the true μ_a , in the sense that for any $\varepsilon > 0$, there exist constants $c_1, c_2 > 0$ such that, for all t with $T_a(t) \geq m$,

$$\begin{aligned} \mathbb{P}(\tilde{\mu}_a^{(t)} > \mu_a + \varepsilon \mid H_{t-1}) &\leq c_1 \exp(-c_2 T_a(t) \varepsilon^2), \\ \mathbb{P}(\tilde{\mu}_a^{(t)} < \mu_a - \varepsilon \mid H_{t-1}) &\leq c_1 \exp(-c_2 T_a(t) \varepsilon^2), \end{aligned}$$

where as in class, we are using H_{t-1} to denote the entire history up to time $t-1$, that is, $H_{t-1} = (A_1, X_1, A_2, X_2, \dots, A_{t-1}, X_{t-1})$.

Hint: Calculate $\hat{\mu}_{a,t} := \mathbb{E}[\tilde{\mu}_a^{(t)} \mid H_{t-1}]$. Then, as in class, split by conditioning onto the events G_a and G_a^c , where $G_a := \{|\hat{\mu}_{a,t} - \mu_a| \leq \frac{\varepsilon}{4}\}$.

4. **(Bounding the selection of the suboptimal arm)** Use the previous item with $\varepsilon = \Delta/2$. For large enough t (when both arms have been pulled m times), provide an upper bound for the probability that we select the bad arm,

$$\mathbb{P}(A_t = \text{bad} \mid H_{t-1})$$

in terms of some constant $c' > 0, \Delta$ and n .

5. **(Final bound on $\mathbb{E}[T_{\text{bad}}(n)]$)** Pick a large enough c and conclude that

$$\mathbb{E}[T_{\text{bad}}(n)] \leq \frac{c \log n}{\Delta^2} + O(1),$$

and hence the regret of Thompson Sampling in this two-armed Bernoulli bandit is bounded by

$$R_n = \Delta \mathbb{E}[T_{\text{bad}}(n)] = O\left(\frac{\log n}{\Delta}\right).$$

Solution 7. (a) Model and conjugacy. For arm $a \in \{1, 2\}$ put prior $\mu_a \sim \text{Beta}(1, 1)$ with density

$$p(\mu_a) = 1, \quad \mu_a \in (0, 1).$$

If by time t arm a has been pulled $N_a(t)$ times with $S_a(t)$ successes and $F_a(t)$ failures, the likelihood is

$$p(\text{data} \mid \mu_a) = \mu_a^{S_a(t)} (1 - \mu_a)^{F_a(t)}.$$

Bayes' rule gives

$$p(\mu_a \mid \text{data}) = \frac{p(\text{data} \mid \mu_a) p(\mu_a)}{\int_0^1 p(\text{data} \mid \mu) p(\mu) d\mu}.$$

Substituting the prior and likelihood,

$$p(\mu_a \mid \text{data}) = \frac{\mu_a^{S_a(t)} (1 - \mu_a)^{F_a(t)}}{\int_0^1 \mu^{S_a(t)} (1 - \mu)^{F_a(t)} d\mu}.$$

The denominator is the normalizing constant (partition function):

$$Z = \int_0^1 \mu^{S_a(t)} (1 - \mu)^{F_a(t)} d\mu = B(1 + S_a(t), 1 + F_a(t)),$$

where $B(\cdot, \cdot)$ is the Beta function.

Thus,

$$p(\mu_a \mid \text{data}) = \frac{1}{B(1 + S_a(t), 1 + F_a(t))} \mu_a^{S_a(t)} (1 - \mu_a)^{F_a(t)}.$$

which is the density of $\text{Beta}(1 + S_a(t), 1 + F_a(t))$.

(b) Time partition. Count separately the pulls of arm 2 made before it has been pulled m times, and those after. Before reaching m pulls, arm 2 can be pulled at most m times. Hence

$$N_2(n) \leq m + \sum_{t: N_2(t-1) \geq m} \mathbf{1}\{A_t = 2\}.$$

Taking expectations gives the desired inequality.

(c) Concentration. After $N_2(t) = m$ pulls of arm 2, its posterior is

$$\mu_2 \mid \mathcal{F}_t \sim \text{Beta}(1 + S_2(t), 1 + F_2(t)),$$

with $S_2(t) + F_2(t) = m$.

Our goal is to show that there exist universal constants $c_1 = 3e$ and $c_2 = \frac{1}{8}$ such that for every $\varepsilon \in (0, 1)$ and every $m \geq 1$,

$$\mathbb{P}(\tilde{\mu}_2 > \mu_2 + \varepsilon) \leq c_1 \exp(-c_2 m \varepsilon^2), \quad (17)$$

and the same holds for arm 1.

Proof. Fix $\varepsilon \in (0, 1)$ and $m \geq 1$. Write

$$\hat{\mu}_m := \frac{S_2(t)}{m} \quad \text{and} \quad \mu_m := \frac{1 + S_2(t)}{m + 2}.$$

Note that μ_m is exactly the posterior mean of $\text{Beta}(1 + S_2(t), 1 + F_2(t))$.

We now split according to a good empirical event. Define the “good” event

$$G_m := \left\{ |\hat{\mu}_m - \mu_2| \leq \frac{\varepsilon}{4} \right\}.$$

By Hoeffding’s inequality for Bernoulli variables,

$$\mathbb{P}(G_m^c) = \mathbb{P}\left(|\hat{\mu}_m - \mu_2| > \frac{\varepsilon}{4}\right) \leq 2 \exp\left(-2m \left(\frac{\varepsilon}{4}\right)^2\right) = 2 \exp\left(-\frac{m \varepsilon^2}{8}\right). \quad (18)$$

We want to bound $\mathbb{P}(\tilde{\mu}_2 > \mu_2 + \varepsilon)$. Split:

$$\begin{aligned} \mathbb{P}(\tilde{\mu}_2 > \mu_2 + \varepsilon) &= \mathbb{P}(\tilde{\mu}_2 > \mu_2 + \varepsilon, G_m) + \mathbb{P}(\tilde{\mu}_2 > \mu_2 + \varepsilon, G_m^c) \\ &\leq \mathbb{P}(\tilde{\mu}_2 > \mu_2 + \varepsilon, G_m) + \mathbb{P}(G_m^c). \end{aligned} \quad (19)$$

Next, we control the posterior mean on the good event. On G_m we have

$$|\hat{\mu}_m - \mu_2| \leq \frac{\varepsilon}{4}.$$

Since

$$\mu_m = \frac{1 + S_2(t)}{m + 2} = \frac{m}{m + 2} \hat{\mu}_m + \frac{1}{m + 2},$$

we can bound, on G_m ,

$$\begin{aligned} |\mu_m - \mu_2| &= \left| \frac{m}{m + 2} \hat{\mu}_m + \frac{1}{m + 2} - \mu_2 \right| \\ &\leq \frac{m}{m + 2} |\hat{\mu}_m - \mu_2| + \frac{1}{m + 2} \\ &\leq \frac{\varepsilon}{4} + \frac{1}{m + 2}. \end{aligned}$$

Now distinguish two sub-cases on G_m .

Case A: $m \geq \frac{4}{\varepsilon}$. Then $\frac{1}{m+2} \leq \frac{\varepsilon}{4}$, hence on G_m ,

$$|\mu_m - \mu_2| \leq \frac{\varepsilon}{2}. \quad (20)$$

In particular,

$$\mu_2 + \varepsilon \geq \mu_m + \frac{\varepsilon}{2}.$$

Next, we use the fact that Beta is sub-Gaussian around its mean. Conditioned on $S_2(t)$ (equivalently on $\mathcal{F}_{t-1}$ in the bandit notation), we have

$$\tilde{\mu}_2 \mid S_2(t) \sim \text{Beta}(1 + S_2(t), 1 + F_2(t)),$$

with total mass

$$(1 + S_2(t)) + (1 + F_2(t)) = m + 2.$$

A standard Chernoff/Hoeffding-type bound for the Beta distribution (it can be proved, e.g., by viewing Beta as a normalized Gamma/Binomial, or by log-concavity) states that there exists an absolute constant $C > 0$ such that for all $u \in (0, 1)$,

$$\mathbb{P}(\tilde{\mu}_2 - \mu_m \geq u \mid S_2(t)) \leq \exp(-C(m + 2)u^2). \quad (21)$$

We now wrap up Case A. On G_m and for $m \geq 4/\varepsilon$, we can use (20) and (21) with $u = \varepsilon/2$ to get

$$\mathbb{P}(\tilde{\mu}_2 > \mu_2 + \varepsilon \mid S_2(t)) \leq \mathbb{P}(\tilde{\mu}_2 > \mu_m + \varepsilon/2 \mid S_2(t)) \leq \exp(-C(m + 2)(\varepsilon/2)^2) \leq \exp(-C'm\varepsilon^2),$$

for some $C' = C/4$. Taking expectation over $S_2(t)$ but keeping G_m ,

$$\mathbb{P}(\tilde{\mu}_2 > \mu_2 + \varepsilon, G_m) \leq \exp(-C'm\varepsilon^2). \quad (22)$$

Plugging (22) and (18) into (19), we get for $m \geq 4/\varepsilon$,

$$\mathbb{P}(\tilde{\mu}_2 > \mu_2 + \varepsilon) \leq 2 \exp\left(-\frac{m\varepsilon^2}{8}\right) + \exp(-C'm\varepsilon^2) \leq 3 \exp\left(-\frac{m\varepsilon^2}{8}\right),$$

after adjusting constants (take $c_2 = 1/8$).

Case B: $1 \leq m < \frac{4}{\varepsilon}$. In this regime we simply use the trivial bound

$$\mathbb{P}(\tilde{\mu}_2 > \mu_2 + \varepsilon) \leq 1.$$

At the same time, since $m < 4/\varepsilon$,

$$\exp\left(-\frac{m\varepsilon^2}{8}\right) \geq \exp\left(-\frac{(4/\varepsilon)\varepsilon^2}{8}\right) = \exp\left(-\frac{\varepsilon}{2}\right) \geq e^{-1/2}.$$

Hence

$$1 \leq 3e \cdot \exp\left(-\frac{m\varepsilon^2}{8}\right).$$

Therefore, for all $1 \leq m < 4/\varepsilon$ the desired bound

$$\mathbb{P}(\tilde{\mu}_2 > \mu_2 + \varepsilon) \leq 3e \cdot \exp\left(-\frac{m\varepsilon^2}{8}\right)$$

also holds.

Finally, we unify the two cases. Combining Case A and Case B, we have shown that for every $m \geq 1$ and every $\varepsilon \in (0, 1)$,

$$\mathbb{P}(\tilde{\mu}_2 > \mu_2 + \varepsilon) \leq 3e \cdot \exp\left(-\frac{m\varepsilon^2}{8}\right).$$

This proves (17) with $c_1 = 3e$ and $c_2 = \frac{1}{8}$. □

(d) Bounding suboptimal selection. Choose $\varepsilon = \Delta/2$. Then, using

$$\{\tilde{\mu}_2^{(t)} > \tilde{\mu}_1^{(t)}\} \subseteq \{\tilde{\mu}_2^{(t)} > \mu_2 + \Delta/2\} \cup \{\tilde{\mu}_1^{(t)} < \mu_1 - \Delta/2\},$$

we get

$$\mathbb{P}(A_t = 2 \mid \mathcal{F}_{t-1}) \leq c_1 \exp\left(-c_2 m \frac{\Delta^2}{4}\right) + c_1 \exp\left(-c_2 m \frac{\Delta^2}{4}\right).$$

Since $m \asymp (\log n)/\Delta^2$, the first term is at most $n^{-c'_1}$ for some $c'_1 > 0$. Under TS, the optimal arm 1 is also pulled $\Omega(\log n)$ times, so the second term is at most $n^{-c'_2}$. Hence

$$\mathbb{P}(A_t = 2 \mid \mathcal{F}_{t-1}) \leq n^{-c'}$$

for some $c' > 0$.

(e) Final bound. Summing over t ,

$$\mathbb{E}[\# \text{ of pulls of arm 2 after } m] \leq \sum_{t=1}^n n^{-c'} = O(1) \quad (\text{for } c' > 1).$$

Therefore

$$\mathbb{E}[N_2(n)] \leq m + O(1) = O\left(\frac{\log n}{\Delta^2}\right),$$

and

$$\mathbb{E}[\text{Regret}(n)] = \Delta \mathbb{E}[N_2(n)] = O\left(\frac{\log n}{\Delta}\right).$$