Foundations of Data Science
Urbanke

Ecole Polytechnique Fédérale, Lausanne: Fall 2022
October 27, 2022

Problem Set 4

For the Exercise Sessions on Oct 28 and Nov 4

Last name	First name	SCIPER Nr	Points

Problem 1: 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 1. Recall the lower bound

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

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

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

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

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

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 2: 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.

Hint: Start with the simplest thing you can think of. If you do not have time to do the math, describe in words the basic idea of your strategy and why it should give us a sublinear regret.

Solution 2. 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,

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

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

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

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 3: 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 $\sqrt{n}$ if we use the gaps when determining m and of order $n^{\frac{2}{3}}$ if we do not use the gaps in order to determine m.

Here is a slightly different algorithm. Let $\epsilon_t = t^{-\frac{1}{3}}$. For each round $t = 1, \ldots,$ 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.

Show that for this algorithm the expected regret at *any* time t is upper bounded by $t^{\frac{2}{3}}$ times terms in t and K of lower order. This is a similar to the worst-case of the explore-then-exploit strategy but here we do not need to know the horizon a priori. Assume that the rewards are in $[0, 1]$.

Solution 3. 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

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

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

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

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 $\hat{u}_j(i-1)$ 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:

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

Where (a) comes from the law of total Expectation and (b) from the σ-sub Gaussianity assumption. Now observe that by the exploration phase the probability that $T_k(i-1) < (i-1)^{2/3}/K$ is low. Indeed, at every step t there is a probability $t^{-1/3}/K$ that we increase it. Let $B_k(t)$ be a Bernouli 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 $\frac{1}{K}\sum_{t=1}^{i-1} t^{-1/3} \leq \frac{3(i-1)^{2/3}}{K}$.

$$
\begin{aligned}
\mathbb{P}\Big[T_k(i-1) < (i-1)^{2/3}/K\Big] &\leq \mathbb{P}\Big[\sum_{t=1}^{i-1} B_k(t) < (i-1)^{2/3}/K\Big]\\
&= \mathbb{P}\Big[-\sum_{t=1}^{i-1} B_k(t) > -(i-1)^{2/3}/K\Big]\\
&\leq \mathbb{P}\Big[\frac{3(i-1)^{2/3}}{K} - \sum_{t=1}^{i-1} B_k(t) > \frac{3(i-1)^{2/3}}{K} - \frac{(i-1)^{2/3}}{K}\Big]\\
&\leq \mathbb{P}\Big[\nu - \sum_{t=1}^{i-1} B_k(t) > \frac{2(i-1)^{2/3}}{K}\Big]\\
&\overset{(\star)}{\leq} \frac{3K}{4(i-1)^{2/3}}
\end{aligned}
$$

where $(\star)$ follows by Chebychev inequality with the function $f(x) = x^2$, *i.e.* $\mathbb{P}[X > a] \leq \mathbb{E}[X^2]/a^2$ and

ν denotes the mean of $\sum_t B_k(t)$. Thus:

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

Hence, the average regret at round t is given by

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

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

Problem 4: These Bandits - Exp3

Consider an adversarial bandit setting with K bandits, where the rewards are arbitrary numbers $x_{t,k} \in [0,1]$ (t stands for the time index and runs from 1 to n and k is the index of the bandit, which goes from 1 to K). You are the adversary, in charge of designing the rewards. You know that the policy that is used is the exp3 algorithm.

Your task is to fill in the numbers. You are given the constraint that the "average" value of all rewards must be $\frac{1}{2}$, where the "average" means the sum over all $n \times K$ entries divided by $n \times K$.

Your aim is to make the expected reward (not regret) of the player as small as possible.

(i) In general, what is the expected reward the player gets in this adversarial setting when using the exp3 algorithm? State the reward normalized by the time n. We are only interested in the first order term, i.e., the constant, and not higher order terms that vanish with n.

(ii) Explain how you fill in the numbers to minimize the expected reward and compute this reward. As before, our interest is in the first order term.

Solution 4. We know that, using the Exp3 algorithm our expected regret (normalized by the time) is sublinear. In other words, we will do almost as good as if we knew the reward matrix and could choose that bandit that has the highest reward, i.e., that row of the matrix whose sum is maximal.

Therefore our task is to make sure that the maximum row sum is as small as possible.

By assumption the "average" value of all rewards is $\frac{1}{2}$. Note that we can compute this average in various ways. In particular we can first compute the reward for each arm and then average over the players.

We conclude that the average reward, when averaged over the bandits must be $\frac{1}{2}$. And our aim is to minimize the maximum reward under this average constraint.

Therefore, if we want to minimize the expected reward that the player will get it is best if we make all bandits to have the same expected reward. Hence, fill in all rows (rewards of a bandit) in such a way that its expected value of this row is $\frac{1}{2}$. To first order, the expected reward that the player will get is then $\frac{1}{2}$.

Problem 5: 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 $\theta = (\theta_1, \cdots, \theta_K)$. Let

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

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 choses 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

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

In the Bayesian approach we proceed as follows. At time 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, $T_k^1(t-1) + T_k^0(t-1) \overset{t\to\infty}{\to} \infty$.]

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 5. 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. Explicity, and dropping the time index as well as the index indicating the arm, we have

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

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 $T_k^1(t-1)+T_k^0(t-1) \overset{t\to\infty}{\to} \infty$ 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.