Foundations of Data Science
Gastpar & Urbanke

Ecole Polytechnique Fédérale, Lausanne: Fall 2023
October 24, 2023

Problem Set 4 **(Graded)** —*Due Tuesday, November 7, before class starts*

For the Exercise Sessions on Oct 24 and Oct 31

Last name	First name	SCIPER Nr	Points

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 $\sqrt{n}$ 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 $t = 1, \dots,$ 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.

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 $[0, 1]$.

Problem 2: 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

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

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

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

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 (1) for our case?

(b) Look at the first of the two terms on the right of (1) 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.

Problem 3: 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

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

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 $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Problem 4: 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.)

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

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

Show that for such a policy we must have

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

(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

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

(d) Show that, if,

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

then,

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

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

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

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

(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.

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