Foundations of Data Science
Gastpar & Urbanke

Ecole Polytechnique Fédérale, Lausanne: Fall 2024
September 30, 2024

Problem Set 3

For the Exercise Sessions on October 1 and 8

Last name	First name	SCIPER Nr	Points

Problem 1: 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.,

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

Show that for such a policy we must have

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

Solution:

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

Suppose that

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

Then

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

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

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

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

Therefore, for some ϵ,

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

which contradicts with the statement that,

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

(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

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

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 $\forall T$,

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

(d) Show that, if,

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

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?)

Solution:

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

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

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

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

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

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

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 2: Bandits

In the course we mentioned *contextual bandits*. E.g., imagine that you suspect that the rewards that you get from the various arms depend on the day of the week (more generally, a particular feature of the input, aka the "context"). In this case it makes sense to run a separate bandit algorithm for each of the possible contexts. As we discussed, the downside of this approach is that we now have less data for each of the bandit algorithms and hence it will take us longer to learn.

Let c be the context, where $c \in \mathcal{C}$. We run a separate bandit algorithm for each of the $|\mathcal{C}|$ possible contexts and the total number of steps we take is n. We further assume that the number of arms is K and the K does not depend on the context. We use the same bandit algorithm for each of the contexts and we assume that this bandit algorithm has an expected regret of $O(\sqrt{Kn\log(n)})$ (when run over n time steps for a fixed context) in the stochastic setting.

Consider the following setup. For each context there is a different stochastic bandit with K arms. At each time a context is sampled independently at random, the player sees the context, chooses and arm, and receives a reward according to the distribution P_i^c (reward distribution for context c and arm i).

Define the expected regret for this setup as the difference between the expected reward that we could have gotten if for each context we would have chosen the arm with the largest mean minus the expected reward we get using our scheme.

1. Show that the expected regret at time n is $O(\sqrt{Kn|\mathcal{C}|\ln(n)})$.

   Discussion: Compared to the case of a single context note that you "only" pay a factor of $\sqrt{|\mathcal{C}|}$. If the number of contexts is small, this might be acceptable. Note further, that this bound is valid, regardless how often each context appears. *Hint:* Start by assuming that each context $c \in \mathcal{C}$ appears n_c times.

   Solution: In the sequel let $\{n_c\}_{c\in\mathcal{C}}$ denote the number of times we see context c. The regret is defined as,

$$R \triangleq \mathbb{E}[\sum_{c\in\mathcal{C}} \max_{i\in[k]} n_c \mu_i^{(c)} - \sum_{t:c_t=c} X_t]$$

   Then, switching the expectation and summation over $\mathcal{C}$,

$$R = \sum_{c\in\mathcal{C}} R_c$$

   where we define R_c as the regret for the times we see context c. Then, the expected regret for the contextual bandit algorithm is equal to

$$\begin{aligned}
\mathbb{E}[R] &= \mathbb{E}[\mathbb{E}[R|\{n_c\}_{c\in\mathcal{C}}]] \\
&\leq \sup_{\{n_c\}_{c\in\mathcal{C}}:\sum_c n_c = n} \sum_{c\in\mathcal{C}} \mathbb{E}[R_c|\{n_c\}_{c\in\mathcal{C}}] \\
&= \sup_{\{n_c\}_{c\in\mathcal{C}}:\sum_c n_c = n} \sum_{c\in\mathcal{C}} O(\sqrt{Kn_c\log(n_c)}) \\
&\leq \sup_{\{n_c\}_{c\in\mathcal{C}}:\sum_c n_c = n} \sum_{c\in\mathcal{C}} O(\sqrt{Kn_c\log(n)}) \\
&\leq O(\sqrt{Kn|\mathcal{C}|\log(n)}),
\end{aligned}$$

   where in the last step we have used the fact that under the condition that $n = \sum_c n_c$ the bound is maximized by the choice $n_c = n/|\mathcal{C}|$. This can be seen by writing the Lagrangian $\sum_c \sqrt{n_c} + \lambda(\sum_c n_c - n)$ and taking the derivative with respect to the n_c. This gives $\frac{\partial L}{\partial n_c} = \frac{1}{2}\frac{1}{\sqrt{n_c}} + \lambda$. Setting those derivatives to 0 we see that all n_c should have the same size, i.e., $n_c = n/|\mathcal{C}|$.

2. Assume now that $\mathcal{C} = [0, 1)$, i.e., $\mathcal{C}$ is very large, and in fact uncountable. In this case we cannot use the strategy above. What would you do in such a case? Under what circumstances will such a scheme likely work? If we assume a fixed number of arms and a very large time horizon, how would you expect the regret to scale?

   Solution: Assume that the means of all the K arms are Lipschitz functions of the context c. Hence, by slightly changing the context c, the means only change slightly. In this case it makes sense to quantize the domain $[0,1)$, i.e., write $[0,1) = \cup_{i=0}^{m}[\delta i, \delta(i+1))$, where $\delta = 1/m$ for some natural number m.

Since the function is Lipschitz, we have for each $1 \leq i \leq m$,

$$E[R|c \in [\delta(i-1), \delta i]] \leq E[R|c = \delta i - \delta/2] + O(n/m).$$

Plugging this upper bound to our derivation in the previous point gives us,

$$E[R] \leq O(\sqrt{Knm\log(n)}) + O(n/m).$$

This bound is optimized when $m \in O\left(\sqrt[3]{\frac{n}{k\log n}}\right)$. The regret due to this choice of m scales as

$$E[R] \leq O(\sqrt[3]{Kn^2\log(n)}),$$

which is still sub-linear.

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

Solution 1.

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.

Problem 4: 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 2. For a simple fixed-horizon scheme consider the following. Take the d orthonormal unit vectors e_i, $1\le i\le d$. 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 θ_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,

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

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.