Sequential Learning: Homework #2

Same comments as for Homework #1. I care about well-written proofs: with sufficient details, with calculations worked out and leading to pleasant and readable bounds. I favor quality of the writing over the quantity of questions answered. I give bonus points for elegant solutions.

Sequential submissions, deadline. Please solve the exercises one by one, in the order indicated below. After solving an exercise, send me your answer in a single PDF file named in the format `HW2-Ex1-YourName.pdf` or `HW2-Pb-YourName.pdf`. Allow me 2 business days to provide feedback on the exercise: whether I am satisfied or not with its solution. Only if I am satisfied you can proceed and send your next solution. Deadline for your final submission is Wednesday, March 31st, at 8pm. Please start early to allow for the first iterations, do not wait for the last minute.

Order to be followed: Exercise 1; then, Exercise 2 or Exercise 3; then, the other exercise between Exercise 2 and Exercise 3; then, Exercise 4; finally, the Problem.

The minimal requirement is to submit two and a half exercises (meaning: two exercices that I am satisfied with, plus something substantial for a third exercise).

Beware: Typos. Most likely the statement comes with typos. This is part of the job. Try to correct them on your own!

Exercise 1: Explore then commit

Consider a stochastic bandit setting with $K = 2$ arms only, each associated with a probability distribution ν_1, ν_2 over $[0, 1]$, with respective expectations μ_1, μ_2. Assume that you have to play for a given horizon $T \geqslant 4$. Explore each arm by pulling it m times, where $1 \leqslant m \leqslant T/2$. Compute the empirical averages $\widehat{\mu}_{1,m}$ and $\widehat{\mu}_{2,m}$ obtained. For the remaining $T - 2m$ steps, play only the arm j with maximal empirical average $\widehat{\mu}_{j,m}$ (ties broken arbitrarily). What is the regret of this strategy (called "explore then commit")?

For the analysis, we will assume with no loss of generality that arm 1 is the optimal arm and we will denote by $\Delta = \mu_1 - \mu_2$ the gap between the expectations associated with the two arms.

1. Show that $\mathbb{P}\big\{\widehat{\mu}_{1,m} < \widehat{\mu}_{2,m}\big\} \leqslant \exp\big(-m\Delta^2/c\big)$ where c is a constant (provide a numerical value).

2. Conclude that the regret is bounded by $m\Delta + (T - 2m)\Delta \exp\big(-m\Delta^2/c\big)$.

3. Assume that T and the range $[0, 1]$ are known. How should we choose m? Show a distribution-free bound on the regret that is a $o(T)$ — but it does not need to be of the typical $\sqrt{T}$ order of magnitude, it can be (much) larger. *Reminder: "distribution-free" means that the bound should only depend on T and on $[0, 1]$, not on the specific bandit problem considered, e.g., not on Δ.*

Exercise 2: Distribution-free lower bound for K–armed bandits

As indicated in class, one of the exercises of the present homework is devoted to proving that in the stochastic K–armed bandit setting, i.e., when K arms with respective distributions $\nu_1, \ldots, \nu_K$ over $[0,1]$ (with expectations denoted by $\mu_1, \ldots, \mu_K$) are available, no strategy $\mathcal{S}$ can have a sharper distribution-free regret bound than one of the order $\sqrt{KT}$.

More precisely, we denote by Y_t the reward obtained at each round, when picking arm I_t; we recall that Y_t is drawn at random according to ν_{I_t} conditionally to I_t. The regret is defined as

$$R_T = T \max_{k=1,\ldots,K} \mu_k - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right].$$

You will prove that for all $K \geqslant 2$ and all $T \geqslant K/5$,

$$R_T^\star = \inf_{\mathcal{S}} \sup_{\underline{\nu}} R_T \geqslant \frac{1}{20}\sqrt{KT},$$

where the defining infimum of $R_T^\star$ is over all strategies $\mathcal{S}$ and the supremum is over all K–tuples of distributions $\underline{\nu} = (\nu_1, \ldots, \nu_K)$ over $[0,1]$.

As the proof will reveal, it actually suffices to consider Bernoulli distributions. Indeed, let $\varepsilon \in (0,1)$ and consider the K–tuples $\underline{\nu}^{(0)}, \underline{\nu}^{(1)}, \ldots, \underline{\nu}^{(K)}$ defined based on the Bernoulli distributions $\mathrm{B}_+ = \mathrm{Ber}(1/2 + \varepsilon/2)$ and $\mathrm{B}_- = \mathrm{Ber}(1/2 - \varepsilon/2)$ as follows:

- In Model 0, all arms are associated with B_-, that is, $\underline{\nu}^{(0)} = (\mathrm{B}_-, \ldots, \mathrm{B}_-)$.
- In Model $i \in \{1, \ldots, K\}$, all arms are associated with B_- except the i–th arm, which is associated with B_+.

We denote by $\mathbb{P}_i$ the probability induced by Model i, for $i \in \{0, 1, \ldots, K\}$, and by $\mathbb{E}_i$ the corresponding expectation. We denote by $N_k(T)$ the number of times arm k was pulled by the considered strategy till round T included.

1. Explain why

$$R_T^\star \geqslant \inf_{\mathcal{S}} \sup_{\varepsilon \in (0,1)} \max_{i \in \{1,\ldots,K\}} \varepsilon\Big(T - \mathbb{E}_i\big[N_i(T)\big]\Big)$$

and why there exists k_0 such that $\mathbb{E}_0\big[N_{k_0}(T)\big] \leqslant T/K$.

2. Use the fundamental inequality for proving lower bounds in stochastic bandit problems and Pinsker's inequality to get, for all strategies $\mathcal{S}$,

$$\mathbb{E}_0\big[N_{k_0}(T)\big]\ \mathrm{KL}(\mathrm{B}_-, \mathrm{B}_+) \geqslant 2\Big(\mathbb{E}_0\big[N_{k_0}(T)/T\big] - \mathbb{E}_{k_0}\big[N_{k_0}(T)/T\big]\Big)^2.$$

3. Combine the results above to derive

$$R_T^\star \geqslant \inf_{\mathcal{S}} \sup_{\varepsilon \in (0,1)} \varepsilon T \left(1 - \frac{1}{K} - \sqrt{\frac{T}{2K}\,\mathrm{KL}(\mathrm{B}_-, \mathrm{B}_+)}\right)$$

and conclude to the desired bound. You may use that

$$\varepsilon \in (0,\, 1/2) \longmapsto 2.5\,\varepsilon^2 - \varepsilon \ln \frac{1+\varepsilon}{1-\varepsilon}$$

takes positive values.

Exercise 3: The (α, ψ)–UCB algorithm

Let $\psi : \mathbb{R} \to \mathbb{R}$ be a convex function such that $\psi(x) = \psi(-x)$ for all $x \in \mathbb{R}$. Consider a bandit model $\mathcal{D}$ such that for all $\nu \in \mathcal{D}$, if X denotes a random variable with distribution ν, then

$$\forall \lambda \geqslant 0, \qquad \max\left\{ \ln \mathbb{E}_\nu\Big[\mathrm{e}^{\lambda(X-\mathbb{E}[X])}\Big], \ln \mathbb{E}_\nu\Big[\mathrm{e}^{\lambda(\mathbb{E}[X]-X)}\Big] \right\} \leqslant \psi(\lambda)\,. \tag{$\star$}$$

For all $x \geqslant 0$, we define the convex conjugate of ψ,

$$\psi^\star(x) = \sup\big\{\lambda x - \psi(\lambda) : \lambda \geqslant 0\big\}\,,$$

and assume that $\psi^\star$ is invertible, with inverse denoted by $(\psi^\star)^{-1}$.

1. Provide such a function ψ for the model $\mathcal{D} = \mathcal{P}\big([0,1]\big)$ of all probability distributions over $[0,1]$. Compute $\psi^\star$ and its inverse.

We generalize the UCB algorithm for stochastic bandits in the following way. We consider the same setting and use the same notation as the ones used in class and in Exercise 3 of the present statement: a stochastic bandit problem is formed by $K \geqslant 2$ probability distributions $\nu_1, \ldots, \nu_K$ in $\mathcal{D}$ with respective expectations μ_k, their maximal expectation is denoted by $\mu^\star$, the gap of arm k is $\Delta_k = \mu^\star - \mu_k$, etc.

(α, ψ)–UCB algorithm

Parameters: $\alpha > 0$ and $\psi : \mathbb{R} \to \mathbb{R}$ with $\psi(x) = \psi(-x)$ for all $x \geqslant 0$

Initialization: Play each arm once, i.e., $I_t = t$ for $t \in \{1, \ldots, K\}$, get a reward $Y_t \sim \nu_t$

For $t \geqslant K+1$,

1. Compute, for all $k \in \{1, \ldots, K\}$,

$$N_k(t-1) = \sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = k\}} \qquad \text{and} \qquad \widehat{\mu}_{k,t-1} = \frac{1}{N_k(t-1)} \sum_{s=1}^{t-1} Y_s \mathbb{1}_{\{I_s = k\}}$$

2. Pick an arm (ties broken arbitrarily)

$$I_t \in \operatorname*{argmax}_{k \in \{1,\ldots,K\}} \left\{ \widehat{\mu}_{k,t-1} + (\psi^\star)^{-1}\left(\frac{\alpha \ln t}{N_k(t-1)} \right) \right\}$$

3. Get a reward $Y_t \sim \nu_{I_t}$ (conditionally to I_t)

We want to upper bound the pseudo-regret of the (α, ψ)–UCB algorithm as follows: for $\alpha > 2$,

$$\overline{R}_T = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^T Y_t\right] \leqslant \sum_{k:\Delta_k > 0} \Delta_k \left(\frac{\alpha}{\psi^\star(\Delta_k/2)} \ln T + \frac{2\alpha}{\alpha - 2} \right). \tag{B}$$

To that end, we first show that for each arm k and $t \geqslant K+1$, an upper confidence bound on μ_k is given by

$$\widehat{\mu}_{k,t-1} + (\psi^\star)^{-1}\left(\frac{\alpha \ln t}{N_k(t-1)} \right).$$

2. Prove that for all $t \geqslant 1$ and all $\lambda \geqslant 0$,

$$\mathbb{E}\left[\exp\Big(-\lambda (Y_t - \mu_k)\, \mathbb{1}_{\{I_t = k\}} \Big) \,\middle|\, \mathcal{F}_{t-1} \right] \leqslant \exp\big(\psi(\lambda)\, \mathbb{1}_{\{I_t = k\}} \big)$$

for a filtration $\mathcal{F} = (\mathcal{F}_t)_{t \geqslant 0}$ to specify explicitly.

Construct an $\mathcal{F}$–adapted supermartingale $(M_t)_{t \geqslant 0}$ based on this inequality.

3. Prove that for all $t \geqslant K+1$, all $\ell \geqslant 1$, and all $\varepsilon > 0$,

$$\mathbb{P}\Big\{\widehat{\mu}_{k,t-1} + \varepsilon \leqslant \mu_k \;\text{ and }\; N_k(t-1) = \ell\Big\} \leqslant \exp\big(-\ell\, \psi^\star(\varepsilon)\big)\,.$$

4. Provide a bound, for $t \geqslant K+1$, on

$$\mathbb{P}\left\{\widehat{\mu}_{k,t-1} + (\psi^\star)^{-1}\left(\frac{\alpha \ln t}{N_k(t-1)}\right) \leqslant \mu_k\right\}.$$

5. Briefly indicate how to bound, for $t \geqslant K+1$,

$$\mathbb{P}\left\{\widehat{\mu}_{k,t-1} - (\psi^\star)^{-1}\left(\frac{\alpha \ln t}{N_k(t-1)}\right) > \mu_k\right\}.$$

To establish the regret bound, we first fix a suboptimal arm j and an optimal arm $a^\star$.

6. Explain why $I_t = j$ for $t \geqslant K+1$ entails one of the following events:

$$\widehat{\mu}_{a^\star,t-1} + (\psi^\star)^{-1}\left(\frac{\alpha \ln t}{N_{a^\star}(t-1)}\right) \leqslant \mu^\star\,,$$

$$\text{or} \qquad \widehat{\mu}_{j,t-1} - (\psi^\star)^{-1}\left(\frac{\alpha \ln t}{N_j(t-1)}\right) > \mu_j\,,$$

$$\text{or} \qquad N_j(t-1) < \frac{\alpha \ln t}{\psi^\star(\Delta_j/2)}\,.$$

7. Establish the regret bound (B).

We conclude this exercice with a discussion of the bound for the model $\mathcal{D} = \mathcal{P}\big([0,1]\big)$.

8. Provide also a distribution-free bound for (α, ψ)–UCB on this model, i.e., a bound over all distributions satisfying $(\star)$. You need first to think of a suitable value for α.

Exercise 4: Stochastic bandits with a continuum of arms

Consider the setting of stochastic bandits with a continuum of arms indexed by $\mathcal{A} = [0,1]$. A bandit problem is given by a mean-payoff function f, which we assume to be continuous, thus bounded; for simplicity we consider $f : [0,1] \to [0,1]$. When the player picks arm $X_t \in [0,1]$ at round t, she gets a payoff $Y_t \in [0,1]$ drawn at random according to a distribution with expectation $f(X_t)$, conditionally to X_t. The question is to upper bound the regret defined as

$$R_T = T \max_{[0,1]} f - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right].$$

Consider the following two-stage methodology. We divide $[0,1]$ into the $K \geqslant 2$ regular intervals $[(i-1)/K,\, i/K]$, for $i \in \{1,\dots,K\}$. An auxiliary algorithm picks an interval index $I_t \in \{1,\dots,K\}$. An arm X_t is then drawn at random within bin I_t and a payoff Y_t is obtained.

We already considered this setting and strategy in an exercise in the lecture notes, with Lipschitz mean-payoff functions and the UCB algorithm.

In the present exercise, we rather assume that the mean-payoff function f is α–Hölder, for some $\alpha > 0$: there exists $L > 0$ such that for all $x, x' \in [0,1]$,

$$\left|f(x) - f(x')\right| \leqslant L|x - x'|^{\alpha}.$$

Also, we rather consider MOSS as the auxiliary algorithm; we recall that its distribution-free regret bound is $K + 45\sqrt{KT}$ against K–tuples of probability distributions over $[0,1]$.

1. Show that for a fixed number of bins $K \geqslant 2$ and a fixed horizon $T \geqslant 2$, the regret of the two-stage strategy above is upper bounded by $K + 45\sqrt{KT} + TL/K^{\alpha}$.
2. Explain how to pick K and which regret bound is obtained when T and α are known.
3. What could we do when α is known but T is unknown? Provide the (order of magnitude of the) corresponding regret bound.

Dealing with an unknown α is much more challenging, but was optimally solved by Hédi Hadiji in his PhD thesis, which he defended in December 2020.

Problem: Adaptation to the range for K–armed bandits

So far we only considered K–armed bandit problems $\nu_1, \dots, \nu_K$ with distributions over a known interval, typically set to $[0,1]$ with no loss of generality. Can the player learn the range? I.e., minimize the regret when the distributions $\nu_1, \dots, \nu_K$ are supported on a bounded range $[m, M]$ but the player ignores m and M? The answer is "Yes" and a strategy to do so can be based on the fully adaptive version of the exponentially weighted average strategy studied in class. We will refer to this strategy as FA-EWA in the sequel.

We use our typical notation: at each round, the player picks an arm I_t, a payoff Y_t is drawn at random according to ν_{I_t} given this choice I_t; expectations are denoted by $\mu_1, \dots, \mu_K$, with maximal value $\mu^\star$; etc.

First case: an element $C \in [m, M]$ is known

We consider an auxiliary strategy outputting probability distributions $p_t = (p_{1,t}, \dots, p_{K,t})$ over the arms, at round $t \geqslant 1$. We also consider a non-increasing sequence $\gamma_t \in (0,\, 1/2]$. We draw the arm I_t at random according to the probability distribution q_t defined by

$$q_{j,t} = (1 - \gamma_t) p_{j,t} + \frac{\gamma_t}{K} \, .$$

The auxiliary strategy is actually given by FA-EWA run on the losses

$$\ell_{j,t} = \frac{-(Y_t - C) \mathbb{1}_{\{I_t = k\}}}{q_{j,t}} + C \, .$$

This strategy indeed has no knowledge of m and M (but requires an element $C \in [m, M]$).

Some useful (in)equalities. First prove the following statements.

1. For all $j \in \{1, \dots, K\}$ and all $t \geqslant 1$,

$$|\ell_{j,t} - C| \leqslant \frac{M - m}{\gamma_t / K} \, .$$

2. Define a filtration $\mathcal{F}$ such that for all $j \in \{1, \dots, K\}$ and all $t \geqslant 1$,

$$\mathbb{E}[\ell_{j,t} \,|\, \mathcal{F}_{t-1}] = \mu_j \, .$$

3. For all $j \in \{1, \dots, K\}$ and all $t \geqslant 1$, we have $\gamma_t \leqslant 1/2$ thus $p_{j,t} \leqslant 2 q_{j,t}$ and

$$\mathbb{E}\big[p_{j,t} (\ell_{j,t} - C)^2\big] \leqslant 2 (M - m)^2 \, .$$

Recall that FA-EWA guarantees that for all ranges $[a, b]$, for all sequences of losses $L_{j,t} \in [a, b]$, for all $T \geqslant 1$,

$$R_T \leqslant 2 \sqrt{\sum_{t=1}^{T} v_t \ln N} + 4 (b - a) \ln N \, ,$$

where R_T is some regret and where the v_t are some variance factors.

4. Recall how R_T and v_t are defined; also pin point the slight simplification performed for the sake of readability in the second-order term $4(b-a) \ln N$ compared to what we proved in class.

Substituting the regret bound of FA-EWA

5. Substitute the regret bound of FA-EWA and some of the useful (in)equalities proved above to get

$$\sum_{t=1}^{T} \sum_{j \in \{1, \dots, K\}} p_{j,t} \ell_{j,t} - \min_{k \in \{1, \dots, K\}} \ell_{k,t} \leqslant 2 \sqrt{\sum_{t=1}^{T} \sum_{j \in \{1, \dots, K\}} p_{j,t} (\ell_{j,t} - C)^2} + \frac{4 (M - m) \ln N}{\gamma_T / K} \, .$$

6. Note that

$$\sum_{j\in\{1,\dots,K\}} q_{j,t}\ell_{j,t} = -Y_t$$

and deduce from the previous question a bound on

$$-\sum_{t=1}^{T} Y_t - \min_{k\in\{1,\dots,K\}} \ell_{k,t}\,.$$

7. Take expectations in the inequality obtained to prove

$$T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] \leqslant 3(M-m)\sqrt{KT\ln K} + 10(M-m)\frac{K\ln K}{\gamma_T} + (M-m)\sum_{t=1}^{T}\gamma_t\,.$$

8. Provide a final regret bound of order $\sqrt{T}$.

Second case: getting rid of the knowledge of C

9. How can the strategy above be adapted so that no knowledge of an element $C \in [m, M]$ is required, without degradating too much the regret bound?