Bandits Assignment
2MMS50 – Stochastic Decision Theory

Jaron Sanders

May 19, 2022

1 Objective

In this assignment you will restrict yourself to the class of *normal bandits* and prove a *lower bound for the pseudo-regret.* Intriguingly, we can derive a lower bound for *any* decision policy.

Generic lower bounds such as this are used to theoretically gauge the performance of specific decision policies. Decision policies are particularly good when an upper bound can be derived for them that (asymptotically) matches the accompanying lower bound. You may think of the squeeze theorem used in calculus.

2 Preliminaries

Let $P = (P_1, \ldots, P_K)$ be the reward distributions associated with one K-armed bandit (say B_P), and let $Q = (Q_1, \ldots, Q_K)$ be the reward distributions associated with another K-armed bandit (say B_Q). Fix some decision policy π. Let $\mathbb{P}_\pi$ and $\mathbb{Q}_\pi$ be the probability measures on the stochastic bandit model induced by the t-round interconnection of π and P, and π and Q, respectively.

The following decomposition of the *Kullback–Leibler divergence between measures* $\mathbb{P}_\pi$ *and* $\mathbb{Q}_\pi$ can then be proven:

$$\mathrm{KL}(\mathbb{P}_\pi, \mathbb{Q}_\pi) = \sum_{k=1}^{K} \mathbb{E}_{\mathbb{P}_\pi}[T_k(t)] \mathrm{KL}(P_k, Q_k).$$

To evaluate the right-hand side of this equation, recall that the *Kullback–Leibler divergence between distributions* P_1 *and* Q_1 *of two continuous random variables* is given by

$$\mathrm{KL}(P_1, Q_1) = \int_{-\infty}^{\infty} p_1(x) \ln \frac{p_1(x)}{q_1(x)} \,\mathrm{d}x.$$

Here, p_1 and q_1 denote the probability density functions of P_1 and Q_1, respectively.

Also recall that in K-armed stochastic bandit problems the forecaster selects at each time $s \in \mathbb{N}_+$ some arm $I_s \in [K]$ and receives a random reward $X_{I_s,s}$ drawn from P_{I_s} which is independent of the past. Recall that for $t \in \mathbb{N}_+$, the *pseudo-regret of policy* π *on bandit* B_P *at time* t is then given by

$$\bar{R}_{B_P,\pi}(t) = \max_{k \in [K]} \mathbb{E}_{\mathbb{P}_\pi}\Big[\sum_{s=1}^{t} X_{k,s} - \sum_{s=1}^{t} X_{I_s,s}\Big].$$

3 Assignment

(20pts) 1. Write a brief essay (of approximately one page) that intuitively describes what the *Kullback–Leibler divergence* is, and what its uses are. The intended audience should be a BSc student at the end of their 3rd year. You are free to refer to existing literature.

Assume that $K > 1$ and $t \geq K - 1$ for the remainder of this assignment.

For $\mu = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^K$, let B_μ refer specifically to a K-armed *normal bandit* which satisfies for $k \in [K] = \{1, \ldots, K\}$, $P_k \overset{(\mathrm{d})}{=} \mathrm{Normal}(\mu_k, 1)$. Thus B_μ is shorthand notation for $B_{(\mathrm{Normal}(\mu_1,1),\ldots,\mathrm{Normal}(\mu_K,1))}$.

2. Let $m > 0$ be any constant.

Consider a first normal bandit B_μ with mean vector $\mu = (\mu_1, \ldots, \mu_K)$, where for $k \in [K]$,

$$\mu_k = \begin{cases} m & \text{if } k = 1, \\ 0 & \text{otherwise.} \end{cases}$$

From here onwards, just as in the preliminaries, let $\mathbb{P}_\pi$ be the probability measure on the stochastic bandit model induced by the t-round interconnection of π and $P = (\mathrm{Normal}(\mu_1, 1), \ldots, \mathrm{Normal}(\mu_K, 1))$.

To avoid pathological situations, we will only consider policies π for which the least used arm under policy π on bandit B_μ, $l^\star(t) \triangleq \arg\min_{l>1} \mathbb{E}_{\mathbb{P}_\pi}[T_l(t)] = l^\star$ say, i.e., is constant. We can then namely construct a second normal bandit B_ν with *fixed* mean vector $\nu = (\nu_1, \ldots, \nu_K)$, where for $k \in [K]$,

$$\nu_k = \begin{cases} m & \text{if } k = 1, \\ 2m & \text{if } k = l^\star, \\ 0 & \text{otherwise.} \end{cases}$$

Again, like before: let $\mathbb{Q}_\pi$ be the probability measure on the stochastic bandit model induced by the t-round interconnection of π and $Q = (\mathrm{Normal}(\nu_1, 1), \ldots, \mathrm{Normal}(\nu_K, 1))$.

(30pts) (a) Prove that for any policy π for which $l^\star(t)$ is constant as a function of t,

$$\bar{R}_{B_\mu,\pi}(t) + \bar{R}_{B_\nu,\pi}(t) \geq \frac{mt}{4} \mathrm{e}^{-\mathrm{KL}(\mathbb{P}_\pi, \mathbb{Q}_\pi)}.$$

Hint 1. Can you prove that $\bar{R}_{B_\mu,\pi}(t) \geq (mt/2)\mathbb{P}_\pi\big[T_1(t) \leq t/2\big]$?

Hint 2. You will need *Bretagnolle–Huber's inequality*: Let $\mathbb{R}$ and $\mathbb{S}$ be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then

$$\mathbb{R}[A] + \mathbb{S}[\Omega \backslash A] \geq \tfrac{1}{2}\mathrm{e}^{-\mathrm{KL}(\mathbb{R},\mathbb{S})}.$$

(50pts) 3. Prove that for any policy π for which $l^\star(t)$ is constant as a function of t, there exists a fixed vector $\xi = (\xi_1, \ldots, \xi_K) \in [0, 1]^K$ such that

$$\bar{R}_{B_\xi,\pi}(t) \geq \frac{1}{16\sqrt{\mathrm{e}}}\sqrt{(K-1)t}.$$

Hint 3. The proof is by construction, which means that for each policy π you should identify a *specific* fixed vector ξ for which the bound holds. **Part a** gives you *two* candidates.

Hint 4. You could derive an upper bound on a Kullback–Leibler divergence that is a function of t, m, and K. For $k \in [K]\backslash\{1\}$, symmetry between the suboptimal arms allows you to give an upper bound on $\mathbb{E}_{\mathbb{P}_\pi}[T_k(t)]$ that is slightly sharper than the easiest upper bound $\mathbb{E}_{\mathbb{P}_\pi}[T_k(t)] \leq t$.

Hint 5. You could always maximize a lower bound over m.

4 Deliverable

Write a report in LaTeX that contains the essay as well as your proofs, and hand in a compiled PDF.

4.1 Knockout criteria

Your report **must** meet the following requirements in order to be marked:

(i) An electronic copy must be submitted in Canvas (PDF), on time.

(ii) It includes a title, author names, student numbers, and bibliography when citing.

(iii) It includes a paragraph in the Appendix, in which every student's specific contributions are explained. In particular, tell us which percentage of (a) analysis and (b) report writing was done by each student. Each student is expected to contribute significantly to both.

(iv) There is a hard page limit of **four A4 papers**, single column. This excludes the bibliography, and assignment contribution statement.
Focus on what is important: it is about quality of content, not quantity.

(v) Margins should be at least 2 cm, and font size at least 10 pt.

(vi) The overall presentation is clean / neat / orderly.

(vii) Your texts are legible, in English, and contain few spelling mistakes.

4.2 Grading

This is a *challenging* theoretical assignment. Try your best, and don't give up. I will be judging for effort and you may report difficulties or even failure. However, your aim should most certainly be a perfect proof for that perfect mark.