CS 747 (Autumn 2023)
Mid-semester Examination

Instructor: Shivaram Kalyanakrishnan

6.30 p.m. – 8.30 p.m., September 20, 2023, LA 201 and LA 202

Note. This exam has **4** questions, given on the pages following this one. Provide justifications/calculations/steps along with each answer to illustrate how you arrived at the answer. You will not receive credit for giving an answer without sufficient explanation.

Steps for submission.

1. Bring your phone in a pouch or bag, and keep it on the table you are using.
2. Before the exam begins, turn on "flight mode" on the phone, so it cannot communicate. Do not touch the phone while you are writing the exam.
3. When you are finished writing, put your pen away and stand up.
4. Remain standing while retrieving your phone, scanning your paper, turning off "flight mode", then uploading the scanned pdf to Moodle.
5. You will get 15 extra minutes after the test end time for scanning and uploading (you can do it earlier if you have finished). If you are unable to scan and upload your paper, you will be given a slot later to do so.
6. Before leaving, you must turn in your answer paper to the invigilators in the room.
7. We will only evaluate submissions for which the scanned copy matches the physical answer paper that has been turned in.

Question 1. We consider $\mathcal{I}$, the set of all n-armed bandit instances, $n \geq 2$, in which arms from the set $A = \{1, 2, \ldots, n\}$ all yield Bernoulli rewards. Each bandit instance $I \in \mathcal{I}$ is thus a vector $(p_1(I), p_2(I), \ldots, p_n(I))$, where $p_a(I)$ for $a \in A$ is the probability that arm a in instance I yields a 1-reward.

Let $\mathcal{L}$ be the set of all *algorithms*, (including both deterministic and randomised algorithms) that operate on instances in $\mathcal{I}$. When an algorithm $L \in \mathcal{L}$ interacts with an instance $I \in I$, suppose that it produces a history h that contains $s_a(h)$ 1-rewards (or successes) and $f_a(h)$ 0-rewards (or failures) for each arm $a \in A$. For example, the history $1, 0, 1, 1, 2, 0, 3, 1, 2, 0$ (read as an "arm, reward" sequence) on a 3-armed bandit instance has one success and one failure for arm 1, two failures for arm 2, and one success for arm 3.

In both the questions below, assume that h is a fixed history, with associated success- and failure-counts $s_a(h)$ and $f_a(h)$, respectively, for $a \in A$. For $I \in \mathcal{I}$, $L \in \mathcal{L}$, let $q_{I,L}(h)$ denote the probability that algorithm L acting on bandit instance I produces history h.

1a. Provide an expression for $\boxed{\max_{L \in \mathcal{L}} q_{I,L}(h)}$ in terms of I and h. [2 marks]

1b. Provide an expression for $\boxed{\max_{I \in \mathcal{I}} \max_{L \in \mathcal{L}} q_{I,L}(h)}$ in terms of h. [3 marks]

Give sufficient justification for both your answers.

Question 2. Consider the usual bandit setup studied in class: n-armed bandits, $n \geq 2$, in which each arm yields Bernoulli rewards. We know that there are algorithms that can yield sub-linear regret (in terms of the horizon) on such bandit instances, and in fact which are asymptotically optimal with respect to Lai and Robbins's lower bound (the ratio of their expected cumulative regret to the lower bound approaches 1 as the horizon goes to infinity).

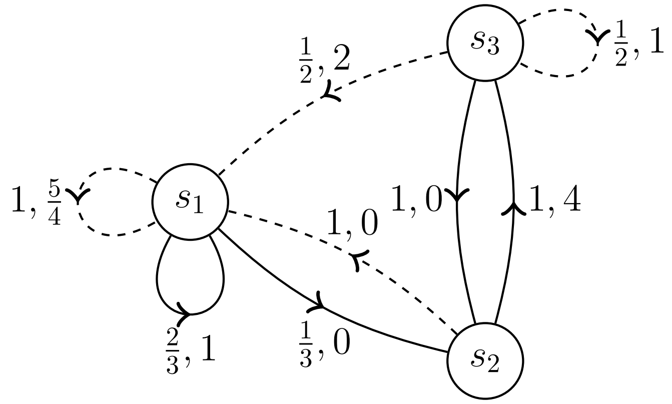
In this question, we consider a family of algorithms $\mathcal{L}(b)$, which are parameterised by an integer $b \geq 1$. Every algorithm $L \in \mathcal{L}(b)$ is constrained as follows. When L selects an arm to pull, it must pull that arm consecutively b times, before getting to decide its next choice of arm afresh. Thus, for example, if $b = 3$, the indices of the arms pulled, in sequence, could be (i) $2, 2, 2, 1, 1, 1, 3, 3, 3, 3, 3, 3, \ldots$, (ii)$1, 1, 1, 4, 4, 4, 2, 2, 2, 1, 1, 1, \ldots$, etc. Sequences that would not be possible would include, among others, (i) $2, 1, 2, 1, 3, 3, \ldots$, and (ii) $1, 1, 1, 4, 4, 4, 2, 2, 1, \ldots$. Observe that the unrestricted family of algorithms we have considered in class is $\mathcal{L}(1)$.

Consider arbitrary $b > 1$. Is there an algorithm $L^\star \in \mathcal{L}(b)$ that achieves sub-linear regret in terms of the horizon on all bandit instances? Is there an algorithm $L^{\star\star} \in \mathcal{L}(b)$ that achieves asymptotically optimal regret with respect to Lai and Robbins's lower bound on all bandit instances? If you claim "yes" for either question, describe $L^\star$ and/or $L^{\star\star}$ and prove your claim. If your answer is "no" to either question, prove why sub-linear and/or asympotically optimal regret is not possible with $b > 1$. For convenience you can assume that the horizon is a multiple of b, and also that the algorithm knows the horizon. You can quote any results stated in class without proving them. [5 marks]

Question 3. This question requires you to compute the optimal value function of an MDP (S, A, T, R) (notations as usual) when rewards are aggregated under the "finite horizon reward" assumption, with no discounting applied. For $s \in S$, $t \geq 1$, let $V^{\star}(s, t)$ denote the maximum expected sum of rewards that the agent can obtain in t steps if starting from state s, and taking actions so as to maximise this expected sum. Thus t is the "steps to go" from s while defining $V^{\star}(s, t)$.

3a. Write down a recursive formula to calculate $V^{\star}(s, t)$, along with the relevant base case. [1 mark]

3b. Consider the MDP shown below, which has a set of states $S = \{s_1,\ s_2,\ s_3\}$, and a set of actions $A = \{a_1,\ a_2\}$. In the figure, transitions of action a_1 are shown as solid lines, while those of a_2 are shown as dashed lines. Only transitions with non-zero probabilities are shown, and annotated with "probability, reward".



Fill out the table below with the values of $V^{\star}(s, t)$ for this MDP, for $s \in S$, $t \in \{1, 2, 3\}$. [4 marks]

s	$t = 3$	$t = 2$	$t = 1$
s_1			
s_2			
s_3			

Question 4. Consider MDP (S, A, T, R, γ), with notations as usual, and in which values are taken as the expectation of an infinite sum of rewards discounted by $\gamma < 1$ (the default assumption in class).

Let $V^\star : S \to \mathbb{R}$ be the optimal value function of this MDP, and let $\overline{V} : S \to \mathbb{R}$ be an ϵ-approximation of $V^\star$ for some $\epsilon \geq 0$. More precisely, $\overline{V}$ satisfies $\|\overline{V} - V^\star\|_\infty \leq \epsilon$.

Now, let $\overline{\pi} : S \to A$ be a greedy policy with respect to $\overline{V}$, in the sense that

$$\overline{\pi}(s) = \operatorname*{argmax}_{a \in A} \sum_{s' \in S} T(s, a, s')\{R(s, a, s') + \gamma \overline{V}(s')\}$$

for $s \in S$, with ties broken arbitrarily in the "argmax". We already know that if $\epsilon = 0$, then $\overline{V} = V^\star$ and $\overline{\pi}$ is an optimal policy. This question investigates by how much $\overline{\pi}$ can possibly deviate from optimality as a function of ϵ.

As usual, let $V^{\overline{\pi}} : S \to \mathbb{R}$ denote the value function of $\overline{\pi}$. Take a moment to understand the distinction between $\overline{V}$ and $V^{\overline{\pi}}$. In general, these functions can be different.

Your task is to show that there exist positive numbers C, m, n such that

$$\|V^{\overline{\pi}} - V^\star\|_\infty \leq C\epsilon^m \left(\frac{1}{1 - \gamma}\right)^n.$$

Provide a step-by-step proof of your claim, with sufficient explanation. You might find the Bellman operator for $\overline{\pi}$ (that is, $B^{\overline{\pi}}$) and/or the Bellman optimality operator (that is, $B^\star$) useful for this exercise, although you do not necessarily have to use either of them. You will receive full credit for any solution that is correct. [5 marks]

Solutions

1a. Suppose $h = a^0, r^0, a^1, r^1, \dots, a^{t-1}, r^{t-1}$. Then

$$q_{I,L}(h) = \prod_{i=0}^{t-1} \mathbb{P}\{L \text{ pulls arm } a^i | a^0, r^0, a^1, r^1, \dots, a^{i-1}, r^{i-1}\} \mathbb{P}\{I \text{ gives reward } r^i \text{ for arm } a^i\}.$$

L can be chosen such that it precisely pulls a^i (given the preceding history that is a prefix of h) for each $0 \leq i \leq t-1$, and puts no probability on any other arm. Hence

$$\begin{aligned}\max_{L \in \mathcal{L}} q_{I,L}(h) &= \prod_{i=0}^{t-1} \mathbb{P}\{I \text{ gives reward } r^i \text{ for arm } a^i\} \\ &= \prod_{i=0}^{t-1} (p_{a^i}(I))^{\mathbf{1}[r^i=1]} (1 - p_{a^i}(I))^{\mathbf{1}[r^i=0]} \\ &= \prod_{a \in A} (p_a(I))^{s_a(h)} (1 - p_a(I))^{f_a(h)},\end{aligned}$$

with the convention that $x^0 = 1$ for $x \in [0, 1]$.

1b. Let $g(a) = p_a(I))^{s_a(h)}(1 - p_a(I))^{f_a(h)}$ denote the contributing factor for arm $a \in A$ in the product above. Maximising $\max_{L \in \mathcal{L}} q_{I,L}(h)$ over I amounts to maximising $g(a)$ for each arm a.

For arm $a \in A$, if $s_a(h) = 0$ or $f_a(h) = 0$, then $g(a)$ can be made 1 by taking I such that $p_a(I) = 0$ or $p_a(I) = 1$, respectively (any choice is okay for $p_a(I)$ if $s_a(h)$ and $f_a(h)$ are both 0). If $s_a(h)$ and $f_a(h)$ are both positive, then clearly $g(a)$ is not maximised at either $p_a(I) = 0$ or $p_a(I) = 1$; we will find the maximiser of $g(a)$ by finding $p_a(I) \in (0, 1)$ that maximises $\ln(g(a)) = s_a(h) \ln p_a(I) + f_a(h) \ln(1 - p_a(I))$. Consider the function

$$\alpha(x) = s_a(h) \ln x + f_a(h) \ln(1 - x)$$

for $x \in (0, 1)$. We have

$$\alpha'(x) = \frac{s_a(h)}{x} - \frac{f_a(h)}{1 - x}, \text{ and } \alpha''(x) = -\frac{s_a(h)}{x^2} - \frac{f_a(h)}{(1 - x)^2}.$$

The unique solution to $\alpha'(x) = 0$ is $x^\star = \frac{s_a(h)}{s_a(h)+f_a(h)}$; also $f''(x^\star) < 0$. We conclude that $x^\star = \frac{s_a(h)}{s_a(h)+f_a(h)}$ is the unique maximiser of $\alpha(\cdot)$. Plugging in this observation, we have

$$\max_{I \in \mathcal{I}} \max_{L \in \mathcal{L}} p_{I,L}(h) = \prod_{a \in A} \left(\frac{s_a(h)}{s_a(h) + f_a(h)} \right)^{s_a(h)} \left(\frac{f_a(h)}{s_a(h) + f_a(h)} \right)^{f_a(h)}.$$

2. We already know that there exist asymptotically optimal algorithms for $b = 1$: for example, KL-UCB and Thompson Sampling. Let us take any of these algorithms as $L^{\star\star}(1)$. For $b > 1$, we construct $L^{\star\star}(b)$ by using $L^{\star\star}(1)$ as a blackbox, such that $L^{\star\star}(b)$ is also asymptotically optimal.

It is instructive to first visualise the construction of $L^{\star\star}(b)$ through a thought experiment. Suppose that siblings X and Y live at home. The sister X has taken CS 747 and knows some optimal algorithms $L^{\star\star}(1)$. Although the brother Y does not know much about bandits, he can still be counted upon to follow simple rules. The bandit interaction is as follows. Whenever an arm is to be pulled, a telephone call must be placed from the siblings' *home* to an *office*. On the call, the caller specifies the arm a to pull; the office responds with b corresponding rewards drawn i.i.d. from the reward distribution of a. As you can see, the home is the agent, and the office is the environment. What is the role played each sibling within their home? X is in charge of implementing $L^{\star\star}(1)$, and Y is in charge of placing phone calls to the office. When X decides to pull a particular arm a, she conveys this to Y, who supplies her the reward. In turn, Y places calls and gets rewards from the office.

The very first time X places a request for any arm a, Y gets b i.i.d. rewards for a by calling the office, and writes these down in his notebook. However, Y still provides X only the first reward from this sequence (which is sufficient for X make her next choice), while retaining the remaining $b - 1$ in his notebook. When X makes another request for the same arm a (possibly much later), Y consults his notebook to pass on the second reward that was received for a. At any point of time, Y's notebook has anything between 0 and $b - 1$ unused rewards for each arm. If for a particular arm that X requests, there are no rewards left in the notebook, Y places a fresh phone call the the office and gets b new rewards, from which one is again conveyed back to X.

Observe that this arrangement enables X to implement any $L^{\star\star}(1)$ algorithm. At any stage, Y's notebook has at most $n(b - 1)$ additional pulls that have been sought but not yet passed on to X. As the number of pulls increases, this "buffer" in Y's notebook vanishes as a fraction of the number of pulls. Since $L^{\star\star}(1)$ is asymptotically optimal, so is $L^{\star\star}(b)$, implemented by the "home" agent. Below is a formal description of $L^{\star\star}(b)$.

```
$L^{\star\star}(b)$

For arm $a \in A$:
    $rewardQueue_a = [\,]$.
$h^0 = \emptyset$.
For $t = 0, 1, \ldots$:
    $a^t \sim L^{\star\star}(1)(h^t)$.
    If $rewardQueue_{a^t}$ is empty:
        Repeat $b$ times:
            Pull arm $a^t$; let r be the reward.
            push($r$, $rewardQueue_a$).
    $r^t$ = pop($rewardQueue_a$).
    $h^{t+1} \leftarrow h^t$.append($a^t, r^t$).
```

Let $LB_T(I)$ denote Lai and Robbins's lower bound on the expected cumulative regret on instance I for horizon T; recall that it grows as $\Omega(\log(T))$. With our usual notation, we observe

$$\frac{R_T(L^{\star\star}(b), I)}{LB_T(I)} \leq \frac{R_T(L^{\star\star}(1), I) + (b-1)\sum_{a \in A}(p^\star - p_a)}{LB_T(I)};$$

$$\lim_{T \to \infty} \frac{R_T(L^{\star\star}(b), I)}{LB_T(I)} \leq \lim_{T \to \infty} \frac{R_T(L^{\star\star}(1), I)}{LB_T(I)} + \lim_{T \to \infty} \frac{(b-1)\sum_{a \in A}(p^\star - p_a)}{LB_T(I)} = 1 + 0 = 1.$$

3a. The general recursion, for $t \geq 2$, is, for $s \in S$:

$$V^\star(s, t) = \max_{a \in A} \sum_{s' \in S} T(s, a, s')\{R(s, a, s') + V^\star(s', t-1)\}.$$

One may use the same recursion for $t = 1$ by defining $V^\star(s, 0) = 0$ for $s \in S$, or alternatively by treating $t = 1$, as base case: for $s \in S$,

$$V^\star(s, 1) = \max_{a \in A} \sum_{s' \in S} T(s, a, s')R(s, a, s').$$

3b.

$$\begin{aligned}
V^\star(s_1, 1) &= \max\left\{\frac{2}{3}(1) + \frac{1}{3}(0), 1(\frac{5}{4})\right\} = \frac{5}{4}.\\
V^\star(s_2, 1) &= \max\{1(4), 1(0)\} = 4.\\
V^\star(s_3, 1) &= \max\left\{1(0), \frac{1}{2}(1) + \frac{1}{2}(2)\right\} = \frac{3}{2}.\\
V^\star(s_1, 2) &= \max\left\{\frac{2}{3}(1 + \frac{5}{4}) + \frac{1}{3}(0 + 4), 1(\frac{5}{4} + \frac{5}{4})\right\} = \frac{17}{6}.\\
V^\star(s_2, 2) &= \max\left\{1(4 + \frac{3}{2}), 1(0 + \frac{5}{2})\right\} = \frac{11}{2}.\\
V^\star(s_3, 2) &= \max\left\{1(4), \frac{1}{2}(1 + \frac{3}{2}) + \frac{1}{2}(2 + \frac{5}{4})\right\} = 4.\\
V^\star(s_1, 3) &= \max\left\{\frac{2}{3}(1 + \frac{17}{6}) + \frac{1}{3}(0 + \frac{11}{2}), 1(\frac{5}{4} + \frac{17}{6})\right\} = \frac{79}{18}.\\
V^\star(s_2, 3) &= \max\left\{1(4 + 4), 1(0 + \frac{17}{6})\right\} = 8.\\
V^\star(s_3, 3) &= \max\left\{1(\frac{11}{2}), \frac{1}{2}(1 + 4) + \frac{1}{2}(2 + \frac{17}{6})\right\} = \frac{11}{2}.
\end{aligned}$$

Here is the consolidated table.

s	$t = 3$	$t = 2$	$t = 1$
s_1	79/18	17/6	5/4
s_2	8	11/2	4
s_3	11/2	4	3/2

4. There are multiple ways to accomplish this proof. First we present the key result. Then we present two possible ways to proceed from the key result to the final claim.

Key result. Our core step is to lower-bound $B^{\overline{\pi}}(V^\star)(s) - V^\star(s)$ for $s \in S$. First, using the fact that $V^\star(s') \geq \overline{V}(s') - \epsilon$ for $s' \in S$, we observe below that

$$\begin{aligned} B^{\overline{\pi}}(V^\star)(s) &= \sum_{s' \in S} T(s, \overline{\pi}(s), s')\{R(s, \overline{\pi}(s), s') + \gamma V^\star(s')\} \\ &\geq \sum_{s' \in S} T(s, \overline{\pi}(s), s')\{R(s, \overline{\pi}(s), s') + \gamma(\overline{V}(s') - \epsilon)\} \\ &= \sum_{s' \in S} T(s, \overline{\pi}(s), s')\{R(s, \overline{\pi}(s), s') + \gamma \overline{V}(s')\} - \gamma\epsilon. \end{aligned} \tag{1}$$

Let $\pi^\star$ be any optimal policy for (S, A, T, R, γ). Since $\overline{\pi}$ is greedy with respect to $\overline{V}$, we have

$$\sum_{s' \in S} T(s, \overline{\pi}(s), s')\{R(s, \overline{\pi}(s), s') + \gamma \overline{V}(s')\} \geq \sum_{s' \in S} T(s, \pi^\star(s), s')\{R(s, \pi^\star(s), s') + \gamma \overline{V}(s')\}. \tag{2}$$

Now we use the fact that $\overline{V}(s') \geq V^\star(s') - \epsilon$ for $s' \in S$, and also that $V^{\pi^\star} = V^\star$, to get

$$\begin{aligned} \sum_{s' \in S} T(s, \pi^\star(s), s')\{R(s, \pi^\star(s), s') + \gamma \overline{V}(s')\} &\geq \sum_{s' \in S} T(s, \pi^\star(s), s')\{R(s, \pi^\star(s), s') + \gamma(V^\star(s') - \epsilon)\} \\ &= V^\star(s) - \gamma\epsilon. \end{aligned} \tag{3}$$

Combining (1), (2), and (3) yields

$$B^{\overline{\pi}}(V^\star)(s) \geq V^\star(s) - 2\gamma\epsilon. \tag{4}$$

(4) is the key result in this proof. There are at least two ways to proceed from here in order to prove our final result.

Method 1. Suppose for $X : S \to \mathbb{R}$ and $Y : S \to \mathbb{R}$ that for $s \in S$, $X(s) \geq Y(s) + c$ for some $c \in \mathbb{R}$. Then we observe that for every $\pi : S \to A$

$$B^\pi(X)(s) - B^\pi(Y)(s) = \gamma \sum_{s' \in S} T(s, \pi(s), s')\{X(s') - Y(s'))\} \geq \gamma c. \tag{5}$$

Indeed we have proven (5) in class for $c = 0$. Repeatedly applying (5) using $\overline{\pi}$ as our policy, to (4), we get the sequence

$$\begin{aligned} B^{\overline{\pi}}(V^\star)(s) &\geq V^\star(s) - 2\gamma\epsilon, \\ (B^{\overline{\pi}})^2(V^\star)(s) &\geq B^{\overline{\pi}}(V^\star)(s) - 2\gamma^2\epsilon, \\ (B^{\overline{\pi}})^3(V^\star)(s) &\geq (B^{\overline{\pi}}(V^\star))^2(s) - 2\gamma^3\epsilon, \\ &\vdots \end{aligned}$$

for $s \in S$, which implies

$$\lim_{l \to \infty} (B^{\overline{\pi}})^l (V^\star)(s) = V^{\overline{\pi}}(s) \geq V^\star(s) - 2\gamma\epsilon(1 + \gamma + \gamma^2 + \dots) = V^\star(s) - \frac{2\gamma\epsilon}{1-\gamma}. \tag{6}$$

On the other hand, since $V^\star$ is the optimal value function, it follows for that for $s \in S$,

$$V^{\overline{\pi}}(s) \leq V^\star(s). \tag{7}$$

(6) and (7) together prove our claim (taking $C = 2, m = 1, n = 1$); in particular, that

$$\|V^{\overline{\pi}} - V^\star\|_\infty \leq \frac{2\epsilon\gamma}{1-\gamma} < \frac{2\epsilon}{1-\gamma}.$$

Method 2. Since $V^\star$ is the optimal value function, we observe from the policy improvement theorem that for every policy $\pi : S \to A$, $V^\star \succeq B^\pi(V^\star)$. We apply this observation to the policy $\overline{\pi}$: for $s \in S$,

$$B^{\overline{\pi}}(V^\star)(s) \leq V^\star(s). \tag{8}$$

In turn, (4) and (8) together imply

$$\|B^{\overline{\pi}}(V^\star) - V^\star\|_\infty \leq 2\gamma\epsilon. \tag{9}$$

Since $B^{\overline{\pi}}$ is a contraction mapping in the $(\mathbb{R}^n, \|\|_\infty)$ Banach space with contraction factor γ, we have

$$\|(B^{\overline{\pi}})^2(V^\star) - B^{\overline{\pi}}(V^\star)\|_\infty \leq \gamma\|B^{\overline{\pi}}(V^\star) - V^\star\|_\infty \leq 2\gamma^2\epsilon,$$

and by proceeding similarly,

$$\|(B^{\overline{\pi}})^l(V^\star) - (B^{\overline{\pi}})^{l-1}(V^\star)\|_\infty \leq 2\gamma\|B^{\overline{\pi}}(V^\star) - V^\star\|_\infty \leq 2\gamma^l\epsilon$$

for integers $l \geq 1$. We conclude that for each integer $l \geq 1$,

$$\begin{aligned}
\|(B^{\overline{\pi}})^l(V^\star) - V^\star\|_\infty &= \|\sum_{i=1}^{l} (B^{\overline{\pi}})^i(V^\star) - (B^{\overline{\pi}})^{i-1}V^\star)\|_\infty \\
&\leq \sum_{i=1}^{l} \|(B^{\overline{\pi}})^i(V^\star) - (B^{\overline{\pi}})^{i-1}V^\star)\|_\infty \\
&\leq \sum_{i=1}^{l} 2\gamma^i\epsilon.
\end{aligned}$$

Taking the limit as $l \to \infty$ on both sides, we get

$$\lim_{l \to \infty} \|(B^{\overline{\pi}})^l(V^\star) - V^\star)\|_\infty = \|V^{\overline{\pi}} - V^\star\|_\infty \leq \frac{2\gamma\epsilon}{1-\gamma},$$

which is the same conclusion as from Method 1.