CS 747 (Autumn 2019): Mid-semester Examination

Instructor: Shivaram Kalyanakrishnan

8.30 a.m. – 10.30 a.m., September 19, 2019, LA 201/202

Total marks: 15

Note. Provide justifications and/or calculations along with each answer to illustrate how you arrived at the answer.

Question 1. An n-armed bandit instance I, where $n \geq 2$, has $p_1, p_2, \ldots, p_n \in (0, 1)$ as the mean rewards of its arms. Each arm implements a Bernoulli distribution (that is, returns 0-1 rewards). Let $r^0, r^1, \ldots, r^{T-1}$ be the rewards obtained in the first T pulls, where $T \geq 1$. Let

$$x^T = r^0 + r^1 + \cdots + r^{T-1}.$$

1a. Consider an algorithm that picks an arm uniformly at random at each round and pulls it (you may think of it as ϵ-greedy sampling with $\epsilon = 1$). If this algorithm is executed, what is the *variance* of x^T?—in other words $\mathbb{E}[(x^T)^2] - (\mathbb{E}[x^T])^2$? [2 marks]

Let

$$y^T = (1 - r^0) + (1 - r^1) + \cdots + (1 - r^{T-1}) = T - x^T.$$

While x^T is the total number of 1-rewards in the first T pulls, y^T is the total number of 0-rewards in the first T pulls. Let

$$z^T = \max(x^T, y^T);$$

that is, z^T counts the total number of the more-frequent reward value in the first T pulls. Suppose we are interested in maximising $\mathbb{E}[z^T]$.

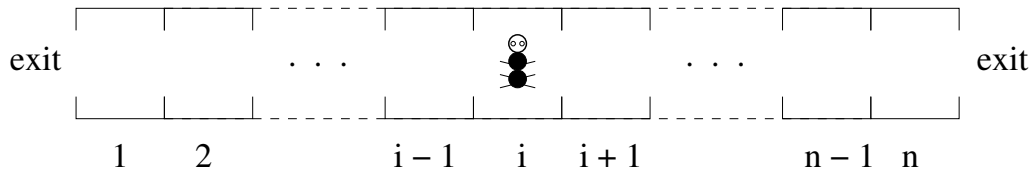
As an analogy, recall the coin-tossing game played in the first class. When we actually played it, the instructor promised you as many Rupees as the number of heads you could gather in T tosses. Now we consider what to do if you were instead promised as many Rupees as either the number of heads or the number of tails, whichever you demanded upon completion (and you would naturally choose to maximise your profit).

1b. Suppose one has knowledge of I: that is, one knows $p_1, p_2, \ldots, p_n$. What algorithm $L_\star$ must one apply in order to maximise $\mathbb{E}[z^T]$? [1 mark]

1c. Let $Z_\star(T)$ denote the maximum value of $\mathbb{E}[z^T]$ that can be attained on I—that is, by playing $L_\star$. Let $Z_L(T)$ denote the value of $\mathbb{E}[z^T]$ achieved by some algorithm L.

Describe an algorithm L that achieves $Z_\star(T) - Z_L(T) \leq C(I) \ln(T)$ for sufficiently large T and for all bandit instances I, where $C(I)$ is a constant depending only on I. You might find it convenient to view $(Z_\star(T) - Z_L(T))$ as the "Z-regret" of L. Provide a proof to back up your claim. You can use results derived in class. [3 marks]

Question 2. An ant is inside a tunnel with n chambers, $n \geq 2$. As shown in the figure below, the chambers are numbered $1, 2, \ldots, n$. The ant seeks your help in exiting the tunnel quickly.



The ant can either take action L (going left) or action R (going right). Due to the winds blowing inside the tunnel, neither action guarantees success; rather, there is an associated probability of making progress.

- From chamber $i \in \{1, 2, 3, \ldots, n\}$, action L takes the ant to chamber $i - 1$ with probability p_L, and retains the ant in chamber i with probability $1 - p_L$, where $p_L \in (0, 1)$.
- From chamber $i \in \{1, 2, 3, \ldots, n\}$, action R takes the ant to chamber $i + 1$ with probability p_R, and retains the ant in chamber i with probability $1 - p_R$, where $p_R \in (0, 1)$.

Chambers 0 and $n+1$, which are not shown in the figure, are outside the tunnel. Upon reaching them the ant is free!

Let $\pi^\star : \{1, 2, \ldots, n\} \to \{L, R\}$ be a policy that minimises the expected number of steps taken by the ant to exit the tunnel, whatever be the starting chamber i.

2a. Show that for $i \in \{1, 2, \ldots, n-1\}$, if $\pi^\star(i) = R$, then $\pi^\star(i+1) = R$. In other words, $\pi^\star$ must take L for chambers $1, 2, \ldots, m$, and R for chambers $m+1, m+2, \ldots, n$, for some $m \in \{0, 1, \ldots, n\}$. [2 marks]

2b. Using the result from 2a, compute $\pi^\star$: that is, express it in terms of n, p_L, and p_R. [2 marks]

Question 3. This question asks you to establish that an approximation of the optimal value function of an MDP induces an approximately-optimal policy. You can use results derived in class as a part of your answer.

Consider an MDP (S, A, T, R, γ). Let Π be the set of all policies for this MDP.

Suppose that the optimal value function $V^\star$ is approximated by $V : S \to \mathbb{R}$. Concretely, we have $\|V - V^\star\|_\infty \leq \epsilon$ for some $\epsilon > 0$. Recall that for $F : S \to \mathbb{R}$, $\|F\|_\infty = \max_{s \in S} |F(s)|$.

A policy $\pi \in \Pi$ is said to be *greedy* with respect to a function $X : S \to \mathbb{R}$ if for all $\pi' \in \Pi$, $B^\pi(X) \succeq B^{\pi'}(X)$. Let $\pi \in \Pi$ be greedy with respect to V.

Show that

$$\|V^\pi - V^\star\|_\infty \leq \frac{2\gamma\epsilon}{1-\gamma}. \text{ [5 marks]}$$

Solutions

1a. Sampling uniformly at random results in a probability $\bar{p} = \frac{p_1+p_2+\cdots+p_n}{n}$ of r^t being 1 and a probability $1-\bar{p}$ of r^t being 0 for $t \in \{0, 1, \ldots, T-1\}$. Hence,

$$\begin{aligned}
\mathbb{E}[x^T] &= \sum_{i=0}^{T} \mathbb{P}\{x^T = i\} i \\
&= \sum_{i=0}^{T} \binom{T}{i} \bar{p}^i (1-\bar{p})^{T-i} i \\
&= \sum_{i=1}^{T} \binom{T}{i} \bar{p}^i (1-\bar{p})^{T-i} i \\
&= \bar{p}T \sum_{i=1}^{T} \binom{T-1}{i-1} \bar{p}^{i-1} (1-\bar{p})^{T-i} \\
&= \bar{p}T \sum_{j=0}^{T} \binom{T-1}{j} \bar{p}^{j} (1-\bar{p})^{T-1-j} \\
&= \bar{p}T.
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[(x^T)^2] &= \sum_{i=0}^{T} \mathbb{P}\{x^T = i\} i^2 \\
&= \sum_{i=0}^{T} \binom{T}{i} \bar{p}^i (1-\bar{p})^{T-i} i^2 \\
&= \sum_{i=1}^{T} \binom{T}{i} \bar{p}^i (1-\bar{p})^{T-i} i^2 \\
&= \sum_{i=1}^{T} \binom{T}{i} \bar{p}^i (1-\bar{p})^{T-i} i(i-1) + \sum_{i=1}^{T} \binom{T}{i} \bar{p}^i (1-\bar{p})^{T-i} i \\
&= \sum_{i=2}^{T} \binom{T}{i} \bar{p}^i (1-\bar{p})^{T-i} i(i-1) + \sum_{i=1}^{T} \binom{T}{i} \bar{p}^i (1-\bar{p})^{T-i} i \\
&= \sum_{i=2}^{T} T(T-1) \binom{T-2}{i-2} \bar{p}^{i-2} (1-\bar{p})^{T-i} + \mathbb{E}[x^T] \\
&= \sum_{i=2}^{T} T(T-1)\bar{p}^2 \binom{T-2}{i-2} \bar{p}^{i-2} (1-\bar{p})^{T-i} + \mathbb{E}[x^T] \\
&= \bar{p}^2 T^2 - \bar{p}^2 T + \bar{p}T.
\end{aligned}$$

Hence, $Var[x^T] = \mathbb{E}[(x^T)^2] - (\mathbb{E}[x^T])^2 = \bar{p}(1-\bar{p})T$.

The easier way to solve the problem is to use the fact that if X and Y are independent, then $Var[X+Y] = Var[X] + Var[Y]$. Note that indeed r^0 and r^1 and $\ldots$ and r^{T-1} are independent and identically distributed; each having variance $\bar{p} - \bar{p}^2$. The answer immediately follows.

1b. One must always play an arm a that maximises $\max(p_a, 1 - p_a)$.

1c. Given an n-armed instance I, we simulate a $2n$-armed instance I'. Each arm a in I is also present in I'; also included in I' is a corresponding arm a' with mean $1 - p_a$. Although we do not know the means of the arms, we can still implement pulls legitimately on I'. If an arm $a \in A$ is pulled, it is pulled in I and the corresponding reward r noted. If an arm $a' \in A'$ is pulled, the corresponding arm $a \in A$ is pulled on I; if r is the reward received on I, then $1 - r$ is recorded as the reward on I'. It is easy to see that the expected reward from arm a is p_a and that of a' is $1 - p_a$.

We run UCB on I'. On I', we know that UCB will play each non-optimal arm at most $C_1(I') \ln(T)$ times up to horizon T. In turn, this means that the number of pulls of non-optimal arms in I (with respect to Z-regret) is at most $C_2(I') \ln(T)$, and the Z-regret itself at most $C_3(I') \ln(T)$. Since I' is fully defined by I, this regret upper bound is of the form $C(I) \ln(T)$.

2a. If $\pi^\star(i)$ and $\pi^\star(i+1)$ are both R, then the only states that can ever be visited starting from i or $i+1$ are i and $i+1$. Clearly the ant cannot exit the tunnel if this is the case. Hence, if $\pi^\star(i) = R$, we must have $\pi^\star(i+1) = R$.

2b. We have established in 2a that either "always take L" or "always take R" will be an optimal control strategy starting from each chamber. For $i \in \{1, 2, \ldots, n\}$, let $X_L(i)$ denote the expected number of steps, if starting from chamber i, to exit, by always taking L. Similarly, let $X_R(i)$ denote the expected number of steps, if starting from chamber i, to exit, by always taking R.

We have $X_L(1) = p_L(1) + (1 - p_L)(1 + X_L(1))$, which gives $X_L(1) = \frac{1}{p_L}$. Using the recurrence for $i \in \{2, 3, \ldots, n\}$ that $X_L(i) = p_L(1 + X_L(i-1)) + (1 - p_L)(1 + X_L(i))$, we get $X_L(i) = \frac{i}{p_L}$. A similar working shows $X_R(i) = \frac{n-i+1}{p_R}$.

Observe that $X_L(i) \leq X_R(i) \iff i \leq \frac{p_L}{p_L + p_R}(n+1)$. Thus, we get (giving ties to L):

$$\pi^\star(i) = \begin{cases} L, \text{ if } i \leq \frac{p_L}{p_L + p_R}(n+1) \\ R, \text{otherwise.} \end{cases}$$

3. First we bound $\|B^\pi(V^\star) - V^\star\|_\infty$. For $s \in S$, we have

$$\begin{aligned}
B^\pi(V^\star)(s) &= Q^\star(s, \pi(s)) \\
&= \sum_{s' \in S} T(s, \pi(s), s')\{R(s, \pi(s), s') + \gamma V^\star(s')\} \\
&\geq \sum_{s' \in S} T(s, \pi(s), s')\{R(s, \pi(s), s') + \gamma(V(s') - \epsilon)\} \\
&= \sum_{s' \in S} T(s, \pi(s), s')\{R(s, \pi(s), s') + \gamma V(s')\} - \gamma\epsilon \\
&= B^\pi(V)(s) - \gamma\epsilon \\
&\geq B^\star(V)(s) - \gamma\epsilon \\
&= \sum_{s' \in S} T(s, \pi^\star(s), s')\{R(s, \pi^\star(s), s') + \gamma V(s')\} - \gamma\epsilon \\
&\geq \sum_{s' \in S} T(s, \pi^\star(s), s')\{R(s, \pi^\star(s), s') + \gamma(V^\star(s') - \epsilon)\} - \gamma\epsilon \\
&= \sum_{s' \in S} T(s, \pi^\star(s), s')\{R(s, \pi^\star(s), s') + \gamma V^\star((s')\} - 2\gamma\epsilon \\
&= V^\star(s) - 2\gamma\epsilon.
\end{aligned}$$

It is also a fact that $B^\pi(V^\star)(s) \leq V^\star(s)$, since $V^\star$ is the optimal value function. In short,

$$0 \leq V^\star(s) - B^\pi(V^\star)(s) \leq 2\gamma\epsilon,$$

which means $\|B^\pi(V^\star) - V^\star\|_\infty \leq 2\gamma\epsilon$. By the repeated application of Banach's Fixed Point Theorem, we get for all $l \geq 1$ that $\|(B^\pi)^l(V^\star) - (B^\pi)^{l-1}(V^\star)\|_\infty \leq 2\gamma^l\epsilon$, which implies

$$\|(B^\pi)^l(V^\star) - V^\star\|_\infty \leq \sum_{i=0}^{l-1} \|(B^\pi)^{i+1}(V^\star) - (B^\pi)^i(V^\star)\|_\infty \leq 2\epsilon(\gamma + \gamma^2 + \cdots + \gamma^l).$$

As we take $l \to \infty$, we get

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