# CS 747 (Autumn 2015): Mid-semester Examination

Instructor: Shivaram Kalyanakrishnan

8.30 a.m. – 10.30 a.m., September 12, 2015, LA 301

Total marks: 15

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

**Question 1.** Consider a 2-armed bandit instance whose arms 1 and 2 yield Bernoulli rewards $p_1$ and $p_2$, respectively, with $1 \geq p_1 > p_2 \geq 0$ (thus arm 1 is the optimal arm). For $t = 1, 2, \ldots$, let $r_t$ denote the 0-or-1 reward obtained by an algorithm on its $t$-th pull. Answer the following questions, which pertain to different sampling algorithms.

1a. Let $A_1$ be an algorithm that at every step, pulls an arm that is selected uniformly at random. What is the probability of getting a run in which the first five rewards obtained by $A_1$ are all 0? [1 mark]

1b. Let $A_2$ be an algorithm that (1) samples each arm once, then (2) selects the arm with the higher empirical mean after these two initial pulls (breaking ties uniformly at random), and (3) samples this selected arm at every time step thereafter. What is $\mathbb{E}[r_{10}]$ for $A_2$? [2 marks]

1c. Algorithm $A_3$ is an $\epsilon$-greedy algorithm, $0 < \epsilon \leq 1$, that (1) samples each arm once, and (2) at every time step thereafter: with probability $1 - \epsilon$, samples the arm with the higher empirical mean (breaking ties uniformly at random), and with probability $\epsilon$, samples an arm that is selected uniformly at random. For $A_3$, what is $\lim_{t \to \infty} \mathbb{E}[r_t]$? [1 mark]

1d. For the UCB algorithm that we analysed in class (denoted UCB1 by Auer *et al.* (2002)), what is $\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r_t]$? [1 mark]

**Question 2.** $M = (S, A, R, T, \gamma)$ is an MDP for which $\pi^\star$ is an optimal policy. Answer the following questions about related MDPs.

2a. Let $M_1$ be the MDP $(S, A, R, T, \gamma')$, where $\gamma' = \frac{\gamma}{2}$. Is $\pi^\star$ necessarily an optimal policy for $M_1$? [1 mark]

2b. Let $M_2$ be the MDP $(S, A, R', T, \gamma)$, where $R'$ is a linear transformation of $R$. That is, for some $\alpha, \beta \in \mathbb{R}$, and for every $s, s' \in S, a \in A$,

$$R'(s, a, s') = \alpha R(s, a, s') + \beta.$$

Is $\pi^\star$ necessarily an optimal policy for $M_2$? [1 mark]

**Question 3.** Consider a MDP $M = (S, A, R, T, \gamma)$, with a set of states $S = \{s_1, s_2\}$; a set of actions $A = \{a_1, a_2\}$; a transition function $T$ and a reward function $R$ as specified in Table 1; and a discount factor $\gamma = \frac{2}{3}$. A state transition diagram corresponding to $M$ is shown in Figure 1.

Table 1: Transition probabilities and rewards (discount factor = 2/3).

| Transition probabilities | Rewards |
|---|---|
| $T(s_1, a_1, s_1) = 1$ | $R(s_1, a_1, s_1) = 0$ |
| $T(s_1, a_1, s_2) = 0$ | $R(s_1, a_1, s_2) = 0$ |
| $T(s_1, a_2, s_1) = 1/2$ | $R(s_1, a_2, s_1) = -1$ |
| $T(s_1, a_2, s_2) = 1/2$ | $R(s_1, a_2, s_2) = 2$ |
| $T(s_2, a_1, s_1) = 1$ | $R(s_2, a_1, s_1) = 1$ |
| $T(s_2, a_1, s_2) = 0$ | $R(s_2, a_1, s_2) = 0$ |
| $T(s_2, a_2, s_1) = 1/4$ | $R(s_2, a_2, s_1) = 0$ |
| $T(s_2, a_2, s_2) = 3/4$ | $R(s_2, a_2, s_2) = 1$ |

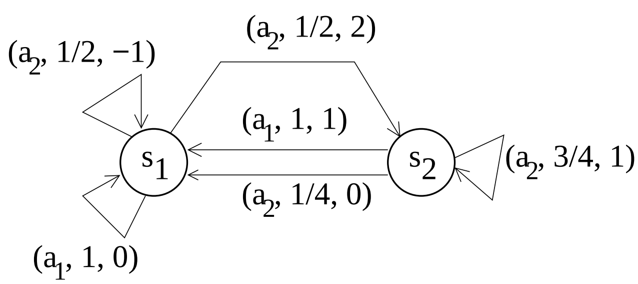


Figure 1: State transition diagram for $M$. Each transition is annotated with (action, transition probability, reward). Transitions with zero probabilities are not shown.

For $i, j \in \{1, 2\}$, let $\pi^{ij}$ denote the deterministic policy that takes action $i$ from state $s_1$ and action $j$ from state $s_2$.

3a. What is the optimal value function for $M$? [4 marks]

3b. Among $\pi^{11}$, $\pi^{12}$, $\pi^{21}$, and $\pi^{22}$, which is an optimal policy? [1 mark]

Consider an agent that at time $t = 1$, is in state $s_1$. At every time step, the agent follows policy $\pi^{22}$.

3c. What is the probability that the agent is in state $s_1$ at time $t = 3$; that is, after it has taken two actions? [1 mark]

3d. What is the probability that the agent is in state $s_1$ at time $t = T$ as $T \to \infty$; that is, after it has taken infinitely many actions? [2 marks]

# Solutions

**1a.** The probability of getting a 0-reward on any given pull is $\frac{1}{2}(1-p_1)+\frac{1}{2}(1-p_2)$; the probability of getting a 0-reward on each of the first five pulls is therefore

$$\left(1 - \frac{p_1 + p_2}{2}\right)^5.$$

**1b.** The arm that gets selected after the first two pulls will be pulled at $t = 10$. The probability that arm 1 gets selected is

$$q_1 = p_1(1-p_2) + \frac{1}{2}p_1p_2 + \frac{1}{2}(1-p_1)(1-p_2).$$

The expected reward at $t = 10$ is $p_1q_1 + p_2(1-q_1)$, which simplifies to:

$$\frac{(p_1-p_2)^2 + p_1 + p_2}{2}.$$

**1c.** As $t \to \infty$, each arm is explored an infinite number of times, and so $\hat{p}_1^t \to p_1$ and $\hat{p}_2^t \to p_2$. Therefore, in the limit, arm 1 is picked for exploiting (with probability $1-\epsilon$), and the arms are picked with equal probability ($\frac{\epsilon}{2}$) while exploring. We get:

$$\lim_{t\to\infty} \mathbb{E}[r_t] = p_1\left(1-\frac{\epsilon}{2}\right) + p_2\frac{\epsilon}{2}.$$

**1d.** Since the expected cumulative *regret* of UCB is $O(\log(T))$ for a horizon of $T$, the expected cumulative *reward* is $p_1T - O(\log(T))$. Thus,

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[r_t] = p_1.$$

**2a.** No; $\pi^\star$ need not be an optimal policy for $M_1$. Consider Figure 2, which shows a state transition diagram for $M$ and $M_1$. $M$ uses a discount factor of $\gamma = \frac{1}{2}$, while $M_1$ uses a discount factor of $\gamma' = \frac{1}{4}$. In both MDPs, it is easy to see that $a_2$ is the optimal action from $s_2$, and $a_2$ is the optimal action from $s_3$. However, the optimal action from $s_1$ is different for $M$ and $M_1$. For $M$: $Q^\star(s_1,a_1) = 2$, and $Q^\star(s_1,a_2) = 1$, and so $a_1$ is the optimal action from $s_1$. For $M_1$; $Q^\star(s_1,a_1) = \frac{2}{3}$, and $Q^\star(s_1,a_2) = 1$, and so $a_2$ is the optimal action from $s_1$.

**2b.** No; $\pi^\star$ need not be an optimal policy for $M_2$. In fact, $\pi^\star$ must be an optimal policy for $M_2$ if $\alpha \geq 0$, but it need not be if $\alpha < 0$. To see this, take $M$ to be a 2-armed bandit with the arms' rewards being 0 and 1. Clearly the optimal arm in $M$ is sub-optimal in $M_2$ if we take $\alpha = -1$. Adding a constant $\beta$ to each reward does not change the relative order among policies, as it increments each value by the same amount ($\frac{\beta}{1-\gamma}$).

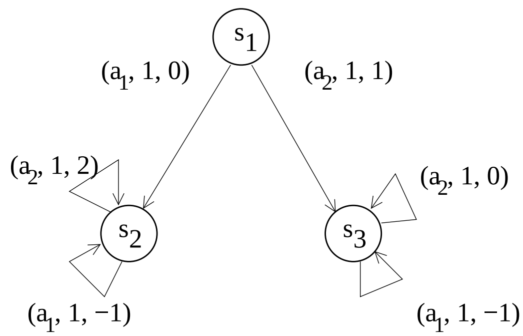


Figure 2: State transition diagram for $M$ ($\gamma = \frac{1}{2}$) and $M_1$ ($\gamma = \frac{1}{4}$). Each transition is annotated with (action, transition probability, reward).

**3a.** Solving Bellman's equations corresponding to the four policies, we get the following results.

$$V^{\pi^{11}}(s_1) = 0; V^{\pi^{11}}(s_2) = 1.$$
$$V^{\pi^{12}}(s_1) = 0; V^{\pi^{12}}(s_2) = \frac{3}{2}.$$
$$V^{\pi^{21}}(s_1) = \frac{15}{8}; V^{\pi^{21}}(s_2) = \frac{9}{4}.$$
$$V^{\pi^{22}}(s_1) = \frac{9}{5}; V^{\pi^{22}}(s_2) = \frac{21}{10}.$$

Clearly, $V^{\pi^{21}}$ dominates all the other value functions, and so:

$$V^{\star}(s_1) = \frac{15}{8}; V^{\star}(s_2) = \frac{9}{4}.$$

**3b.** From the answer to the previous question, we see that $\pi^{21}$ is an optimal policy, while $\pi^{11}$, $\pi^{12}$, and $\pi^{22}$ are not.

**3c.** Let $x_t$ denote the probability that at time $t$, the agent is in state $s_1$. Thus, the probability of being in state $s_2$ at time $t$ is $1 - x_t$. We are given that $x_1 = 1$; based on the transition dynamics of the MDP and the agent's policy, we get:

$$x_t = \frac{1}{2}x_{t-1} + \frac{1}{4}(1 - x_{t-1}) = \frac{x_{t-1} + 1}{4}.$$

Thus, $x_2 = \frac{1}{2}$, and $x_3 = \frac{3}{8}$.

**3d.** From the previous answer, we see that

$$x_t = \frac{x_1}{4^{t-1}} + \sum_{\tau=1}^{t-1} \frac{1}{4^{\tau}},$$

and therefore,

$$\lim_{t \to \infty} x_t = 0 + \frac{1/4}{1 - 1/4} = \frac{1}{3}.$$