CS 747 (Autumn 2016): Mid-semester Examination

Instructor: Shivaram Kalyanakrishnan

8.30 a.m. – 10.30 a.m., September 10, 2016, 101/103/105 New CSE Building

Total marks: 15

Note. For questions carrying 2 or more marks, provide brief justifications and/or calculations along with each answer to illustrate how you arrived at the answer. For questions carrying 1 mark, you need not describe the intermediate steps.

Question 1. Consider an n-armed bandit instance whose arms $a_1, a_2, \ldots, a_n$ yield Bernoulli (that is, 0 or 1) rewards with means $p_1, p_2, \ldots, p_n$, respectively, such that $p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_n$ (thus, a_1 is an 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 relating to the application of different sampling algorithms on this bandit instance.

1a. Let A_1 be an algorithm that (1) samples the n arms in a round-robin fashion *once*, and (2) for $t \geq n+1$, (2.1) if t is a perfect square $(1, 4, 9, 25, \ldots)$, samples an arm that is picked uniformly at random, else (2.2) samples an arm with the highest empirical mean (breaking ties uniformly at random). For A_1, what is $\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r^t]$? [1 mark]

1b. Let A_2 be an ϵ-greedy algorithm, $0 < \epsilon \leq 1$, that (1) samples each arm once, and (2) at every time step thereafter: with probability $1 - \epsilon$, samples an arm with the highest empirical mean (breaking ties uniformly at random), and with probability ϵ, samples an arm that is picked uniformly at random. For A_2, what is $\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r^t]$? [1 mark]

1c. Let A_3 be an algorithm that samples arms uniformly at random. For $t = 2, 4, 6, \ldots$, what is the probability that after t pulls made by A_3, the number of 0-rewards received and the number of 1-rewards received are *equal* (that is, the sequence $r^1, r^2, \ldots, r^t$ has $\frac{t}{2}$ 0's and $\frac{t}{2}$ 1's)? [1 mark]

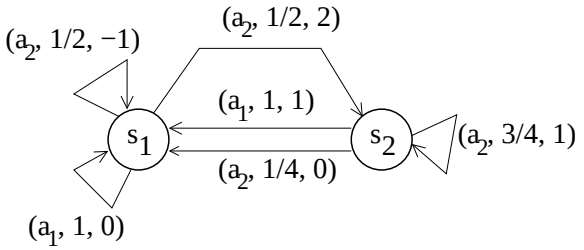
1d. Denote the probability described in 1c by $Q(t)$. Is $Q(t)$ a monotonically increasing function of t, a monotonically decreasing function of t, or neither? [1 mark]

1e. Describe a condition on our bandit instance (that is, on $p_1, p_2, \ldots, p_n$) such that $Q(t)$ is maximal. [1 mark]

Question 2. While discussing stochastic multi-armed bandits in class, we made the assumption of *stationarity*: that is, that the true means of the arms in the bandit instance being sampled do not change over time. On the contrary, in many practical applications (such as on-line advertising), the true means (click-though rates of ads) do gradually change with time. How would you design a regret-minimisation algorithm (or revise an existing one) for the *nonstationary* setting? Provide a high-level sketch in a few lines; do not write equations or pseudocode. [1 mark]

Question 3. Consider an 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 the table below; and a discount factor $\gamma = \frac{2}{3}$. A state transition diagram corresponding to M is shown below. In the figure, each transition is annotated with (action, transition probability, reward); transitions with zero probabilities are not shown. (You should be familiar with this MDP if you looked at last year's mid-semester paper: it is the MDP from Question 3 in that paper!)

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



Answer the following questions about this MDP.

3a. Consider a run of **Value Iteration**, which is initialised with $V^0(s_1) = -3; V^0(s_2) = 0$. What is the next iterate?: that is, what are $V^1(s_1)$ and $V^1(s_2)$? [2 marks]

3b. Consider an agent that at time $t = 1$, is in state s_2. At every time step, the agent follows policy π^{22}, by which it takes action a_2 from state s_1 and action a_2 from state s_2. What is the expected sum of the first three rewards that the agent accrues? Equivalently, using notation from class, what is $\mathbb{E}[r^1 + r^2 + r^3]$? [2 marks]

Question 4. Let $\mathcal{M}$ be the set of all MDPs with n states and 2 actions: that is, each element of $\mathcal{M}$ is an MDP (S, A, R, T, γ), such that $|S| = n$ and $|A| = 2$. For $M \in \mathcal{M}$, let $N(M)$ denote the number of deterministic *optimal* policies for M. Provide a set L of integers such that

1. For every $M \in \mathcal{M}$, $N(M)$ is an element of L, and
2. For every $l \in L$, there exists $M \in \mathcal{M}$ such that $N(M) = l$.

In other words, you must provide the set of all possible values $N(M)$ can take when M is drawn from $\mathcal{M}$. Prove that your answer is correct. [5 marks]

Solutions

1a. At time t, the number of times each arm has been pulled is at least in the order of $\sqrt{t}/n$, and so, by Hoeffding's inequality and union bounds, the probability that a suboptimal arm has the highest empirical mean is at most in the order of $t^A \cdot \exp(-B \cdot \sqrt{t}/n)$, for some constants $A > 0, B > 0$. Consequently the fraction of suboptimal "exploitation" pulls vanishes as $t \to \infty$. The fraction of "exploration" pulls is anyway in the order of $1/\sqrt{t}$, which also vanishes as $t \to \infty$. We are only left with optimal pulls. Hence, $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r^t] = p_1$.

1b. Since at time t, the number of times each arm has been pulled is at least in the order of $\epsilon t/n$, the probability that a suboptimal arm has the highest empirical mean is at most in the order of $t^A \cdot \exp(-B \cdot t/n)$, for some constants $A > 0, B > 0$. Consequently the fraction of suboptimal "exploitation" pulls vanishes as $t \to \infty$. In contrast with 1a, the fraction of "exploration" pulls remains a *constant* ϵ. Hence, $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r^t] = p_1(1-\epsilon) + (\epsilon/n) \sum_{i=1}^{n} p_i$.

1c. Since the probability of getting a 1-reward on a pull is $\frac{1}{n} \sum_{i=1}^{n} p_i$ (and the probability of getting a 0-reward is the remainder), the probability of getting an equal number of 0- and 1-rewards is:

$$Q(t) \stackrel{\text{def}}{=} \binom{t}{t/2} \left(\frac{1}{n}\sum_{i=1}^{n} p_i\right)^{t/2} \left(1 - \frac{1}{n}\sum_{i=1}^{n} p_i\right)^{t/2}.$$

1d. Observe that $Q(t)$ is uniformly 0 in the case that the means are all 0 or the means are all 1. If not, we have

$$\frac{Q(t+2)}{Q(t)} = 4 \left(\sum_{i=1}^{n} p_i\right) \left(1 - \sum_{i=1}^{n} p_i\right) \left(\frac{t+1}{t+2}\right) < 1.$$

Thus, when both 0- and 1-rewards are possible, $Q(t)$ is a monotonically decreasing function of t.

1e. From the expression for $Q(t)$, we observe that it is maximal when $\sum_{i=1}^{n} p_i = \frac{n}{2}$.

2. If the means of the arms change with time, we cannot trust "old" samples to be indicative of the current means: we must either discard them completely (by keeping a recency window) or discount them based on age. To make up for lack of information about recently unsampled arms, we also need to explore and learn at a non-vanishing—rather than decaying—rate in the nonstationary setting. If the drift in the means can be modeled, the prediction of the model can be used to guide exploration. At an extreme, if the means are changing quickly and unpredictably, we can run bandit algorithms for the adversarial (rather than stochastic) setting.

3a. By applying the Bellman Optimality Operator to V^0, we get:

$$V^1(s_1) = \max\left\{1\left(0 + \gamma V^0(s_1)\right), \frac{1}{2}\left(-1 + \gamma V^0(s_1)\right) + \frac{1}{2}\left(2 + \gamma V^0(s_2)\right)\right\} = -\frac{1}{2},$$

$$V^1(s_2) = \max\left\{1\left(1 + \gamma V^0(s_1)\right), \frac{1}{4}\left(0 + \gamma V^0(s_1)\right) + \frac{3}{4}\left(1 + \gamma V^0(s_2)\right)\right\} = \frac{1}{4}.$$

3b. Below are the possible 3-step trajectories, along with their rewards and probabilities.

s^1	s^2	s^3	s^4	$r^1 + r^2 + r^3$	Probability of trajectory
s_2	s_1	s_1	s_1	$0 + (-1) + (-1) = -2$	$\frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}$
s_2	s_1	s_1	s_2	$0 + (-1) + 2 = 1$	$\frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}$
s_2	s_1	s_2	s_1	$0 + 2 + 0 = 2$	$\frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{32}$
s_2	s_1	s_2	s_2	$0 + 2 + 1 = 3$	$\frac{1}{4} \cdot \frac{1}{2} \cdot \frac{3}{4} = \frac{3}{32}$
s_2	s_2	s_1	s_1	$1 + 0 + (-1) = 0$	$\frac{3}{4} \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{3}{32}$
s_2	s_2	s_1	s_2	$1 + 0 + 2 = 3$	$\frac{3}{4} \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{3}{32}$
s_2	s_2	s_2	s_1	$1 + 1 + 0 = 2$	$\frac{3}{4} \cdot \frac{3}{4} \cdot \frac{1}{4} = \frac{9}{64}$
s_2	s_2	s_2	s_2	$1 + 1 + 1 = 3$	$\frac{3}{4} \cdot \frac{3}{4} \cdot \frac{3}{4} = \frac{27}{64}$

From this table, we get

$$\mathbb{E}[r^1 + r^2 + r^3] = (-2)\frac{1}{16} + (1)\frac{1}{16} + (2)\frac{1}{32} + (3)\frac{3}{32} + (0)\frac{3}{32} + (3)\frac{3}{32} + (2)\frac{9}{64} + (3)\frac{27}{64} = \frac{135}{64}.$$

It is a good idea to verify the answer through an alternative argument. The expected number of time steps (among $\{1, 2, 3\}$) that the agent starts (and takes an action) from s_2 is $1 + \frac{3}{4} + \frac{1}{4} \cdot \frac{1}{2} + \frac{3}{4} \cdot \frac{3}{4} = \frac{39}{16}$. Therefore, the expected number of time steps it starts from s_1 is $3 - \frac{39}{16} = \frac{9}{16}$. The expected reward obtained by starting at s_2 and taking action a_2 is $\frac{3}{4}(1) + \frac{1}{4}(0) = \frac{3}{4}$, and the expected reward from starting at s_1 and taking action a_2 is $\frac{1}{2}(-1) + \frac{1}{2}(2) = \frac{1}{2}$. In all, the expected three-step reward is therefore $\frac{39}{16} \cdot \frac{3}{4} + \frac{9}{16} \cdot \frac{1}{2} = \frac{135}{64}$.

4. $L = \{1, 2, 4, 8, \ldots, 2^n\}$. In other words, (1) the number of optimal policies in any element of $\mathcal{M}$ is of the form 2^i, where $i \in \{0, 1, 2, \ldots, n\}$, and indeed (2) for every $i \in \{0, 1, 2, \ldots, n\}$, there exists an MDP $M \in \mathcal{M}$ such that $N(M) = i$. We shall prove (1) and (2) in turn. We find the following two definitions useful for our proof.

- Since A has exactly two actions, for $a \in A$, denote by a^c the element of $A \setminus \{a\}$.
- For $\pi_1, \pi_2 \in \Pi$, denote by $\text{Diff}(\pi_1, \pi_2)$ the set of states on which π_1 and π_2 take different actions: $\text{Diff}(\pi_1, \pi_2) \stackrel{\text{def}}{=} \{s \in S : \pi_1(s) \neq \pi_2(s)\}$.

(1) Consider an arbitrary MDP $M \in \mathcal{M}$. We know that every MDP has at least one optimal policy, a unique optimal value function $V^\star$, and a unique action value function $Q^\star$. Let $\pi^\star$ be an arbitrary optimal policy for M, which we shall fix as a basis. Now, let $S^\star_=$ be the set of states in which the *complementary* action to $\pi^\star$ has an *equal* Q-value under $\pi^\star$:

$$S^\star_= \stackrel{\text{def}}{=} \{s \in S : Q^\star(s, (\pi^\star(s))^c) = Q^\star(s, \pi^\star(s))\}.$$

It is easily verified that for any policy $\pi \in \Pi$, if $\text{Diff}(\pi, \pi^\star) \subseteq S^\star_=$, then $B^\pi(V^\star) = V^\star$. By a working similar to that in the proof of the policy improvement theorem, it follows that $V^\pi = V^\star$. On the other hand, if $\text{Diff}(\pi, \pi^\star) \setminus S^\star_= \neq \emptyset$, we find that $V^\star \succ B^\pi(V^\star)$, and therefore, $V^\star \succ V^\pi$. We conclude that a policy $\pi \in \Pi$ is optimal if and only if $\text{Diff}(\pi, \pi^\star) \subseteq S^\star_=$. Thus, the number of optimal policies in M is exactly the number of subsets of $S^\star_=$, which is $2^{|S^\star_=|}$.

(2) Given $i \in \{0, 1, 2, \ldots, n\}$, we describe an MDP $M \in \mathcal{M}$ such that $N(M) = 2^i$. Take $M = (S, A, R, T, \gamma)$, wherein $S = \{s_1, s_2, \ldots, s_n\}$, $A = \{a_1, a_2\}$, and $\gamma = 0$. M will only contain *self-loops* at each state: for $s \in S, a \in A$, $T(s, a, s) = 1$ (and so, for $s' \in S, s' \neq s$, we get $T(s, a, s') = 0$). By this construction, the action value function is completely determined by the reward function, which we manipulate to our end. We define, for $s, s' \in S, a \in A$:

$$R(s, a, s') = \begin{cases} 1 & a = a_1; \\ 1 & a = a_2 \wedge s \in \{s_1, s_2, \ldots, s_i\}; \text{ and} \\ 0 & a = a_2 \wedge s \in \{s_{i+1}, s_{i+2}, \ldots, s_n\}. \end{cases}$$

For this MDP, we observe that every policy that picks either a_1 or a_2 in states $s_1, s_2, \ldots, s_i$, and picks action a_1 in states $s_{i+1}, s_{i+2}, \ldots, s_n$, is an optimal policy. Every policy that picks a_2 in even one state among $s_{i+1}, s_{i+2}, \ldots, s_n$ is not an optimal policy. Thus, the number of optimal policies for this MDP is 2^i.