# CS 747 (Autumn 2017): Mid-semester Examination

Instructor: Shivaram Kalyanakrishnan

8.30 a.m. – 10.30 a.m., September 12, 2016, 103 New CSE Building

Total marks: 15

**Question 1.** A coin that generates 1-rewards (heads) with probability $p = 0.6$ and 0-rewards (tails) with probability $(1 - p) = 0.4$ is to be tossed $u = 4$ times. After the 4 tosses, let $\hat{p}$ be the resulting empirical mean (number of heads obtained/4).

1a. Apply Hoeffding's Inequality to obtain an upper bound on the probability that $\hat{p} > 0.7$. Your answer can be in terms of $e$. [1 mark]

1b. What is the probability that $\hat{p} > 0.7$? [1 mark]

**Question 2.** Consider a 2-armed bandit instance whose arms yield Bernoulli (that is, 0 or 1) rewards. Interestingly, this bandit instance is such that the mean reward of one of the arms is positive, while that of the other arm is exactly 0.

2a. Specify an algorithm for sampling this bandit instance, with the aim of minimising the expected cumulative regret. The algorithm can use the fact that one arm has a positive mean and the other a zero mean, but the algorithm neither knows (1) which arm has which mean, nor knows (2) the value of the positive mean. [2 marks]

2b. Assuming a sufficiently large horizon $T > 0$ (so terms of the form $x^T$ can be ignored for $x \in (0, 1)$), calculate the expected cumulative regret of your algorithm as a function of the positive mean $p > 0$. [2 marks]

2c. Discuss your answer to 2b in the light of Lai and Robbins's lower bound. Do you notice a contradiction?—why or why not? [1 mark]

**Question 3.** Consider an MDP $M = (S, A, T, R, \gamma)$, in which $|S| = n \geq 2$ and $|A| = 2$. If for policy $\pi : S \to A$, the set of improvable states is denoted $I(\pi)$, recall that a policy $\pi' : S \to A$ is said to locally improve upon $\pi$ if for some non-empty set $S' \subseteq I(\pi)$:

1. for all $s \in S'$, $\pi'(s) = (\pi(s))^c$, where for $a \in A$, $a^c$ denotes the action in $A$ other than $a$ (recall that $|A| = 2$), and
2. for all $s \in S \setminus S'$, $\pi'(s) = \pi(s)$.

It follows that in general, there could be multiple locally improving policies for a given policy $\pi$. In this question, assume "policy" to mean "deterministic policy". (Question 4, however, explicitly asks you about a stochastic policy.)

3a. Show that for every policy $\pi$, there exists a sequence of $m$ policies $\pi_1, \pi_2, \ldots, \pi_m$ such that

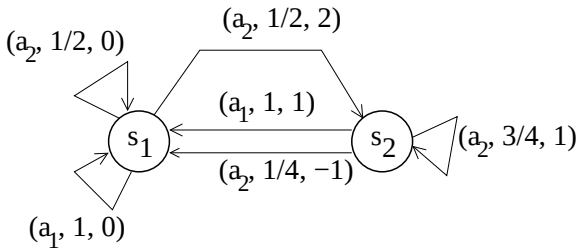
1. $\pi_1 = \pi$,
2. for $i \in \{1, 2, \ldots, m-1\}$, $\pi_{i+1}$ locally improves upon $\pi_i$,
3. $\pi_m$ is an optimal policy, and
4. $m \leq n + 1$.

In other words, show that starting with any policy $\pi$, there exists a sequence of policy improvement operations that will lead to an optimal policy in no more than $n$ steps. [4 marks]

3b. The tightest upper bounds shown to date on the number of iterations taken by Policy Iteration are exponential in $n$. Does the proof you are asked to furnish in 3a imply a linear bound for Policy Iteration? Why or why not? [1 mark]

**Question 4.** Consider an MDP $M = (S, A, T, R, \gamma)$, 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 $\gamma = \frac{1}{2}$. 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.

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




This question pertains to the *stochastic* policy $\pi$ that takes actions with equal probability in both states: that is,

$$\pi(s_1, a_1) = \pi(s_1, a_2) = \pi(s_2, a_1) = \pi(s_2, a_2) = \frac{1}{2}.$$

What is $Q^\pi(s_1, a_2)$?

4a. First write down *exactly* how $Q^\pi(s_1, a_2)$ is related to $T$, $R$, and $\gamma$. It is okay to do so using multiple steps. [2 marks]

4b. Substitute numeric values and simplify to obtain the answer. [1 mark]

# Solutions

**1a.** $e^{-2\times 4\times (0.7-0.6)^2} = e^{-0.08}$ ($\approx 0.9232$).

**1b.** For $\hat{p}$ to exceed 0.7, either 3 or all 4 of the tosses must yield heads, the probability of which happening is $4 \times 0.6^3 \times 0.4 + 0.6^4 = 0.4752$.

**2a.** Given the structure of the bandit instance, it is clear that once an arm gives a 1-reward, the right strategy is to pull that arm (the non-zero-mean arm) at every time step thereafter. Until a 1-reward is obtained, the arms must be sampled roughly equally, in order to maximise the number of pulls given to the yet unknown positive-mean arm. Round robin sampling is a good idea. A slight advantage can be obtained by randomising the choice of the arm pulled at the beginning (rather than starting with an arbitrary arm). Here is our proposed algorithm.

1. Pull an arm uniformly at random. If a 1-reward is obtained, sample it for ever.
2. Pull the arm not pulled in the previous step. If a 1-reward is obtained, sample it for ever.
3. Go to 2.

**2b.** Assume the arms are sampled at time steps $1, 2, 3, \ldots$. Let $t \geq 1$ be the first time step at which a 1-reward is obtained. If the arm that gave this reward is sampled at every time step thereafter, then no expected regret is incurred for time steps $t+1$ and after. The total regret obtained over the first $t-1$ steps is $p(t-1)$, since 0-rewards were obtained in all those steps. The regret from the $t$-th time step is $p-1$ (which could be negative).

The expected cumulative regret of the algorithm becomes

$$\mathbb{E}[p(t-1)] + p - 1 + 0 = p\mathbb{E}[t] - 1.$$

Essentially, we have to calculate $\mathbb{E}[t]$. Observe that the arms are sampled in pairs until a success is recorded. The probability of registering a success in any given pair is exactly $p$, and so the expected number of *unsuccessful pairs* seen before the first successful one is

$$p(0) + p(1-p)(1) + p(1-p)^2(2) + \cdots + p(1-p)^{T-1}(T-1) = \frac{1-p}{p}$$

for large $T$. The expected number of pulls in the first successful pair before obtaining a 1-reward is $\frac{1}{2}(0) + \frac{1}{2}(1) = \frac{1}{2}$. It follows that

$$\mathbb{E}[t] = 2\left(\frac{1-p}{p}\right) + \frac{1}{2},$$

and the regret of the algorithm is

$$p\mathbb{E}[t] - 1 = 2 - 1.5p.$$

Had we not randomised the choice of the first arm to pull, the only difference is that in the bad case, we would have had to incur 1 pull (rather than $\frac{1}{2}$ a pull) in the first successful pair—which would lead to an overall regret bound of $2 - p$.

**2c.** We designed our algorithm specifically for instances in which one mean is known to be exactly zero, and we have shown it achieves constant regret on such instances. Notice that the algorithm

stops sampling an arm as soon as the other arm gives a 1-reward. Hence, the algorithm can incur linear regret on bandit instances in which both means are positive. Lai and Robbins's result is that if an algorithm incurs sub-polynomial regret on every bandit instance, then it must incur at least logarithmic regret on every instance. Clearly our algorithm does not qualify the precondition of Lai and Robbins's statement, and so the logarithmic lower bound does not apply.

**3a.** Let $\pi^\star : S \to A$ be an optimal policy for the MDP, which we fix for the purpose of this proof. For every policy $\pi : S \to A$, let $D(\pi)$ be the set of states on which $\pi$ *differs* from $\pi^\star$: that is,

$$D(\pi) = \{s \in S, \pi(s) \neq \pi^\star(s)\}.$$

Also, for convenience, let us denote $d(\pi) = |D(\pi)|$. Observe that $D(\pi^\star) = \emptyset$, and also that $d(\pi)$ is at most $n$. We accomplish the proof by furnishing, for any given non-optimal policy $\pi$, a locally improving policy $\pi'$, which is such that $d(\pi') = d(\pi) - 1$. It follows that an optimal policy must be reached after at most $n$ such local improvements.

Consider a non-optimal policy $\pi$, and recall that $I(\pi)$ denotes the set of improvable states for $\pi$. We observe that $I(\pi) \cap D(\pi) \neq \emptyset$. On the contrary, if every state on which $\pi$ and $\pi^\star$ differ is not in $I(\pi)$, it follows from the Policy Deprovement Theorem (after switching all the states in $D(\pi)$) that $\pi \succeq \pi^\star$. Clearly this result is impossible, since $\pi$ is non-optimal and $\pi^\star$ is optimal.

Having established that $I(\pi) \cap D(\pi) \neq \emptyset$, we choose an arbitrary state $s_{\text{switch}}$ from the intersection. We define a policy $\pi' : S \to A$ as follows:

$$\pi'(s_{\text{switch}}) = (\pi(s_{switch}))^c, \text{ and for } s \in S \setminus \{s_{\text{switch}}\}, \pi'(s) = \pi(s).$$

In other words, only the state $s_{\text{switch}}$ is switched to go from $\pi$ to $\pi'$. Since $s_{\text{switch}} \in I(\pi)$, it follows that $\pi'$ locally improves upon $\pi$. Since $s_{\text{switch}} \in D(\pi)$, we have $d(\pi') = d(\pi) - 1$. Our proof is done.

**3b.** Observe that in our proof, the single state $s_{\text{switch}}$ that is switched for each non-optimal policy $\pi$ is picked based on the knowledge of $D(\pi)$. Thus, while such a state exists, it is not feasible for a PI algorithm to know $D(\pi)$ or its intersection with $I(\pi)$ (unless $|I(\pi)| = 1$). Given a policy and an improvement set, a PI algorithm picks a subset of the improvement set to switch—it executes the same choice regardless of the MDP on which is it being run. Hence, while there exists a chain of policy improvements that can reach an optimal policy in $n$ or fewer steps from every policy for every given MDP, a PI algorithm, in general, need not implement such a chain.

**4a.** Bellman's Equations are modified appropriately to account for stochasticity in action selection: for $s \in \{s_1, s_2\}, a \in \{a_1, a_2\}$,

$$V^\pi(s) = \sum_{a \in A} \pi(s, a) \sum_{s' \in S} T(s, a, s')\{R(s, a, s') + \gamma V^\pi(s')\}.$$

Once we solve for $V^\pi$, we can obtain $Q^\pi$. Specifically,

$$Q^\pi(s_1, a_2) = \sum_{s' \in S} T(s_1, a_2, s')\{R(s_1, a_2, s') + \gamma V^\pi(s')\}.$$

**4b.** Solving for $V^\pi$ gives $V^\pi(s_1) = \frac{16}{15}$ and $V^\pi(s_2) = \frac{4}{3}$, and thus $Q^\pi(s_1, a_2) = \frac{8}{5}$.