CS 747 (Autumn 2018): Mid-semester Examination

Instructor: Shivaram Kalyanakrishnan

1.30 p.m. – 3.30 p.m., September 12, 2018, LA 301/302

Total marks: 15

Note. Provide 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 a_1 and a_2 have means p_1 and p_2, respectively, with $1 > p_1 > p_2 > 0$. Each arm yields i.i.d. Bernoulli rewards with the corresponding mean. Hence, each reward obtained is either 0 or 1.

An algorithm $\mathcal{L}$ is applied to this bandit instance. At every step, $\mathcal{L}$ pulls whichever arm has obtained the *least* number of 0-rewards up to then, breaking ties uniformly at random. Thus, the very first pull is equally likely to come from a_1 and a_2. Suppose a_2 was pulled and it gives a 1-reward, then again both arms are equally likely to be picked. If a_1 is now pulled, and it gives a 0-reward, then a_2 will be pulled next, and repeatedly until it gives a 0-reward. At this point, both arms will again have an equal number of 0-rewards, and therefore be equally likely to be pulled, and so on.

For $T \geq 1$, let z^T denote the number of 0-rewards obtained in the first T pulls. Answer the following questions, providing calculations in terms of the mean rewards p_1 and p_2. If it is convenient, you can also use $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

1a. What is $\mathbb{E}[z^2]$? [2 marks]

1b. What is $\lim_{T\to\infty} \frac{\mathbb{E}[z^T]}{T}$? [4 marks]

1c. Let R^T denote the cumulative regret after T pulls. What is $\lim_{T\to\infty} \frac{\mathbb{E}[R^T]}{T}$? You might find it useful to first express R^T in terms of z^T, and then use your answer from 1b. [2 marks]

Question 2. Fix an MDP (S, A, T, R, γ). For fixed $\epsilon \in [0, 1]$, let Π_ϵ be the set of all stochastic policies π_ϵ such that $\forall s \in S, \forall a \in A$: $\pi_\epsilon(s, a) \geq \frac{\epsilon}{|A|}$. Here $\pi_\epsilon(s, a)$ denotes the probability with which π_ϵ takes action a from state s. You may view the elements of Π^ϵ as policies that are constrained to take actions uniformly at random with with probability ϵ, and with the remaining probability, can pick actions arbitrarily (deterministically or stochastically) from each state. We refer to such policies as ϵ-soft policies. (Note that Π_0 does contain deterministic policies.)

2a. Prove Π_ϵ contains an *optimal* ϵ-soft policy $\pi^\star_\epsilon$, such that $\forall s \in S, \forall \pi_\epsilon \in \Pi_\epsilon$: $V^{\pi^\star_\epsilon}(s) \geq V^{\pi_\epsilon}(s)$. Recall that we proved this statement in class for $\epsilon = 0$. Your easiest way to answer the question might be to follow a similar approach. [5 marks]

2b. Consider $\epsilon, \epsilon' \in [0, 1]$ such that $\epsilon < \epsilon'$. Let $\pi^\star_\epsilon$ be an optimal ϵ-soft policy, and $\pi^\star_{\epsilon'}$ an optimal ϵ'-soft policy. Show that $\forall s \in S$: $V^{\pi^\star_\epsilon}(s) \geq V^{\pi^\star_{\epsilon'}}(s)$. [2 marks]

Solutions

1a. The probability that the first pull is given to a_1 is $\frac{1}{2}$; the probability that the second pull is given to a_1 is $\frac{1}{2} + \frac{1}{2} \cdot p_1 \cdot \frac{1}{2} + \frac{1}{2} \cdot p_2 \cdot \frac{1}{2} + \frac{1}{2} \cdot q_2$. Hence, the expected number of pulls of a_1 in the first two pulls is $\frac{1}{2}(1 + \frac{p_1}{2} + \frac{p_2}{2} + q_2)$. Similarly, the expected number of pulls of a_2 among the first two pulls is $\frac{1}{2}(1 + \frac{p_1}{2} + \frac{p_2}{2} + q_1)$. The expected number of 0-rewards is

$$\frac{1}{2}(1 + \frac{p_1}{2} + \frac{p_2}{2} + q_2) \cdot q_1 + \frac{1}{2}(1 + \frac{p_1}{2} + \frac{p_2}{2} + q_1) \cdot q_2,$$

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

1b. We observe from the definition of the algorithm that it performs a series of "compound pulls", wherein in each compound pull, either (1) arm a_1 is pulled until it returns a 0-reward, and thereafter arm a_2 pulled until it yields a 0-reward, or (2) arm a_2 is pulled until it returns a 0-reward, and thereafter arm a_1 pulled until it yields a 0-reward. The cases are equally likely, since each compound pull begins when the arms have both registered the same number of 0-rewards.

Let n^T denote the number of compound pulls that have been completed within T (atomic) pulls. Clearly each completed compound pull has exactly two 0-rewards, while a compound pull that is underway might have either a single 0-reward or none at all. Thus, if T exactly coincides with the completion of a compound pull, then z^T is exactly $2n^T$; otherwise z^T is either $2n^T$ or $2n^T + 1$. Regardless, notice that $\lim_{T\to\infty} \frac{\mathbb{E}[z^T]}{T} = 2 \lim_{T\to\infty} \frac{\mathbb{E}[n^T]}{T}$. We calculate $\lim\limits_{T\to\infty} \dfrac{\mathbb{E}[n^t]}{T}$ by accounting for the length (the number of atomic pulls) of each compound pull.

The expected number of pulls of a_1 in a compound pull is $1(q_1) + 2(p_1 q_1) + 3(p_1 p_1 q_1) + \cdots = \frac{1}{q_1}$, and similarly the expected number of pulls of a_2 in a compound pull is $\frac{1}{q_2}$. Hence, the expected length of a compound pull is $\frac{1}{q_1} + \frac{1}{q_2}$. As $T \to \infty$, the law of large numbers implies that the empirical mean length of compound pulls converges to $\frac{1}{q_1} + \frac{1}{q_2}$, and therefore $\frac{\mathbb{E}[n^T]}{T}$ to $\frac{1}{\frac{1}{q_1} + \frac{1}{q_2}}$. It follows that

$$\lim_{T\to\infty} \frac{\mathbb{E}[z^T]}{T} = \frac{2q_1 q_2}{q_1 + q_2}.$$

1c. The cumulative reward from T pulls is exactly $T - z^T$, and so $R^T = Tp_1 - (T - z^T)$. Therefore,

$$\lim_{T\to\infty} \frac{\mathbb{E}[R^T]}{T} = p_1 - 1 + \lim_{T\to\infty} \frac{\mathbb{E}[z^T]}{T} = \frac{q_1(q_2 - q_1)}{q_1 + q_2} = \frac{(1 - p_1)(p_1 - p_2)}{2 - p_1 - p_2}.$$

2a. There are at least two distinct ways to prove the result, both of which shed light on the nature of the problem. The first solution, which is more comprehensive, involves generalising the proof of the policy improvement theorem. The second approach is to reduce the given problem to a known one (the case of $\epsilon = 0$) and use existing results. We present both solutions.

Below we consider ϵ to be a fixed element of $(0, 1)$: the case of $\epsilon = 0$ was already solved in class, and the case of $\epsilon = 1$ is trivial since Π_1 contains only one policy.

Solution 1. Define "extremal" ϵ-soft policies to be those that allocate, for each state, exactly $\frac{\epsilon}{|A|}$ probability to some $|A|-1$ actions, and $1-\epsilon+\frac{\epsilon}{|A|}$ probability to the remaining action (which could vary from state to state). Our first step is to show that every ϵ-soft policy $\pi_\epsilon \in \Pi_\epsilon$, there is an extremal ϵ-soft policy $\pi'_\epsilon \in \Pi_\epsilon$ such that $V^{\pi'_\epsilon} \succeq V^{\pi_\epsilon}$. In fact, for each state $s \in S$, let

$$\bar{a} = \arg\max_{a \in A} Q^{\pi_\epsilon}(s,a), \text{ breaking ties arbitrarily, and take}$$

$$\pi'_\epsilon(s,a) = \begin{cases} 1-\epsilon+\frac{\epsilon}{|A|}, \text{ if } a = \bar{a}, \\ \frac{\epsilon}{|A|}, \text{ otherwise.} \end{cases}$$

With this definition, it can be verified that $B^{\pi'_\epsilon}(V^{\pi_\epsilon}) \succeq V^{\pi_\epsilon}$; for $s \in S$:

$$B^{\pi'_\epsilon}(V^{\pi_\epsilon})(s) = \sum_{a \in A} \pi'_\epsilon(s,a) Q^{\pi_\epsilon}(s,a) \geq \sum_{a \in A} \pi_\epsilon(s,a) Q^{\pi_\epsilon}(s,a) = V^{\pi_\epsilon}(s).$$

As we have already seen in the proof of the policy improvement theorem, the result implies $\pi'_\epsilon \succeq \pi_\epsilon$. Also notice that if π_ϵ allocates more than $\frac{\epsilon}{|A|}$ probability from some state to any action that does not maximise its action value function, then $\pi'_\epsilon \succ \pi_\epsilon$. In fact, even if π_ϵ was itself extremal, but not ϵ-greedy with respect to it action value function, we see from the same working that there exists an extremal ϵ-soft policy π'_ϵ such that $\pi'_\epsilon \succ \pi_\epsilon$.

Observe that the set of extremal ϵ-soft policies is finite. Hence, it now suffices for us to show that if there is an extremal ϵ-soft policy π^g_ϵ that is ϵ-greedy with respect to its action value function, then it is at least as rewarding from each state as every other ϵ-soft policy π_ϵ. This is immediately apparent when we expand $B^{\pi_\epsilon}(V^{\pi^g_\epsilon})$: we verify that $V^{\pi^g_\epsilon} \succeq B^{\pi_\epsilon}(V^{\pi^g_\epsilon})$, which, in turn, yields $\pi^g_\epsilon \succeq \pi_\epsilon$.

Our proof is curiously similar to the one for $\epsilon = 0$; what exactly is the difference? In the $\epsilon = 0$ case, we test a state s under policy π for improvability by checking if there is an action a such that $Q^{\pi(s,a)} > V^\pi(s)$.

For $\epsilon > 0$ and ϵ-soft policy π_ϵ, the test is *not* if there is an action a such that $Q^{\pi_\epsilon(s,a)} > V^{\pi_\epsilon}(s)$, but rather, if there is a policy π_ϵ such that $B^{\pi'_\epsilon}(V^{\pi_\epsilon})(s) > V^{\pi_\epsilon}(s)$. In general, we ask: is there a different *policy*, which we follow for one step before adopting the original policy, which will increase expected long-term return? If the policy in question is deterministic (as we conveniently take when $\epsilon = 0$), the question incidentally reduces to a comparison of action values. In general it does not when the improving policy is stochastic, as it is here.

Solution 2. Let $M = (S, A, T, R, \gamma)$ be the given MDP, and π_ϵ an arbitrary ϵ-soft policy. For $s \in S, a \in A$, we may view $\pi_\epsilon(s,a)$ as the sum of two probabilities: $\frac{\epsilon}{|A|}$, which it is mandatory for π_ϵ to allocate each action, and $\rho(s,a) \in [0, 1-\epsilon]$, which is the portion "deliberately" allocated by π_ϵ. If we remove the constraint of ϵ-softness, a natural way to change π_ϵ would be to take action a from state s solely with the "deliberate" portion of the probability, scaled appropriately. Indeed let the policy π_0 do exactly that: for $s \in S, a \in A$:

$$\pi_0(s,a) = \frac{\rho(s,a)}{1-\epsilon} = \frac{\pi_\epsilon(s,a) - \frac{\epsilon}{|A|}}{1-\epsilon}.$$

In essence we have a 1-to-1 correspondence between elements of Π_ϵ and Π_0. Our strategy is to define a new MDP $\bar{M} = (S, A, \bar{T}, \bar{R}, \gamma)$ such that for every $\pi_\epsilon \in \Pi_\epsilon$ and $\pi_0 \in \Pi_0$ that map to each other, the value function of π_ϵ on M is *identical* to the value function of π_0 on $\bar{M}$. We already know that Π_0 must contain an optimal policy for every MDP with state space S and action space A; let $\pi\star_0 \in \Pi_0$ be an optimal policy for $\bar{M}$. It follows that the corresponding ϵ-soft policy $\pi\star_\epsilon \in \Pi_\epsilon$ must be an optimal ϵ-soft policy for M.

We define $\bar{M} = (S, A, \bar{T}, \bar{R}, \gamma)$ by essentially transferring the ϵ-fraction of uniform action selection from the agent to the environment. Hence, it will be as though the environment ($\bar{T}$, $\bar{R}$)

implements uniform transitions with probability ϵ, accounting appropriately for the expected rewards. We define $\forall s, s' \in S, a \in A$:

$$\bar{T}(s,a,s') = (1-\epsilon)T(s,a,s') + \frac{\epsilon}{|A|}\sum_{b\in A} T(s,b,s'), \text{ and}$$

$$\bar{R}(s,a,s') = \sum_{x\in S}\left((1-\epsilon)T(s,a,x)R(s,a,x) + \frac{\epsilon}{|A|}\sum_{b\in A} T(s,b,x)R(s,b,x)\right).$$

It is easily verified that $\bar{T}$ is a genuine transition function: that is, each element is non-negative, and the elements corresponding to every state-action pair sum to 1. Observe that $\bar{R}(s,a,s')$ does not depend on s'. This choice simplifies our calculaitons; in general we can also construct consistent reward functions that do depend on the next states of transitions.

It remains to be shown that the value function of π_ϵ on M is the same as that of π_0 on $\bar{M}$. To do so, we use the fact that the value function is the unique solution of Bellman's Equations, and show that π_0 has the same set of equations on $\bar{M}$ as π_ϵ on M. For $s \in S$:

$$\begin{aligned}
V^{\pi_0}_{\bar{M}}(s) &= \sum_{a\in A}\pi_0(s,a)\sum_{s'\in S}\bar{T}(s,a,s')\left(\bar{R}(s,a,s') + \gamma V^{\pi_0}_{\bar{M}}(s')\right)\\
&= \sum_{a\in A}\frac{\pi_\epsilon(s,a) - \frac{\epsilon}{|A|}}{1-\epsilon}\sum_{s'\in S}\bar{T}(s,a,s')\left(\bar{R}(s,a,s') + \gamma V^{\pi_0}_{\bar{M}}(s')\right)\\
&= G_1 + G_2, \text{ where}
\end{aligned}$$

$$\begin{aligned}
G_1 &= \sum_{a\in A}\frac{\pi_\epsilon(s,a) - \frac{\epsilon}{|A|}}{1-\epsilon}\sum_{s'\in S}\bar{T}(s,a,s')\bar{R}(s,a,s')\\
&= \sum_{a\in A}\frac{\pi_\epsilon(s,a) - \frac{\epsilon}{|A|}}{1-\epsilon}\sum_{x\in S}\left((1-\epsilon)T(s,a,x)R(s,a,x) + \frac{\epsilon}{|A|}\sum_{b\in A}T(s,b,x)R(s,b,x)\right)\\
&= \sum_{a\in A}\left(\pi_\epsilon(s,a) - \frac{\epsilon}{|A|}\right)\sum_{x\in S}\left(T(s,a,x)R(s,a,x) + \frac{\epsilon}{(1-\epsilon)|A|}\sum_{b\in A}T(s,b,x)R(s,b,x)\right)\\
&= \sum_{a\in A}\pi_\epsilon(s,a)\sum_{x\in S}T(s,a,x)R(s,a,x) - \frac{\epsilon}{|A|}\sum_{a\in A}\sum_{x\in S}T(s,a,x)R(s,a,x) +\\
&\frac{\epsilon}{(1-\epsilon)|A|}\left(\sum_{a\in A}\left(\pi_\epsilon(s,a) - \frac{\epsilon}{|A|}\right)\right)\left(\sum_{x\in S}\sum_{b\in A}T(s,b,x)R(s,b,x)\right)\\
&= \sum_{a\in A}\pi_\epsilon(s,a)\sum_{x\in S}T(s,a,x)R(s,a,x), \text{ and}
\end{aligned}$$

$$
\begin{aligned}
G_2 &= \gamma \sum_{a \in A} \frac{\pi_\epsilon(s,a) - \frac{\epsilon}{|A|}}{1-\epsilon} \sum_{s' \in S} \bar{T}(s,a,s') V_{\bar{M}}^{\pi_0}(s') \\
&= \gamma \sum_{a \in A} \left(\pi_\epsilon(s,a) - \frac{\epsilon}{|A|} \right) \sum_{s' \in S} \left(T(s,a,s') + \frac{\epsilon}{(1-\epsilon)|A|} \sum_{b \in A} T(s,b,s') \right) V_{\bar{M}}^{\pi_0}(s') \\
&= \gamma \sum_{a \in A} \pi_\epsilon(s,a) \sum_{s' \in S} T(s,a,s') V_{\bar{M}}^{\pi_0}(s') - \frac{\gamma\epsilon}{|A|} \sum_{a \in A} \sum_{s' \in S} T(s,a,s') V_{\bar{M}}^{\pi_0}(s') + \\
&\frac{\gamma\epsilon}{(1-\epsilon)|A|} \left(\sum_{a \in A} \left(\pi_\epsilon(s,a) - \frac{\epsilon}{|A|} \right) \right) \left(\sum_{s' \in S} \sum_{b \in A} T(s,b,s') V_{\bar{M}}^{\pi_0}(s') \right) \\
&= \gamma \sum_{a \in A} \pi_\epsilon(s,a) \sum_{s' \in S} T(s,a,s') V_{\bar{M}}^{\pi_0}(s'), \text{ which gives}
\end{aligned}
$$

$$
V_{\bar{M}}^{\pi_0}(s) = \sum_{a \in A} \pi_\epsilon(s,a) \sum_{s' \in S} T(s,a,s') \left(R(s,a,s') + \gamma V_{\bar{M}}^{\pi_0}(s') \right).
$$

Since we know that $V_M^{\pi_\epsilon}$ is the unique solution of this set of equations, it is clear that $V_{\bar{M}}^{\pi_0}$ and $V_M^{\pi_\epsilon}$ are identical.

2b. Note that $\epsilon < \epsilon'$. It follows from the definition that since $\pi^\star_{\epsilon'}$ is an ϵ'-soft policy, it is also an ϵ-soft policy. And since $\pi^\star_\epsilon$ is an optimal ϵ-soft policy, we get from 2a that $\pi^\star_\epsilon \succeq \pi^\star_{\epsilon'}$.