CMPUT 605: Theoretical Foundations of Reinforcement Learning, Winter 2023
Homework #3

Instructions

Submissions You need to submit a single PDF file, named `p03_<name>.pdf` where `<name>` is your name. The PDF file should include your typed up solutions (we strongly encourage to use pdfLaTeX). Write your name in the title of your PDF file. We provide a LaTeXtemplate that you are encouraged to use. To submit your PDF file you should send the PDF file via private message to Vlad Tkachuk on Slack before the deadline.

Collaboration and sources Work on your own. You can consult the problems with your classmates, use books or web, papers, etc. Also, the write-up must be your own and you must acknowledge all the sources (names of people you worked with, books, webpages etc., including class notes.) Failure to do so will be considered cheating. Identical or similar write-ups will be considered cheating as well. Students are expected to understand and explain all the steps of their proofs.

Scheduling Start early: It takes time to solve the problems, as well as to write down the solutions. Most problems should have a short solution (and you can refer to results we have learned about to shorten your solution). Don't repeat calculations that we did in the class unnecessarily.

Deadline: March 12 at 11:55 pm

Average vs. mixed policies

Fix policies $\pi^{(1)}, \dots, \pi^{(k)}$ of some finite discounted MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$. There are two ways of combining these policies with some weights $\alpha \in \mathcal{M}_1([k])$. The first way is to choose one of the policies at random from the multinomial parameterized by α and then follow the resulting policy for all the time steps. Formally, one would choose an index $I \in [k]$ at random such that $\mathbb{P}(I = i) = \alpha_i$ and then follow the policy $\pi^{(I)}$ for whichever state one encounters. The second way is to choose the policy to follow at random in each time step. Call the policy that is obtained following the first method the (α-weighted) **mixture of** $\pi^{(1)}, \dots, \pi^{(k)}$. Call the policy that is obtained following the second method the (α-weighted) **average of** $\pi^{(1)}, \dots, \pi^{(k)}$.

Intuitively, a distribution $\mu \in \mathcal{M}_1(\mathcal{S})$ over the states and the interconnection of a mixture policy and M gives rise to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that carries the random elements $I, S_0, A_0, S_1, A_1, \dots$ with $I \in [k]$, $S_t \in \mathcal{S}$ and $A_t \in \mathcal{A}$ for $t \geq 0$ and such that for $H_t = (S_0, A_0, S_1, \dots, A_{t-1}, S_t)$,

1. $\mathbb{P}(S_0 = s | I) = \mu(s)$ for all $s \in \mathcal{S}$,
2. $\mathbb{P}(A_t = a | I, H_t) = \pi_t^{(I)}(a|H_t)$ for all $a \in \mathcal{A}, t \geq 0$,
3. $\mathbb{P}(S_{t+1} = s' | I, H_t, A_t) = P_{A_t}(S_t, s')$ for all $s' \in \mathcal{S}$, and
4. $\mathbb{P}(I = i) = \alpha_i$ for all $i \in [k]$.

Note that all first three criteria are modified to express that the laws that govern S_0, the action distribution and the next state distribution are as before even when conditioning on I. A new, fourth criterion is added that expresses that the distribution of I follows the multinomial distribution with parameter α. That the probability distribution $\mathbb{P}$ with the above properties exists is guaranteed again by the Ianescu-Tulcea theorem. As usual, when needed, we use $\mathbb{P}_\mu$ to indicate the dependence of $\mathbb{P}$ on μ.

Finally some notation: For a probability measure $\mathbb{P}$ on a measurable space $(\Omega, \mathcal{F})$ and a sub-sigma algebra $\mathcal{G}$ of $\mathcal{F}$, let $\mathbb{P}|_\mathcal{G}$ be the probability measure on $(\Omega, \mathcal{G})$ obtained from $\mathbb{P}$ by restricting it to $\mathcal{G}$: $\mathbb{P}|_\mathcal{G}(U) = \mathbb{P}(U)$ for any $U \in \mathcal{G}$.

Question 1. Unless otherwise specified let $\pi^{(1)}, \dots, \pi^{(k)}$ be arbitrary policies of M and let $\alpha \in \mathcal{M}_1([k])$, $\mu \in \mathcal{M}_1(\mathcal{S})$ be also arbitrary. Also, let $(\Omega, \mathcal{F}, \mathbb{P})$ as above (we shall also use $\mathbb{P}_\mu$ when the dependence on μ is important). Let $Z = (S_0, A_0, S_1, A_1, \dots)$. Show that the following hold:

1. Z is random element between $(\Omega, \mathcal{F})$ and $((\mathcal{S} \times \mathcal{A})^{\mathbb{N}}, \mathcal{G}')$ where $\mathcal{G}'$ is the product σ-algebra on $(\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$ induced by the discrete topology on $\mathcal{S} \times \mathcal{A}$.

   5 points

2. Show that there is a policy $\bar{\pi}$ of the MDP M such that for any $\mu \in \mathcal{M}_1(\mathcal{S})$, the pushforward of $\mathbb{P}_\mu$ under Z, $(\mathbb{P}_\mu)_Z$ satisfies
   $$(\mathbb{P}_\mu)_Z = \mathbb{P}_\mu^{\bar{\pi}}$$
   where $\mathbb{P}_\mu^{\bar{\pi}}$ is the unique probability measure on the canonical space $((\mathcal{S} \times \mathcal{A})^{\mathbb{N}}, \mathcal{G}')$ induced by the interconnection of $\bar{\pi}$ and the MDP, given the initial state distribution μ. That is, a mixture policy induces a policy $\bar{\pi}$ of the MDP M.

   20 points

3. Let $R = \sum_{t=0}^{\infty} \gamma^t r_{A_t}(S_t)$ and let $\mathbb{P}$ be as above with the choice $\mu = \delta_s$. Let $\mathbb{E}$ be the expectation operator corresponding to $\mathbb{P}$. Show that $v(s) = \mathbb{E}[R]$ is well-defined: That is, for any $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\Omega, \mathcal{F}, \mathbb{P}')$ as long as $\mathbb{P}$ and $\mathbb{P}'$ satisfy the above four properties, $\mathbb{E}[R] = \mathbb{E}'[R]$ where $\mathbb{E}'$ is the expectation operator underlying $\mathbb{P}'$.

   10 points

4. Show that $v(s) = v^{\bar{\pi}}(s)$.

   5 points

5. Let $\mathbb{P}_\mu^{\pi^{(i)}}$ ($\mathbb{P}_\mu^{\bar{\pi}}$) be the probability measures induced on the canonical space $((\mathcal{S} \times \mathcal{A})^{\mathbb{N}}, \mathcal{G}')$ by the initial state distribution μ and the interconnection of $\pi^{(i)}$ (respectively, $\bar{\pi}$) with the MDP M. Show that $\mathbb{P}_\mu^{\bar{\pi}} = \sum_{i=1}^{k} \alpha_i \mathbb{P}_\mu^{\pi^{(i)}}$.

   10 points

6. Mixing is guaranteed to keep performance bounds: if for some $v : \mathcal{S} \to \mathbb{R}$ and for all $i \in [k]$, $v^{\pi^{(i)}} \geq v$ then $v^{\bar{\pi}} \geq v$.

   5 points

7. Averaging is not guaranteed to keep performance bounds: For any $\gamma > 1/2$ there exists an MDP with state space $\mathcal{S}$, $k \geq 2$, policies $\pi_1, \dots, \pi_k$, a function $v : \mathcal{S} \to \mathbb{R}$ and $\alpha \in \mathcal{M}_1([k])$ such that $v^{\pi_i} \geq v$ holds for all $i \in [k]$, yet if π is the α-average of $\pi_1, \dots, \pi_k$ then $v^{\pi} < v$.

   10 points

Hint: Recall the change-of-variables formula: For a random element X taking values in some measurable set $\mathcal{X}$, the pushforward $\mathbb{P}_X$ of X satisfies

$$\mathbb{E}[f(X)] = \int f(x)\mathbb{P}_X(dx)\,.$$

Recall also that integration is linear in measures. In particular, for any measures $\mathbb{P}_i$ and nonnegative coefficients α_i, $i \in [k]$ and f which is $(\sum_{i=1}^k \alpha\mathbb{P}_i)$-integrable, $\int f d(\sum_{i=1}^k \alpha\mathbb{P}_i) = \sum_{i=1}^k \alpha_i \int f d\mathbb{P}_i$ (this also extends to signed measures, but we won't need this extension).

Total: **65 points**

Solution. Let $s, a, s_0, a_0, s_1, a_1, \dots$ be an arbitrary sequence of state-actions pairs.

1. We need to check that for $U \in \mathcal{G}'$, $Z^{-1}(U) \in \mathcal{F}$. Since $\mathcal{G}'$ is a product σ-algebra, it suffices to check this for the "simple" cylinder sets, i.e., when U is either of the form

$$C = \{s_0\} \times \{a_0\} \times \{s_1\} \dots \{s_t\} \times \Omega\,, \quad \text{or, of the form}$$
$$C' = \{s_0\} \times \{a_0\} \times \{s_1\} \dots \{s_t\} \times \{a_t\} \times \Omega\,.$$

   For the first case, $Z^{-1}(C) = \{S_0 = s_0, A_0 = a_0, S_1 = s_1, \dots, S_t = s_t\}$, which is in $\mathcal{F}$ because $S_0, \dots, S_t$ and $A_0, \dots, A_{t-1}$ are $\mathcal{F}$-measurable. The same holds for the second case for identical reasons, just add that A_t is also $\mathcal{F}$-measurable. In this case, $Z^{-1}(C') = \{S_0 = s_0, A_0 = a_0, S_1 = s_1, \dots, S_t = s_t, A_t = a_t\}$.

2. Fix μ and let $\mathbb{P} = \mathbb{P}_\mu$. We show that $\mathbb{P}$ satisfies the criteria that define the probability measure $\mathbb{P}_\mu^{\bar\pi}$ with a suitable policy $\bar\pi$. It follows that $\mathbb{P}_Z$ also satisfies these criteria (because the criteria are concerned with events in $\sigma(Z)$). Hence, $\mathbb{P}_Z = \mathbb{P}_\mu^{\bar\pi}$ follows by the uniqueness of the canonical probability space. Fix any $t \geq 0$. For the first criterion, by the tower rule,

$$\mathbb{P}(S_0 = s) = \mathbb{E}[\mathbb{P}(S_0 = s|I)] = \mathbb{E}[\mu(s)] = \mu(s)\,.$$

   The second criterion will be verified by defining $\bar\pi_t$ as

$$\bar\pi_t(a|h_t) = \begin{cases} \mathbb{P}(A_t = a|H_t = h_t)\,, & \text{if } \mathbb{P}(H_t = h_t) > 0\,; \\ \pi_0(a)\,, & \text{otherwise}\,, \end{cases}$$

   where $h_t = (s_0, a_0, \dots, a_{t-1}, s_t)$ is arbitrary and π_0 is an arbitrary distribution over the actions. This indeed defines a policy: $\bar\pi = (\bar\pi_t)$; $\bar\pi_t$ maps histories to distributions. Indeed, this is clear when $\mathbb{P}(H_t = h_t) = 0$. Otherwise,

$$\begin{aligned} &\sum_{a\in\mathcal{A}} \bar\pi_t(a|h_t) \\ &\qquad = \sum_{a\in\mathcal{A}} \mathbb{P}(A_t = a|H_t = h_t) \\ &\qquad = \mathbb{P}(A_t \in \mathcal{A}|H_t = h_t) = 1\,. \end{aligned}$$

   We now claim that $\bar\pi_t$ is independent of μ ($\mathbb{P}$ hides its dependence on μ). Again, this is clear when $\mathbb{P}(H_t = h_t) = 0$ since π_0 does not depend on μ. When $\mathbb{P}(H_t = h_t) > 0$ we have

$$\begin{aligned} \bar\pi_t(a|h_t) &= \mathbb{P}(A_t = a|H_t = h_t) \\ &= \sum_i \mathbb{P}(A_t = a|H_t = h_t, I = i)\mathbb{P}(I = i|H_t = h_t) = \sum_i \pi_t^{(i)}(a|h_t)\mathbb{P}(I = i|H_t = h_t)\,, \end{aligned}$$

where the last equality follows because if $\mathbb{P}(H_t = h_t, I = i) = 0$ then, by definition, $\mathbb{P}(I = i|H_t = h_t) = 0$, and hence $\mathbb{P}(A_t = a|H_t = h_t, I = i)\mathbb{P}(I = i|H_t = h_t) = 0 = \pi_t^{(i)}(a|h_t)\mathbb{P}(I = i|H_t = h_t)$.

It remains to show that $\mathbb{P}(I = i|H_t = h_t)$ does not depend on μ. Again, this is clear when $\mathbb{P}(H_t = h_t) = 0$ since in this case $\mathbb{P}(I = i|H_t = h_t) = 0$. For the case when $\mathbb{P}(I = i|H_t = h_t) > 0$, we have

$$\mathbb{P}(I = i|H_t = h_t) = \frac{\mathbb{P}(H_t = h_t, I = i)}{\mathbb{P}(H_t = h_t)}\,.$$

Based on the properties of $\mathbb{P}$, with repeated conditioning, we calculate,

$$\begin{aligned}\mathbb{P}(H_t = h_t, I = i) = \alpha_i \mu(s_0)\, \pi_0^{(i)}(a_0|s_0)\pi_1^{(i)}(a_1|s_0, a_0, s_1)\dots\pi_{t-1}^{(i)}(a_{t-1}|s_0, a_0, \dots, s_{t-1}) \times \\ P_{a_0}(s_0, s_1)\dots P_{a_{t-1}}(s_{t-1}, s_t)\,.\end{aligned} \tag{1}$$

Hence,

$$\mathbb{P}(I = i|H_t = h_t) =$$

$$\frac{\pi_0^{(i)}(a_0|s_0)\pi_1^{(i)}(a_1|s_0, a_0, s_1)\dots\pi_{t-1}^{(i)}(a_{t-1}|s_0, a_0, \dots, s_{t-1})\ \cancel{\mu(s_0)P_{a_0}(s_0, s_1)\dots P_{a_{t-1}}(s_{t-1}, s_t)}}{\sum_i \alpha_i \pi_0^{(i)}(a_0|s_0)\pi_1^{(i)}(a_1|s_0, a_0, s_1)\dots\pi_{t-1}^{(i)}(a_{t-1}|s_0, a_0, \dots, s_{t-1})\ \cancel{\mu(s_0)P_{a_0}(s_0, s_1)\dots P_{a_{t-1}}(s_{t-1}, s_t)}}\,,$$

which is independent of μ as required.

For the third criterion, we have

$$\mathbb{P}(S_{t+1} = s|H_t, A_t) = \mathbb{E}[\mathbb{P}(S_{t+1} = s|H_t, A_t, I)|H_t, A_t] = \mathbb{E}[P_{A_t}(S_t, s)|H_t, A_t] = P_{A_t}(S_t, s)\,,$$

where the first equality uses the tower rule, the second uses Property 2 of $\mathbb{P}$, the third uses that $P_{A_t}(S_t, s)$ is a constant given H_t, A_t, hence it can be moved outside of the expectation (formally, $P_{A_t}(S_t, s)$ is $\sigma(H_t \times A_t)$ measurable). Hence $\mathbb{P}$ satisfies the three criteria of measures induced by the interconnection of $\bar{\pi}$, the MDP M and the initial distribution μ, finishing the proof.

3. Noting that $R = f(Z)$ where f is defined via $f(s_0, a_0, s_1, a_1, \dots) = \sum_{t=0}^{\infty} \gamma^t r_{a_t}(s_t)$ is a measurable function from $((\mathcal{S} \times \mathcal{A})^{\mathbb{N}}, \mathcal{G}')$ to $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, it suffices to show that $\mathbb{P}_Z = \mathbb{P}'_Z$ because then, by the change-of-variables-formula,

$$\mathbb{E}[R] = \mathbb{E}[f(Z)] = \int f(z)\mathbb{P}_Z(dz) = \int f(z)\mathbb{P}'_Z(dz) = \mathbb{E}'[f(Z)] = \mathbb{E}'[R]\,.$$

Now, for $U \in \mathcal{G}'$ we have

$$\mathbb{P}_Z(U) = \mathbb{P}(Z \in U) = \mathbb{P}'(Z \in U) = \mathbb{P}'_Z(U)\,,$$

where the second equality follows because, as it can be easily seen, equality here holds for all simple cylinder sets U, hence $\mathbb{P}_Z = \mathbb{P}'_Z$ also holds and the proof is finished.

4. By Part 2, $\mathbb{P}_Z = \mathbb{P}_s^{\bar{\pi}}$. Then, with f as above,

$$v(s) = \int f(z)\mathbb{P}_Z(dz) = \int f(z)\mathbb{P}_s^{\bar{\pi}}(dz) = v^{\bar{\pi}}(s)\,.$$

5. By Eq. (1) and the construction of $\mathbb{P}_\mu^{\pi^{(i)}}$,

$$\begin{aligned}\mathbb{P}(H_t = h_t, I = i) &= \alpha_i \mu(s_0) \prod_{j=0}^{t-1} \pi_j^{(i)}(a_j|s_0, a_0, \dots, s_j) \prod_{j=0}^{t-1} P_{a_j}(s_j, s_{j+1}) \\ &= \alpha_i \mathbb{P}_\mu^{\pi^{(i)}}(H_t = h_t)\,,\end{aligned}$$

and, similarly,

$$\mathbb{P}(H_t = h_t, A_t = a_t, I = i) = \alpha_i \mathbb{P}_\mu^{\pi^{(i)}}(H_t = h_t, A_t = a_t)\,.$$

Summing these up for $i \in [k]$, we get

$$\mathbb{P}(H_t = h_t) = \sum_{i=1}^k \alpha_i \mathbb{P}_\mu^{\pi^{(i)}}(H_t = h_t)\,,$$

$$\mathbb{P}(H_t = h_t, A_t = a_t) = \sum_{i=1}^k \alpha_i \mathbb{P}_\mu^{\pi^{(i)}}(H_t = h_t, A_t = a_t)\,.$$

Since h_t, a_t are arbitrary, $\mathbb{P}_Z = \sum_{i=1}^k \alpha_i \mathbb{P}_\mu^{\pi^{(i)}}$ (again, verifying this for simple cylinder sets). By Part 2, $\mathbb{P}_Z = \mathbb{P}_\mu^{\bar{\pi}}$. Putting things together, we get $\mathbb{P}_\mu^{\bar{\pi}} = \sum_{i=1}^k \alpha_i \mathbb{P}_\mu^{\pi^{(i)}}$.

6. With f as in the previous parts,

$$v^{\bar{\pi}}(s) = \int f(z)\mathbb{P}_s^{\bar{\pi}}(dz) = \sum_i \alpha_i \int f(z)\mathbb{P}_s^{\pi^{(i)}}(dz) = \sum_i \alpha_i v^{\pi^{(i)}}(s)\,.$$

Hence, if $v^{\pi^{(i)}} \geq v$ then multiplying both sides by $\alpha_i \geq 0$, integrating with respect to $\mathbb{P}_s^{\pi^{(i)}}$ and summing up we get $v^{\bar{\pi}} \geq v$.

7. It is enough to consider a 2-state, 2-action MDP with $\mathcal{S} = \mathcal{A} = [2]$ such that action $i \in [2]$ sets the next state to i (deterministically). Further, make staying at any of the states incur a reward of 1, while make transitioning between the states incur a reward of zero. Choose $k = 2$. Policy π_i uses action i (moving to state i) everywhere. The value of both π_1 and π_2 is above $\gamma/(1-\gamma)$. The uniform average chooses the actions at random at both states. The value of the averaged policy π at both states is $\frac{1}{2(1-\gamma)}$, which is lower than $\gamma/(1-\gamma)$ provided that $\gamma > 1/2$.

□

Finding needles with high probability

The high-probability needle lemma is as follows:

Lemma 1 (High-probability needle lemma)**.** *Any algorithm that correctly identifies the single nonzero entry in any binary array of length k with probability at least* 0.91 *has the property that on some input the expected number of queries that the algorithm uses is at least $\Omega(k)$.*

Question 2. Prove Lemma 1. Note that the algorithms are allowed to randomize.

Total: **30 points**

Solution. We give two solutions, each of which have their own merits. The idea of the first solution is rather simple: by repeatedly running it, any algorithm that is correct with positive probability can be turned into an algorithm which is always correct at the expense of only increasing the runtime inversely proportionally to the success probability. However, the formal argument relies on familiarity with Wald's identity. In contrast, the second solution is direct and elementary, but it is special to the problem at hand.

Solution 1: In what follows we will identify the possible inputs over k element arrays with the integers $i \in [k]$. We prove a stronger claim that for any algorithm that returns solutions that are correct with at least probability p, for any $k \geq 2$, if $q_{k.i}$ is the runtime of algorithm when it is used on input $i \in [k]$,

$$\max_{i \in [k]} q_{k,i} \geq p\left(\frac{k+1}{2} - \frac{1}{k}\right) - 1\,,$$

Fix $k \geq 2$. Fix any algorithm A. This algorithm gives rise to an algorithm A' that knows when it is correct and A' uses at most one extra query compared to A: When A stops and chooses item I, at the expense of at most one extra query, A' can verify whether $I = i$. Thus, A' will know whether it was successful and not. Since the number of queries issued by A is at best one less than that of A', it suffices to show that A' uses $\Omega(k)$ queries on inputs of length k. Hence, in what follows, we restrict ourselves to algorithm that also output an indicator of their own success.

Let $Q \in \{0, 1, \dots\}$ denote the random number of queries used and let $S \in \{0, 1\}$ be the indicator whether A finds the nonzero entry in its input. As agreed, we may assume that S is the output of A. On input $i \in [k]$, algorithm A induces some distribution $P_{k,i} \in \mathcal{M}_1(\{0, 1, \dots\} \times \{0, 1\})$ over these pairs. Let $q_{k,i}$ be the expected number of queries used by A on input i. Further, by assumption, $p_{k,i}$, the probability that algorithm A succeeds on input i is at least p:

$$p_{k,i} \geq p\,. \tag{2}$$

Let $\mathbb{P}_{k,i}$ be the probability distribution over interaction sequences of infinitely many independent runs of A on input i. Define A'' as the algorithm that runs A (every time freshly initialized) until A succeeds when it returns the item returned by A on it last call. Clearly, when A'' stops it finds the correct item. We claim the following: Let $i \in [k]$ be arbitrary.

1. If N is the number of times A'' runs A, $\mathbb{P}_{k,i}(N < \infty) = 1$, that is, A'' stops with probability one;
2. Letting Q be the number of queries used by A'',

$$\mathbb{E}_{k,i}[Q] = \mathbb{E}_{k,i}[N] q_{k,i} \leq \frac{q_{k,i}}{p}\,. \tag{3}$$

If the above two claims are established, it follows that A is a randomized algorithms which always finds the correct entry. Thus, by the first problem on homework 0, for some $i \in [k]$,

$$\frac{k+1}{2} - \frac{1}{k} \leq \mathbb{E}_{k,i}[Q]\,.$$

Putting this together with (3) gives $p(\frac{k+1}{2} - \frac{1}{k}) \leq q_{k,i}$. Thus,

$$\max_{i \in [k]} q_{k,i} \geq p\left(\frac{k+1}{2} - \frac{1}{k}\right)\,,$$

finishing the proof.

It remains to establish the above two claims. Fixing k, i allows us to reduce clutter by writing $\mathbb{E}$ in place of $\mathbb{E}_{k,i}$ and $\mathbb{P}$ in place of $\mathbb{P}_{k,i}$.

To prove the claims, introduce (Q_t, S_t) as the pair where Q_t is the number of queries used in call $t \geq 1$ of algorithm A and where $S_t \in \{0, 1\}$ indicates whether this call was successful. By construction, $((Q_t, S_t))_{t \geq 1}$ is an i.i.d. sequence, with common distribution $P_{k,i}$. Also, by definition,

$$N = \min\{n \geq 1 \,:\, S_n = 1\}\,.$$

As is well known, N has a geometric distribution with parameter $p_{k,i}$: $\mathbb{P}(N = n) = p_{k,i}(1 - p_{k,i})^{n-1}$ and $\mathbb{P}(N \geq n) = (1 - p_{k,i})^{n-1}$. As $\mathbb{P}(N < \infty) = 1 - \lim_{n \to \infty} \mathbb{P}(N \geq n) = 1$, establishing the first claim.

As to the second claim, note that by definition,

$$Q = \sum_{n=1}^{N} Q_n\,.$$

We intend to use Wald's identity to get our desired result. To be able to use this identity, we need to check that the following are satisfied:

1. $(Q_n)_{n\geq 1}$ share the same finite-mean;
2. $\mathbb{E}[N] < \infty$;
3. $\mathbb{E}[Q_n \mathbb{I}\{N \geq n\}] = \mathbb{E}[Q_n]\mathbb{P}(N \geq n)$ for all $n \geq 1$;
4. $\sum_{n=1}^{\infty} \mathbb{E}[|Q_n| \mathbb{I}\{N \geq n\}] < \infty$.

If these conditions hold, Wald's identity gives

$$\mathbb{E}[Q] = \mathbb{E}[N]\mathbb{E}[Q_1]\,.$$

Then, using that $\mathbb{E}[Q_1] = q_{k,i}$ and that, as is well known,

$$\mathbb{E}[N] = \sum_{n\geq 1} \mathbb{P}(N \geq n) = \frac{1}{p_{k,i}}\,, \tag{4}$$

combined with (2) gives

$$\mathbb{E}[Q] \leq \frac{q_{k,i}}{p}$$

as required.

It remains to verify the stated conditions. The first condition follows from the definitions (the common mean is $q_{k,i}$). For the second condition, we already noted that $\mathbb{E}[N] = 1/p_{k,i}$ which is finite. For the third condition, note that $\{N \geq n\} = \{S_1 = 0, \ldots, S_{n-1} = 0\}$ whose indicator is independent of Q_n (since Q_n and $(S_1, \ldots, S_{n-1})$ are independent). Hence,

$$\mathbb{E}[Q_n \mathbb{I}\{N \geq n\}] = \mathbb{E}[Q_n \mathbb{I}\{S_1 = 0, \ldots, S_{n-1} = 0\}] = \mathbb{E}[Q_n]\mathbb{E}[\mathbb{I}\{S_1 = 0, \ldots, S_{n-1} = 0\}] = \mathbb{E}[Q_n]\mathbb{P}(N \geq n)\,,$$

as required. The fourth condition follows from the third: $\sum_{n\geq 1} \mathbb{E}[|Q_n| \mathbb{I}\{N \geq n\}] = \sum_{n\geq 1} \mathbb{E}[Q_n \mathbb{I}\{N \geq n\}] = \sum_{n\geq 1} \mathbb{E}[Q_n]\mathbb{P}(N \geq n) = q_{k,i}\mathbb{E}[N] < \infty$.

Solution 2: Let $\mathrm{Perm}([k])$ denote the permutations on $[k]$. WLOG we may restrict ourselves to randomized algorithms that query the entries in a random order, say $P \in \mathrm{Perm}([k])$, querying first $P(1)$, then $P(2)$, etc. Indeed, as argued in homework 0, algorithms that query entries twice or more, are dominated. Similarly, we may assume that the algorithm stops whenever it receives 1 as the response or when it queried $k-1$ entries. In general, an algorithm may also decide to stop after $M \in [k-1]$ queries were issued: In this case, again, WLOG, we may assume that it outputs a random element R from the entries not yet queried: $R \in \{P(M+1), \ldots, P(k)\}$. Thus, an arbitrary, non-dominated randomizing algorithm is fully described by the joint distribution of (P, M, R).

Fix now such an algorithm. Let C be the output (entry returned by the algorithm). Further, let Q be the number of queries the algorithm uses. Thus, on instance $i \in [k]$, $C = i$ if $P^{-1}(i) \leq M$, otherwise $C = R$. (Note that $P^{-1}(i) \leq M$ is equivalent to $i \in \{P(1), \ldots, P(M)\}$.) Further, on instance i, $Q = \min(P^{-1}(i), M)$. Let $I \in [k]$ be a random index that is uniformly chosen, independently of the choice of (P, M, R).

Let $\mathbb{P}_i$ be the probability distribution induced on (C, Q, I) by running algorithm on instance i. Further, let $\mathbb{P}$ be the probability distribution induced on (C, Q, I) by running the algorithm on a random index $I \in [k]$ with a uniform distribution. As $\mathbb{P}_i(\cdot) = \mathbb{P}(\cdot | I = i)$ and I is uniformly distributed, $\mathbb{P} = \frac{1}{k}\sum_{i=1}^{k} \mathbb{P}_i$. We denote by $\mathbb{E}_i$ the expectation operator underlying $\mathbb{P}_i$, and by $\mathbb{E}$ the expectation operator underlying $\mathbb{P}$.

Assume that the expected query cost of the algorithm is "small":

$$\max_i \mathbb{E}_i[Q] \leq ck$$

for $c > 0$ to be chosen later, while the algorithm is guaranteed to return the correct answer with "high probability":

$$\min_i \mathbb{P}_i(C = i) \geq 0.91\,.$$

Fix $i \in [k]$. By Markov's inequality,

$$\mathbb{P}_i(Q > 100ck) \leq \frac{\mathbb{E}_i[Q]}{100ck} \leq \frac{1}{100}\,.$$

Hence,

$$\mathbb{P}_i(C = i, Q \leq 100ck) \geq \mathbb{P}_i(C = i) - \mathbb{P}_i(Q > 100ck) \geq 0.91 - 0.01 = 0.9\,.$$

Taking the average over $i = 1, \ldots, k$, it follows that

$$\mathbb{P}(C = I, Q \leq 100ck) \geq 0.9\,.$$

By the tower rule, $\mathbb{P}(C = I, Q \leq 100ck) = \mathbb{E}[\mathbb{P}(C = I, Q \leq 100ck|P, M, R)] \geq 0.9$, from which it follows that for some $p \in \mathrm{Perm}([k])$, $m \in [k-1]$, $r \in [k]$ with

$$r \in \{p(m+1), \ldots, p(k)\}\,,$$

it holds that

$$\mathbb{P}(C = I, Q \leq 100ck|P = p, M = m, R = r) \geq 0.9\,.$$

Now,

$$\begin{aligned}
&\mathbb{P}(C = I, Q \leq 100ck|P = p, M = m, R = r) \\
&\quad \leq \mathbb{P}(p^{-1}(I) \leq 100ck|P = p, M = m, R = r) + \mathbb{P}(p^{-1}(I) > 100ck, C = I, Q \leq 100ck|P = p, M = m, R = r) \\
&\quad \leq 100c + \mathbb{P}(p^{-1}(I) > 100ck, C = I, Q \leq 100ck|P = p, M = m, R = r)\,,
\end{aligned}$$

where the second inequality used that I and P, M, R are independent and that $\lceil 100ck \rceil \leq 100ck$. Considering the last term note that if $p^{-1}(I) > 100ck \geq Q$ then $Q = \min(p^{-1}(I), m) = m$ and thus $C = r$. Thus,

$$\begin{aligned}
&\mathbb{P}(p^{-1}(I) > 100ck, C = I, Q \leq 100ck|P = p, M = m, R = r) \\
&\quad \leq \mathbb{P}(p^{-1}(I) > 100ck, I = r|P = p, M = m, R = r) \leq \frac{k - \lceil 100ck + 1 \rceil}{k} \leq \frac{(k - 100ck)}{k} = 1 - 100c\,,
\end{aligned}$$

where we used again the independence of I and P, M, R. Choosing $c = 0.002$ we see that

$$0.9 \leq \mathbb{P}(C = I, Q \leq 100ck|P = p, M = m, R = r) \leq 0.8\,,$$

which is a contradiction. Hence, with this choice of c there is no algorithm with the above two properties. □

Fitted Value Iteration

Assume that the rewards belong to the $[0, 1]$ interval and fix the discount factor γ. Let $H_\gamma = 1/(1-\gamma)$. Assume we are given a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ which spans $\mathbb{R}^d$. Let $\mathcal{F} = \{f_\theta : f_\theta(s, a) = \phi(s, a)^\top \theta, \theta \in \mathbb{R}^d\}$ be the span of the features. Let $C \subset \mathcal{Z} := \mathcal{S} \times \mathcal{A}$ be the set whose existence is guaranteed by the Kiefer-Wolfowitz theorem for the feature map ϕ and let $\rho : C \to [0, 1]$ be the corresponding weighting function. In particular, $|C| \leq d(d+1)/2$, $\sum_{z \in C} \rho(z) = 1$ and with $G_\rho = \sum_{z \in C} \rho(z)\phi(z)\phi(z)^\top$, $\max_{z \in \mathcal{Z}} \|\phi(z)\|_{G_\rho^{-1}} \leq \sqrt{d}$.

For $k \geq 1$, $(s,a) \in \mathcal{S} \times \mathcal{A}$, let $C_k(s,a) = [S'_1(k,s,a), \ldots, S'_m(k,s,a)]$ be so that all the $(C_k(s,a))_{k,s,a}$ are independent of each other, and for any k, s, a, $S'_1(k,s,a), \ldots, S'_m(k,s,a) \overset{\text{iid}}{\sim} P_a(s)$. For $k \geq 1$ let $\hat{T}_k : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}$ be defined by

$$(\hat{T}_k q)(s,a) = r_a(s) + \frac{\gamma}{m} \sum_{s' \in C_k(s,a)} M q(s') .$$

Further, let $\Pi : \mathbb{R}^{\mathcal{Z}} \to \mathbb{R}^{\mathcal{Z}}$ be defined by $(\Pi f)(z) = \max(\min(f(z), H_\gamma), 0)$: In words, Π truncates the values of its argument to the $[0, H_\gamma]$ interval.

Consider the following procedure, which we call fitted q iteration (FQI).[1]

1. $\theta_0 = \mathbf{0}$
2. `for` $k = 1, 2, \ldots, K$ `do`
3. $\theta_k = \text{argmin}_{\theta \in \mathbb{R}^d} \sum_{z \in C} \rho(z)(f_\theta(z) - (\hat{T}_k \Pi f_{\theta_{k-1}})(z))^2$
4. `return` θ_K

Let $\varepsilon_{\text{apx}} = \sup_\theta \inf_{\theta'} \|f_{\theta'} - T\Pi f_\theta\|_\infty$.

Question 3. Prove that the following hold:

1. The computation cost of FQI is $O(Kd^3mA)$ and it needs $O(d^2)$ space (all in the RAM model of computation). The query cost is $O(Kd^2m)$. Explain how you get the bounds.

5 points

2. Fix $k \geq 0$. Let $q_k = \Pi f_{\theta_k}$. For $k > 0$, let $\epsilon_k : \mathcal{Z} \to R$ and $\theta^*_k \in \mathbb{R}^d$ be such that $Tq_{k-1} = f_{\theta^*_k} + \epsilon_k$ and $\|\epsilon_k\|_\infty \leq \varepsilon_{\text{apx}}$. Show that ϵ_k and θ^*_k are well-defined (i.e., they exist).

10 points

3. Show that for any $k \geq 1$, $0 \leq \zeta \leq 1$, with probability at least $1 - \zeta$,

$$\|q_k - Tq_{k-1}\|_\infty \leq (1 + \sqrt{d})\varepsilon_{\text{apx}} + \sqrt{d} H_\gamma \sqrt{\frac{\log\left(\frac{2|C|}{\zeta}\right)}{2m}} .$$

10 points

4. Show that, on the same event as in the previous part, the policy π that is greedy with respect to q_K is δ-optimal with

$$\delta \leq 2H_\gamma^2 \left\{ (1 + \sqrt{d})\varepsilon_{\text{apx}} + \gamma^K + \sqrt{d} H_\gamma \sqrt{\frac{\log\left(\frac{2|C|K}{\zeta}\right)}{2m}} \right\} .$$

10 points

[1] A terrible name.

5. Fix $\epsilon > 0$. Argue that K, m and ζ can be chosen as a polynomial function of $H_\gamma, d, 1/\epsilon$ so that the expected suboptimality of the policy π is bounded by $2H_\gamma^2(1+\sqrt{d})\varepsilon_{\text{apx}} + 2\epsilon$. Show the choices you made.

5 points

6. Argue that with a query, runtime and space cost that is polynomial in $H_\gamma, d, 1/\epsilon, A$, the procedure obtains a policy π that is at most δ-optimal with $\delta = 2H_\gamma^2(1+\sqrt{d})\varepsilon_{\text{apx}} + 2\epsilon$.

5 points

7. The MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ is called linear in ϕ if it holds that with some $\theta_r \in \mathbb{R}^d$, $r_a(s) = f_{\theta_r}(s,a)$ holds for all (s,a) and if for some $\mu : \mathcal{S} \to \mathbb{R}^d$, for any (s,a), $P_a(s,s') = \langle \phi(s,a), \mu(s') \rangle$. Show that if $\mathcal{M}$ is linear in ϕ then $\varepsilon_{\text{apx}} = 0$.

10 points

Total: **55 points**

Solution.

1. Note that

$$\theta_k = G_\rho^{-1} \underbrace{\sum_{z \in C} \rho(z)\phi(z)Y_k(z)}_{=:B_k}, \tag{5}$$

where $Y_k(z) = \hat{T}_k \Pi f_{\theta_{k-1}}(z)$. For z fixed, $Y_k(z)$ can be computed in $O(mAd)$ steps. All $Y_k(\cdot)$ is computed in $O(|C|mAd) = O(d^3 mA)$ steps. Given these $O(d^2)$ values, $B_k \in \mathbb{R}^d$ can be calculated in time $O(|C|d) = O(d^3)$ and thus the total cost of calculating B_k is $O(d^3 mA)$. The matrix inverse G_ρ^{-1} needs only to be computed once, at the cost of, say $O(d^3)$. The cost of matrix vector multiplication is $O(d^2)$. The total cost of calculating θ_k is dominated by $O(d^3 mA)$. Multiply this by K to get the total cost of the procedure.

For storage, one can invert a matrix in place. Besides the matrix G_ρ^{-1}, one needs to store only d-dimensional vectors. Hence, the storage cost is $O(d^2)$.

The query complexity of calculating comes from the need to access $C_k(z)$ for $z \in C$. Hence, the query cost is $O(d^2 m)$. Multiply this by K to get the total number of queries.

2. Choose θ_k^* as the minimizer of $\theta \mapsto g(\theta) := \|Tq_{k-1} - f_\theta\|_\infty$. We argue that this exists. Indeed, g is continuous and nonnegative. Hence, there exists a sequence $(\theta_i)_i$ such that $g(\theta_i) \to \inf_\theta g(\theta)$. Note that G_ρ is full rank because ϕ spans $\mathbb{R}^d$. There are two cases: Either $\sup_i \|\theta_i\|_{G_\rho}$ is finite, or it is infinite. If it is finite, by the completeness of $\mathbb{R}^d$, a subsequence of θ_i converges to a minimizer of g by the continuity of g. In the opposite case, from $\|\theta_i\|_{G_\rho}^2 = \sum_{z \in C} \rho(z) f_{\theta_i}^2(z)$ we see that, $(f_{\theta_i}^2(z))_i$ must be unbounded for at least one $z \in C$. Hence, for this z,

$$g(\theta_i) = \|f_{\theta_i} - Tq_{k-1}\|_\infty \geq |f_{\theta_i}(z)| - |Tq_{k-1}(z)| \,.$$

Hence,

$$\limsup_{i \to \infty} g(\theta_i) \geq \limsup_{i \to \infty} |f_{\theta_i}(z)| - |Tq_{k-1}(z)| = \infty \,,$$

which contradict to that $\limsup_{i\to\infty} g(\theta_i) = \inf_\theta g(\theta) \leq g(0) < \infty$.

3. We have

$$\begin{aligned}\|q_k - Tq_{k-1}\|_\infty &= \|\Pi f_{\theta_k} - Tq_{k-1}\|_\infty \\ &\le \|f_{\theta_k} - (f_{\theta_k^*} + \epsilon_k)\|_\infty \\ &\le \|f_{\theta_k} - f_{\theta_k^*}\|_\infty + \varepsilon_{\mathrm{apx}} \\ &\le \|f_{\theta_k} - f_{\theta_k^*}\|_\infty + \varepsilon_{\mathrm{apx}} \\ &\le \sqrt{d} \max_{z \in C} |\epsilon(z)| + \varepsilon_{\mathrm{apx}} \,,\end{aligned}$$

where the first equality uses the definition of q_k, the next inequality uses that $Tq_{k-1} \in [0, 1/(1-\gamma)]$, hence dropping the truncation can only increase the values (at the same place we also used the definition of θ_k^* and ϵ_k). The next inequality uses the triangle inequality and that by definition $\|\epsilon_k\|_\infty \le \varepsilon_{\mathrm{apx}}$, and for the last inequality we use an appropriately defined function $\epsilon : C \to \mathbb{R}$. For the definition recall the corollary of Lecture 8 that states that $\|f_{\hat\theta} - f_\theta\|_\infty \le \sqrt{d} \max_{z \in C} |\epsilon(z)|$ holds for $\hat\theta = G_\rho^{-1} \sum_{z \in Z} \rho(z)\phi(z)(f_\theta(z) + \epsilon(z))$. Now, recall the definition of θ_k from (5). Writing

$$Y_k(z) = (Tq_{k-1})(z) + \hat\epsilon(z) = f_{\theta_k^*}(z) + \hat\epsilon(z) + \epsilon_k(z) \,,$$

where the first equality defines $\hat\epsilon(z)$, we see that above we can use $\epsilon(z) = \hat\epsilon(z) + \epsilon_k(z)$. Now, note from Hoeffding's inequality that with probability $1 - \zeta$,

$$\max_{z \in C} |\hat\epsilon(z)| \le H_\gamma \sqrt{\frac{\log\left(\frac{2|C|}{\zeta}\right)}{2m}} \,,$$

where we use that $\mathbb{E}[Y_k(z)|q_{k-1}] = (Tq_{k-1})(z)$ and that $S_1'(k,z), \ldots, S_m'(k,z)$ are independent given q_{k-1}, hence, $(Mq_{k-1}(S_j'(k,z))_j$ is an i.i.d. sequence, and it takes values in the interval $[0, H_\gamma]$. We also have

Cs: this independence is not quite well explained.

$$|\epsilon(z)| \le \varepsilon_{\mathrm{apx}} + H_\gamma \sqrt{\frac{\log\left(\frac{2|C|}{\zeta}\right)}{2m}} \,,$$

Putting everything together gives the desired claim.

4. Let $\delta_k = \|q_k - q^*\|_\infty$. For $k > 0$ we have

$$\delta_k \le \|q_k - Tq_{k-1}\|_\infty + \|Tq_{k-1} - Tq^*\|_\infty \le \|q_k - Tq_{k-1}\|_\infty + \gamma \|q_{k-1} - q^*\|_\infty \le \|q_k - Tq_{k-1}\|_\infty + \gamma\delta_{k-1} \,.$$

Unfolding this and using $\delta_0 \le H_\gamma$,

$$\delta_K \le \gamma^K H_\gamma + H_\gamma \max_{1 \le k \le K} \|q_k - Tq_{k-1}\|_\infty \,.$$

Taking a union bound over $k \in [K]$ and plugging in the bound from the previous item, we get

$$\delta_K \le H_\gamma \gamma^K + H_\gamma \left\{ (1 + \sqrt{d})\varepsilon_{\mathrm{apx}} + \sqrt{d} H_\gamma \sqrt{\frac{\log\left(\frac{2|C|K}{\zeta}\right)}{2m}} \right\} .$$

Finally, by our policy error bound,

$$\delta \le 2H_\gamma^2 \left\{ (1 + \sqrt{d})\varepsilon_{\mathrm{apx}} + \gamma^K + \sqrt{d} H_\gamma \sqrt{\frac{\log\left(\frac{2|C|K}{\zeta}\right)}{2m}} \right\} .$$

5. Let $\hat{\pi}$ be the random policy computed by the algorithm. Let $\mathcal{E}$ be the event of the previous part. By the previous part, on $\mathcal{E}$,

$$v^* - v^{\hat{\pi}} \le \delta \mathbf{1}\,.$$

Now, for any fixed $s \in \mathcal{S}$,

$$\begin{aligned}\mathbb{E}[v^*(s) - v^{\hat{\pi}}(s)] &= \mathbb{E}[(v^*(s) - v^{\hat{\pi}}(s))\mathbb{I}_{\mathcal{E}}] + \mathbb{E}[(v^*(s) - v^{\hat{\pi}}(s))\mathbb{I}_{\mathcal{E}^c}] \\ &\le \mathbb{E}[\delta \mathbb{I}_{\mathcal{E}}] + \mathbb{E}[H_\gamma \mathbb{I}_{\mathcal{E}^c}] \\ &= \delta \mathbb{P}(\mathcal{E}) + H_\gamma \mathbb{P}(\mathcal{E}^c) \\ &\le \delta + H_\gamma \zeta\,,\end{aligned}$$

where the last inequality used Q3.

Now, from the result of Q4,

$$\delta + H_\gamma \zeta \le 2H_\gamma^2(1+\sqrt{d})\varepsilon_{\text{apx}} + 2\underbrace{H_\gamma^2\left[\gamma^K + \sqrt{d}H_\gamma\sqrt{\frac{\log\left(\frac{2|C|K}{\zeta}\right)}{2m}} + \frac{\zeta}{2H_\gamma}\right]}_{\le \epsilon} \le 2H_\gamma^2(1+\sqrt{d})\varepsilon_{\text{apx}} + 2\epsilon.$$

Letting each of the last three terms above to be less than $\epsilon/3$, we get the following conditions:

$$\begin{aligned}K &\ge \frac{\log\left(3H_\gamma^2/\epsilon\right)}{\log(1/\gamma)}, \\ \zeta &\le 2\epsilon/(3H_\gamma), \quad \text{and} \\ m &\ge \frac{9H_\gamma^6 d}{2\epsilon^2}\left[\log(2|C|) + \log K + \log\left(3H_\gamma/(2\epsilon)\right)\right].\end{aligned}$$

Recalling that $|C| \le d(d+1)/2$ gives us the desired result.

6. From Q1, we know that the query cost $O(Kd^2m)$, the runtime complexity $O(Kd^3mA)$, and the space complexity $O(d^2)$ are all polynomial in A, d, K, and m. Therefore, the result follows from Q5, which shoes that both K and m themselves have polynomial dependence on H_γ, d, and $1/\epsilon$, and that policy is δ-suboptimal with $\delta = 2H_\gamma^2(1+\sqrt{d})_{\varepsilon_{\text{apx}}} + 2\epsilon$.

7. It suffices to see that for any $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, $Tq \in \mathcal{F}_\phi$. Indeed, then for any θ, $T\Pi f_\theta \in \mathcal{F}_\phi$, which means that $\inf_{\theta'} \|f_{\theta'} - T\Pi f_\theta\|_\infty = 0$.

Fix now $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. Let $v = Mq$. Letting $Z \in \mathbb{R}^{d\times \mathrm{S}}$ be defined using $Z(i,s') = \mu_i(s')$, notice that for $P \in \mathbb{R}^{\mathrm{SA}\times\mathrm{S}}$ it holds that $P = \Phi Z$ while $r = \Phi\theta_r$. Hence,

$$Tq = r + \gamma P v = \Phi\theta_r + \gamma\Phi Z v = \Phi(\theta_r + \gamma Z v)\,,$$

which shows that $Tq \in \mathcal{F}_\phi$, finishing the proof.

Total for all questions: 150. Of this, 30 are bonus marks (i.e., 120 marks worth 100% on this problem set).