# 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

## Tightness of performance bounds of greedy policies

Error bounds for greedy policies are at the heart of many of the upper bounds we obtained. Here you will be asked to show that these bounds are unimprovable. For example, in Lecture 6, the following is stated in Part II of the "Policy error bound - I." lemma:

**Lemma 1.** *Let $\pi$ be a memoryless policy and choose a function $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $\epsilon \geq 0$. Then, if $\pi$ is greedy with respect to $q$ then*

$$v^\pi \geq v^* - \frac{2\|q - q^*\|_\infty}{1-\gamma}\mathbf{1}\,.$$

The first problem is to show that this bound is tight:

**Question 1.** Show that for any $\gamma \in [0, 1)$ and $\varepsilon > 0$ there is a finite discounted MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ and $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that the following hold:

1. $\|q - q^*\|_\infty = \varepsilon$;
2. There is policy $\pi$ that is greedy with respect to $q$ such that $\|v^\pi - v^*\|_\infty = \frac{2\varepsilon}{1-\gamma}$.

Total: **10 points**

---

*Solution.* The MDP will have a single state, call it $s$ (i.e., $\mathcal{S} = \{s\}$) and two actions, say, $\mathcal{A} = \{1, 2\}$. Since there is only a single state, of course $p_a(s|s) = 1$ for both actions $a$. Let

$$0 \leq r_1(s) \leq r_2(s) = 1\,.$$

Then

$$v^*(s) = \frac{1}{1-\gamma}\,.$$

For the policy $\pi$ that chooses action 1 in state $s$, we have

$$v^\pi(s) = \frac{r_1(s)}{1-\gamma}\,.$$

Hence,

$$\|v^\pi - v^*\|_\infty = \frac{1-r_1(s)}{1-\gamma}$$

and thus $\|v^\pi - v^*\|_\infty = \frac{2\varepsilon}{1-\gamma}$ if $r_1(s) = 1-2\varepsilon$. Hence, choose this for the value of $r_1(s)$. Thus,

$$\begin{aligned} q^*(s,1) &= r_1(s) + \gamma\frac{1}{1-\gamma} = \frac{1}{1-\gamma} - 2\varepsilon\,, \\ q^*(s,2) &= \frac{1}{1-\gamma} \end{aligned}$$

and hence if

$$q(s,1) = q^*(s,1) + \varepsilon \text{ and } q(s,2) = q^*(s,2) - \varepsilon$$

then $q(s,1) = q(s,2) = \frac{1}{1-\gamma} - \epsilon$, $\|q - q^*\|_\infty = \varepsilon$ and policy $\pi$ is greedy with respect to $q$ (because any policy is greedy with respect to $q$ as it assigns the same value for both actions). By our previous argument, $\|v^\pi - v^*\|_\infty = 2\varepsilon/(1-\gamma)$, thus, finishing the proof. □

## 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 $\alpha$ 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 ($\alpha$-weighted) **mixture of** $\pi^{(1)}, \dots, \pi^{(k)}$. Call the policy that is obtained following the second method the ($\alpha$-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 $\alpha$. 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 $\mu$.

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 2.** 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 $\mu$ 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 $\sigma$-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 $\mu$. 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 $\mu$ 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 $\pi$ is the $\alpha$-average of $\pi_1, \dots, \pi_k$ then $v^\pi < v$.

   **10 points**

8. The state-wise uniform average of all deterministic ML policies and the uniform mixture of all deterministic ML policies both give the policy that is uniform over all the actions.

   **5 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 $\alpha_i$, $i \in [k]$ 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: **70 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 $\sigma$-algebra, it suffices to check this for the "simple" cylinder sets, i.e., when $U$ is either of the form

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

   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 $\mu$ 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 $\pi_0$ 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 $\mu$ ($\mathbb{P}$ hides its dependence on $\mu$). Again, this is clear when $\mathbb{P}(H_t = h_t) = 0$ since $\pi_0$ does not depend on $\mu$. 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 $\mu$. 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,

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

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 $\mu$ 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 $\mu$, 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

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

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

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 $\pi_i$ uses action $i$ (moving to state $i$) everywhere. The value of both $\pi_1$ and $\pi_2$ is above $\gamma/(1-\gamma)$. The uniform average chooses the actions at random at both states. The value of the averaged policy $\pi$ at both states is $\frac{1}{2(1-\gamma)}$, which is lower than $\gamma/(1-\gamma)$ provided that $\gamma > 1/2$.

8. This question was incorrect, it did not hold for the mixing case. For averaging it is trivial.

For mixing we show a counter example. Our goal is to show $\mathbb{P}(A_t = a | H_t = h_t) = 1/A$ if we restrict ourselves to uniform mixtures of all deterministic "ML" policies. Let $\mathcal{S} = \{s_1\}$ and $\mathcal{A} = \{a_1, a_2\}$, then pick $h_t = (s_1, a_1, s_1)$. Only two deterministic ML policies exist, define them such that $\pi^{(1)}(a_1|s_1) = 1$ and $\pi^{(2)}(a_2|s_1) = 1$. Then we have

$$\begin{aligned}\mathbb{P}(A_1 = a_1 | H_1 = (s_1, a_1, s_1)) &= \sum_i \mathbb{P}(A_1 = a_1 | H_1 = (s_1, a_1, s_1), I = i)\mathbb{P}(I = i | H_1 = (s_1, a_1, s_1))\\ &= \sum_i \pi^{(i)}(A_1 = a_1 | H_1 = (s_1, a_1, s_1))\mathbb{P}(I = i | H_1 = (s_1, a_1, s_1))\\ &= \pi^{(1)}(A_1 = a_1 | H_1 = (s_1, a_1, s_1))\mathbb{P}(I = 1 | H_1 = (s_1, a_1, s_1)) +\\ &\quad \pi^{(2)}(A_1 = a_1 | H_1 = (s_1, a_1, s_1))\mathbb{P}(I = 2 | H_1 = (s_1, a_1, s_1))\\ &= 1 * 1 + 0 * 0 = 1 \neq 1/2\end{aligned}$$

Basically, if it is not the first time encountering a state in the trajectory then the action that that will be selected is deterministic (exactly the same action that was selected in that state the last time is was seen).

□

## Finding needles with high probability

The high-probability needle lemma is as follows:

**Lemma 2** (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 3.** Prove Lemma 2. 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 $p(\frac{k+1}{2} - \frac{1}{k}) \le q_{k,i}$. Thus,

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

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 $t \ge 1$ of algorithm $A$ and where $S_t \in \{0, 1\}$ indicates whether this call was successful. By construction, $((Q_t, S_t))_{t\ge1}$ is an i.i.d. sequence, with common distribution $P_{k,i}$. Also, by definition,

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

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 $\mathbb{P}(N \ge n) = (1 - p_{k,i})^{n-1}$. As $\mathbb{P}(N < \infty) = 1 - \lim_{n\to\infty} \mathbb{P}(N \ge n) = 1$, 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. $(Q_n)_{n\ge1}$ share the same finite-mean;
2. $\mathbb{E}[N] < \infty$;
3. $\mathbb{E}[Q_n \mathbb{I}\{N \ge n\}] = \mathbb{E}[Q_n]\mathbb{P}(N \ge n)$ for all $n \ge 1$;
4. $\sum_{n=1}^{\infty} \mathbb{E}[|Q_n|\mathbb{I}\{N \ge n\}] < \infty$.

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,

$$\mathbb{E}[N] = \sum_{n\ge1} \mathbb{P}(N \ge n) = \frac{1}{p_{k,i}} , \tag{4}$$

combined with (2) gives

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

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 $\{N \ge n\} = \{S_1 = 0, \ldots, S_{n-1} = 0\}$ whose indicator is independent of $Q_n$ (since $Q_n$ and $(S_1, \ldots, S_{n-1})$ are independent). Hence,

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

as required. The fourth condition follows from the third: $\sum_{n\ge1} \mathbb{E}[|Q_n|\mathbb{I}\{N \ge n\}] = \sum_{n\ge1} \mathbb{E}[Q_n \mathbb{I}\{N \ge n\}] = \sum_{n\ge1} \mathbb{E}[Q_n]\mathbb{P}(N \ge n) = q_{k,i}\mathbb{E}[N] < \infty$.

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: $R \in \{P(M+1), \dots, P(k)\}$. 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 $P^{-1}(i) \le M$, otherwise $C = R$. (Note that $P^{-1}(i) \le M$ is equivalent to $i \in \{P(1), \dots, P(M)\}$.) 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":

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

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

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

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

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

Hence,

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

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

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

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

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

it holds that

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

Now,

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

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,

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

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. □

---

**Total for all questions: 110**. Of this, up to 20 can be bonus marks. You can receive bonus marks by asking/upvoting questions, for a total of 20 bonus marks! You must ask at least one question in one of the Lecture Discussion Threads by the Assignment 3 deadline to receive 12 bonus marks. You can also receive 2 bonus marks for upvoting at least one question before 8am on the day of each lecture, for a maximum of 2 marks x 4 lectures = 8 marks for upvoting. Your assignment will be marked out of 110 minus the bonus marks you received.