CMPUT 605: Theoretical Foundations of Reinforcement Learning, Winter 2023
Homework #1

Instructions

Submissions You need to submit a single PDF file, named `p01_<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: January 29 at 11:55 pm

Problems

Unless otherwise stated, for the problem described below all policies, value functions, etc. are for a discounted, finite MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$. That is, $\mathcal{S}$ and $\mathcal{A}$ are finite, $0 \le \gamma < 1$. Also, without the loss of generality, $\mathcal{S} = [\mathrm{S}] = \{1, \dots, \mathrm{S}\}$ and $\mathcal{A} = [\mathrm{A}] = \{1, \dots, \mathrm{A}\}$. Below we use notation introduced in the lecture without redefining it, e.g., $\mathbb{P}_\mu^\pi$, $\mathbb{E}_\mu^\pi$, v^π, v^*, T_π, T, etc. All these objects are to be understood in the context of the fixed $\mathcal{M}$.

Question 1. Show that for any policy π (not necessarily memoryless) and distribution $\mu \in \mathcal{M}_1(\mathcal{S})$ over the states, $v^\pi(\mu) = \sum_{s \in \mathcal{S}} \mu(s) v^\pi(s)$.
Hint: Read the end-notes to Lecture 2. Use the canonical probability space for MDPs and the cylinder sets to show that $\mathbb{P}_\mu = \sum_{s \in \mathcal{S}} \mu(s) \mathbb{P}_s$.

Total: **10 points**

Question 2. Recall that for a memoryless policy π, P_π is the $\mathrm{S} \times \mathrm{S}$ matrix whose (s, s')th entry is

$$\sum_{a \in \mathcal{A}} \pi(a|s) P_a(s, s') \,.$$

Show that for any $s, s' \in \mathcal{S}$ and $t \ge 1$, $(P_\pi^t)_{s,s'} = \mathbb{P}_s^\pi(S_t = s')$.
Hint: Use the properties of $\mathbb{P}_s$ (the tower rule of conditional expectations may be useful, too, especially if you do not want to write a lot).

Total: **10 points**

Question 3. Prove that for any memoryless policy π, $v^\pi = \sum_{t \ge 0} \gamma^t P_\pi^t r_\pi$.
Hint: You may want to reuse the result of the previous exercise.

Total: **10 points**

Question 4. Prove that for any memoryless policy π, v^π is the fixed point of T_π: $v^\pi = T_\pi v^\pi$.

Total: **5 points**

Question 5. Let $w \in (0,\infty)^{\mathrm{S}}$ be an S-dimensional vector whose entries are all positive. Let $\tilde{v}^*$ be a solution to the optimization problem

$$\max_{v \in \mathbb{R}^{\mathrm{S}}} w^\top v \qquad \text{s.t.} \qquad v \le Tv \,. \tag{1}$$

Show that $\tilde{v}^* = v^*$. That is, the unique solution to the problem stated in (1) is v^*.

Total: **5 points**

Question 6. Let $w \in (0,\infty)^{\mathrm{S}}$ be an S-dimensional vector whose entries are all positive. Let $\tilde{v}^*$ be a solution to the optimization problem

$$\min_{v \in \mathbb{R}^{\mathrm{S}}} w^\top v \qquad \text{s.t.} \qquad v \ge Tv \,. \tag{2}$$

Show that $\tilde{v}^* = v^*$. That is, the unique solution to the problem stated in (2) is v^*.

Total: **5 points**

Question 7. A linear program is a constrained optimization problem with a linear objective and linear constraints. Which of (1) or (2) is equivalent to a linear program? Give the linear program and show the equivalence.

Total: **5 points**

Question 8. Show that for any policy π and distribution $\mu \in \mathcal{M}_1(\mathcal{S})$ there is a memoryless policy π' such that $\nu_\mu^\pi = \nu_\mu^{\pi'}$ (i.e., memoryless policies exhaust the set of all discounted state-action occupancy measures). **Hint**: For arbitrary π, μ, let $\tilde{\nu}_\mu^\pi(s) = \sum_{a \in \mathcal{A}} \nu_\mu^\pi(s,a)$. Define $\pi'(a|s) = \nu_\mu^\pi(s,a)/\tilde{\nu}_\mu^\pi(s)$ when the denominator is nonzero, and otherwise let $\pi'(\cdot|s)$ be an arbitrary distribution. Show that $\tilde{\nu}_\mu^\pi = \mu + \gamma \tilde{\nu}_\mu^\pi P_{\pi'}$ (treating $\tilde{\nu}_\mu^\pi$ and μ as row-vectors) to conclude that $\tilde{\nu}_\mu^\pi = \tilde{\nu}_\mu^{\pi'}$. To conclude, use the definition of π' and that for memoryless policies π'', $\tilde{\nu}_\mu^{\pi''}(s)\pi''(a|s) = \nu_\mu^{\pi''}(s,a)$.

Total: **15 points**

For the next questions, define the operators

$$P : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}\,, \quad M : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}}\,, \quad M_\pi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}}$$

via

$$(Pv)(s,a) = \langle P_a(s), v \rangle\,, \qquad (Mq)(s) = \max_{a \in \mathcal{A}} q(s,a)\,, \qquad (M_\pi q)(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q(s,a)\,,$$

where $(s,a) \in \mathcal{S} \times \mathcal{A}$, $v \in \mathbb{R}^{\mathcal{S}}$, $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and π is an arbitrary memoryless policy. Further, let $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be defined by $r(s,a) = r_a(s)$, $(s,a) \in \mathcal{S} \times \mathcal{A}$. It is easy to see that for any $v \in \mathbb{R}^{\mathcal{S}}$ the following hold:

$$Tv = M(r + \gamma P v)\,, \tag{3}$$
$$T_\pi v = M_\pi(r + \gamma P v)\,. \tag{4}$$

Question 9. Let π be a memoryless policy. Show that T_π is a γ-contraction with respect to the max-norm.

Total: **5 points**

Question 10. Show that M, M_π and P as defined above are non-expansion when there domains and ranges are equipped with the maximum norm. That is, show that for all $q, q' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $v, v' \in \mathbb{R}^{\mathcal{S}}$,

$$\begin{aligned} \|Mq - Mq'\|_\infty &\le \|q - q'\|_\infty\,, \\ \|M_\pi q - M_\pi q'\|_\infty &\le \|q - q'\|_\infty\,, \\ \|Pv - Pv'\|_\infty &\le \|v - v'\|_\infty\,. \end{aligned}$$

Hint: To show that M is a non-expansion, consider proving that $|\max_a q(a) - \max_b q'(b)| \le \|q - q'\|_\infty$ holds for any $q, q' \in \mathbb{R}^{\mathcal{A}}$.

Total: **10 points**

Question 11. Let $\tilde{T} : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be defined using $\tilde{T}q = r + \gamma PMq$. Show that $\tilde{T}$ is a γ-contraction with respect to the max-norm.

Total: **5 points**

Question 12. Let q^* be the fixed point of $\tilde{T}$ defined in Question 11. Show that $v^* = Mq^*$.

Total: **8 points**

Question 13. Let q^* be the fixed point of $\tilde{T}$ as before. Show that $q^* = r + \gamma P v^*$.

Total: **5 points**

Question 14. Show that if $q^* \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is the fixed-point of $\tilde{T}$ and if π is a memoryless policy that chooses actions maximizing q^* (i.e. $M_\pi q^* = Mq^*$) then π is an optimal policy and any memoryless optimal policy can be found this way.

Total: **5 points**

Question 15. Let π be a memoryless policy and $\epsilon > 0$. Call π *ϵ-optimizing* $M_\pi q^* \ge v^* - \epsilon\,\mathbb{1}$. Show that if π is ϵ-optimizing then π is $\epsilon/(1-\gamma)$-optimal, that is, $v^\pi \ge v^* - \frac{\epsilon}{1-\gamma}\,\mathbb{1}$.

Total: **10 points**

Question 16. Show that if $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is such that $\|q - q^*\|_\infty \le \epsilon$ and π is greedy with respect to q (i.e., $M_\pi q = Mq$) then π is $2\epsilon/(1-\gamma)$ optimal.
Hint: Aim for reusing the answer to Question 15.

Total: **5 points**

Question 17. Let π be a memoryless policy that selects ϵ-optimal actions with probability at least $1-\zeta$ in each state (i.e., $\sum_{a:q^*(s,a)\geq v^*(s)-\epsilon} \pi(a|s) \geq 1-\zeta$). Show that π is at least $(\epsilon + 2\zeta\|q^*\|_\infty)/(1-\gamma)$ optimal. Only assume that the reward is deterministic and bounded (i.e. do not assume it is in $[0,1]$). **Hint**: Aim for showing first that π is $(\epsilon + 2\zeta\|q^*\|_\infty)$-optimizing.

Total: **5 points**

Total for all questions: 123. Of this, 23 are bonus marks. Your assignment will be marked out of 100.