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**

Solution. We follow the hint. Let $((\mathcal{S}\times\mathcal{A})^{\mathbb{N}}, \mathcal{F})$ be the canonical probability space of trajectories over which the measures $(\mathbb{P}_\mu^\pi)_{\mu,\pi}$ are defined. Fix π. We let $\mathbb{P}_s = \mathbb{P}_s^\pi$ and let $\mathbb{P}_\mu = \mathbb{P}_\mu^\pi$ (suppressing dependence on π). The statement will follow from

$$\mathbb{P}_\mu = \sum_{s\in\mathcal{S}} \mu(s)\mathbb{P}_s\,, \tag{1}$$

since if this holds then

$$\begin{aligned}
v^\pi(\mu) &= \int R(\omega)\mathbb{P}_\mu(d\omega) && \text{(definition of } v^\pi(\mu)\text{)}\\
&= \sum_{s\in\mathcal{S}} \mu(s) \int R(\omega)\mathbb{P}_s(d\omega) && \text{(Eq. (1) and linearity of integrals)}\\
&= \sum_{s\in\mathcal{S}} \mu(s) v^\pi(s)\,, && \text{(definition of } v^\pi(s)\text{)}
\end{aligned}$$

where $R(\omega)$ is the return on $\omega \in \Omega$. To show Eq. (1) recall that a measure over the product measurable space $(\Omega, \mathcal{F})$ is uniquely defined based on what probabilities it assigns to the cylinder sets that take either the form

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

By the properties of $\mathbb{P}_\mu$,

$$\mathbb{P}_\mu(C) = \mu(s_0)\pi_0(a_0|s_0)P_{a_0}(s_0, s_1) \dots P_{a_{t-1}}(s_{t-1}, s_t)\,,$$
$$\mathbb{P}_\mu(C') = \mu(s_0)\pi_0(a_0|s_0)P_{a_0}(s_0, s_1) \dots P_{a_{t-1}}(s_{t-1}, s_t)\pi_t(a_t|s_0, a_0, \dots, s_t)\,,$$

Applying this with $\mu = \delta_s$, we get that indeed

$$\mathbb{P}_\mu(C) = \sum_s \mu(s)\mathbb{P}_s(C) \qquad \text{and}$$
$$\mathbb{P}_\mu(C') = \sum_s \mu(s)\mathbb{P}_s(C')\,.$$

Since C and C' were arbitrary cylinder sets of the above form, Eq. (1) holds. □

Question 2. Recall that for a memoryless policy π, P_π is the S × 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 $t \geq 1$, $(P_\pi^t)_{s,s'} = \mathbb{P}_s^\pi(S_t = s')$.
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**

Solution. Fix any $t \geq 0$. Fix also π and $s_0 \in \mathcal{S}$. We abbreviate $\mathbb{P}_{s_0}^\pi$ to $\mathbb{P}$ in what follows (we changed s to s_0 so that there is no clash with indexing of states below while we can reduce clutter). Detto for $\mathbb{E}_{s_0}^\pi$ and $\mathbb{E}$. Recall that $H_t = (S_0, A_0, \dots, S_{t-1}, A_{t-1}, S_t)$. Fix $s' \in \mathcal{S}$. By the tower rule of conditional expectations (applied twice),

$$\mathbb{P}(S_{t+1} = s') = \mathbb{E}[\mathbb{E}[\mathbb{P}(S_{t+1} = s'|H_t, A_t)|H_t]]\,.$$

For the innermost expectation (probability, actually) we have

$$\mathbb{P}(S_{t+1} = s'|H_t, A_t) = P_{A_t}(S_t, s')$$

by the construction of $\mathbb{P}$. Now,

$$\mathbb{E}[P_{A_t}(S_t, s')|H_t] = \sum_{a \in \mathcal{A}} \pi_t(a|H_t)P_a(S_t, s')$$

and since π is memoryless, $\pi_t(a|H_t) = \pi(a|S_t)$. Hence, the expression in the right-hand side is $P_\pi(S_t, s')$. Plugging this in, using the law of total expectations,

$$\mathbb{E}[P_\pi(S_t, s')] = \sum_{s \in \mathcal{S}} P_\pi(s, s')\mathbb{P}(S_t = s)\,.$$

Putting everything together we see that

$$\mathbb{P}(S_{t+1} = s') = \sum_{s \in \mathcal{S}} P_\pi(s, s') \mathbb{P}(S_t = s)$$

which, together with $\mathbb{P}(S_0 = s) = \delta_{s_0}(s)$ implies the desired statement. Indeed, for $t = 0$ we get

$$\mathbb{P}(S_1 = s') = P_\pi(s_0, s') = P_\pi(s_0, s')\,,$$

and hence, by induction,

$$\mathbb{P}(S_{t+1} = s') = \sum_{s \in \mathcal{S}} P_\pi^t(s_0, s) P_\pi(s, s') = P_\pi^{t+1}(s_0, s')\,.$$

□

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

Total: **10 points**

Solution. Let $\mathbb{P} = \mathbb{P}_s^\pi$. By the result of the previous exercise, for $t \geq 1$, $\mathbb{P}(S_t = s') = P_\pi^t(s, s')$. By the tower rule and Lebesgue's dominated convergence theorem,

$$v^\pi(s) = \sum_{t \geq 0} \gamma^t \mathbb{E}[\mathbb{E}[r_{A_t}(S_t)|H_t]]\,.$$

For the innermost expectation we have

$$\mathbb{E}[r_{A_t}(S_t)|H_t] = \sum_{a \in \mathcal{A}} \pi_t(a|H_t) r_a(S_t) = \sum_{a \in \mathcal{A}} \pi(a|S_t) r_a(S_t) = r_\pi(S_t)\,,$$

because π is memoryless. Plugging this in and using the law of total expectations we get

$$\mathbb{E}[\mathbb{E}[r_{A_t}(S_t)|H_t]] = \sum_{s' \in \mathcal{S}} \mathbb{P}(S_t = s') r_\pi(s')\,.$$

Since for $t = 0$, $\mathbb{P}(S_t = s') = 1$ iff $s' = s$, this together with the result of the previous problem gives that

$$v^\pi(s) = \sum_{t \geq 0} \gamma^t \sum_{s' \in \mathcal{S}} P_\pi^t(s, s') r_\pi(s')$$

(recall that A^0 is the identity matrix for any square matrix A). Using a matrix-vector notation we can write the above display as

$$v^\pi = \sum_{t \geq 0} \gamma^t P_\pi^t r_\pi$$

□

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**

Solution. Fix $s \in \mathcal{S}$ and π. From the solution of the previous problem,

$$v^\pi = r_\pi + \gamma P_\pi \underbrace{\sum_{t \geq 0} \gamma^t P_\pi^t r_\pi}_{v^\pi} = r_\pi + \gamma P_\pi v^\pi \,,$$

finishing the proof. □

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

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

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

Total: **5 points**

Solution. Let v be any feasible point: $v \leq Tv$. Let π be greedy w.r.t. v. Hence, $T_\pi v = Tv \geq v$ and by induction on $k \geq 0$, for any $k \geq 0$, $T_\pi^k v \geq v$ and hence $v^\pi \geq v$. Since π was arbitrary memoryless, by the fundamental theorem, $v^* = \sup_{\pi \in \mathrm{ML}} v^\pi \geq v$.

Now, let v be the solution of the optimization problem. Hence, v is feasible and thus $v^* \geq v$. If there is a state $s_0 \in \mathcal{S}$ such that $v^*(s_0) > v(s_0)$ then $\mathbb{1}^\top v^* > \mathbb{1}^\top v$, contradicts that v is an optimal solution since v^* is also a feasible point of the optimization problem. Therefore $v^* = v$. □

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

$$\min_{v \in \mathbb{R}^{\mathrm{S}}} w^\top v \qquad \text{s.t.} \qquad v \geq Tv \,. \tag{3}$$

Show that $\tilde{v}^* = v^*$. That is, the unique solution to the problem stated in (3) is v^*.

Total: **5 points**

Solution. Let v be any feasible point: $v \geq Tv$. Iterating with T we get that $v \geq v^*$. Now, let v be the solution of the optimization problem. Hence, v is feasible and thus $v \geq v^*$. If there is a state $s_0 \in \mathcal{S}$ such that $v(s_0) > v^*(s_0)$ then $\mathbb{1}^\top v > \mathbb{1}^\top v^*$, which contradicts that v is an optimal solution since v^* is also a feasible point. Therefore $v^* = v$. □

Question 7. A linear program is a constrained optimization problem with a linear objective and linear constraints. Which of (2) or (3) is equivalent to a linear program? Give the linear program and show the equivalence.

Total: **5 points**

Solution. It is (3). The linear program takes the form

$$\min_{v \in \mathbb{R}^{\mathrm{S}}} w^\top v \qquad \text{s.t.} \qquad v \geq T_a v, \qquad a \in \mathcal{A} \,. \tag{4}$$

This is equivalent to (3) because they have the same objective and same feasibility sets. In particular, if v is feasible for (3) then $v \geq Tv \geq T_a v$ for any $a \in \mathcal{A}$. Hence, $v \geq T_a v$ holds for all $a \in \mathcal{A}$ and v is feasible for (4). In the reverse direction, if v is feasible for (4), $v(s) \geq (T_a v)(s)$ holds for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$. Taking the maximum of both sides over a, we get $v(s) \geq \max_{a \in \mathcal{A}} (T_a v)(s) = (Tv)(s)$ where the last equality used the definitions of T_a and T. Since this holds for any $s \in \mathcal{S}$, $v \geq Tv$, i.e., we get that v is also feasible for (3). □

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**

Solution. For arbitrary μ, π, define

$$\tilde{\nu}_\mu^\pi(s) = \sum_{a \in \mathcal{A}} \nu_\mu^\pi(s,a)\,,$$

the 'marginal' of $\nu_\mu^\pi \in \mathcal{M}_1(\mathcal{S} \times \mathcal{A})$ over the states. Clearly,

$$\tilde{\nu}_\mu^\pi(s) = \sum_{t \geq 0} \gamma^t \mathbb{P}_\mu^\pi(S_t = s)\,.$$

When π is a memoryless policy, since $\mathbb{P}_\mu^\pi(S_t = s) = \mu P_\pi^t e_s$,

$$\tilde{\nu}_\mu^\pi = \mu \sum_{t \geq 0} (\gamma P_\pi)^t\,. \tag{5}$$

Now fix μ, π as in the theorem and pick an arbitrary distribution $\pi_0 \in \mathcal{M}_1(\mathcal{A})$ over the actions. Define π' as follows:

$$\pi'(a|s) = \begin{cases} \frac{\nu_\mu^\pi(s,a)}{\tilde{\nu}_\mu^\pi(s)}\,, & \text{if } \tilde{\nu}_\mu^\pi(s) \neq 0\,; \\ \pi_0(a)\,, & \text{otherwise}\,. \end{cases}$$

We will now argue that this policy is indeed suitable. In particular, we will show that

$$\tilde{\nu}_\mu^\pi = \mu + \gamma \tilde{\nu}_\mu^\pi P_{\pi'}, \tag{6}$$

which implies the result since viewing this as a (linear) equation in $\tilde{\nu}_\mu^\pi$, the unique solution to this equation is $\tilde{\nu}_\mu^{\pi'}$, the discounted occupancy measure of π' over the states (see Eq. (5)). Thus, $\tilde{\nu}_\mu^\pi = \tilde{\nu}_\mu^{\pi'}$ and thus,

$$\nu_\mu^\pi(s,a) = \tilde{\nu}_\mu^\pi(s)\pi'(a|s) = \tilde{\nu}_\mu^{\pi'}(s)\pi'(a|s) = \nu_\mu^{\pi'}(s,a)\,,$$

where the last equality follows from the definitions of $\tilde{\nu}_\mu^{\pi'}$ and ν_μ^π and the fact that π' is memoryless.

It remains to show that (6) holds. For this, we have

$$\begin{aligned}
\tilde{\nu}_\mu^\pi(s) &= \sum_{t \geq 0} \gamma^t \mathbb{P}_\mu^\pi(S_t = s) \\
&= \mu(s) + \gamma \sum_{t \geq 0} \gamma^t \mathbb{P}_\mu^\pi(S_{t+1} = s) \\
&= \mu(s) + \gamma \sum_{s_{\mathrm{prev}}, a} \underbrace{\sum_{t \geq 0} \gamma^t \mathbb{P}_\mu^\pi(S_t = s_{\mathrm{prev}}, A_t = a)}_{\nu_\mu^\pi(s_{\mathrm{prev}}, a)} P_a(s_{\mathrm{prev}}, s) \\
&= \mu(s) + \gamma \sum_{s_{\mathrm{prev}}} \tilde{\nu}_\mu^\pi(s_{\mathrm{prev}}) \sum_a \pi'(a|s_{\mathrm{prev}}) P_a(s_{\mathrm{prev}}, s) \\
&= \mu(s) + \gamma \sum_{s_{\mathrm{prev}}} \tilde{\nu}_\mu^\pi(s_{\mathrm{prev}}) P_{\pi'}(s_{\mathrm{prev}}, s)\,,
\end{aligned}$$

which is equivalent to (6). Here, the last equality follows from the definition of $P_{\pi'}$. □

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:

$$Tv = M(r + \gamma P v), \tag{7}$$
$$T_\pi v = M_\pi(r + \gamma P v). \tag{8}$$

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**

Solution. Let $v, v' \in \mathbb{R}^{\mathrm{S}}$. By the triangle inequality we have $|(P_\pi(v-v'))(s) \le \sum_{s'\in\mathcal{S}} P_\pi(s,s')|v(s)-v(s')| \le \|v-v'\|_\infty \sum_{s'\in\mathcal{S}} P_\pi(s,s') = \|v-v'\|_\infty$. Taking the maximum over s, $\|P_\pi(v-v')\|_\infty \le \|v-v'\|_\infty$. Finally, $\|T_\pi v - T_\pi v'\|_\infty = \|r_\pi + \gamma P_\pi v - (r_\pi + \gamma P_\pi v')\|_\infty = \gamma\|P_\pi(v-v')\|_\infty \le \gamma\|v-v'\|_\infty$. □

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 $v, v' \in \mathbb{R}^{\mathcal{S}}$,

$$\|Mq - Mq'\|_\infty \le \|q-q'\|_\infty,$$
$$\|M_\pi q - M_\pi q'\|_\infty \le \|q-q'\|_\infty,$$
$$\|Pv - Pv'\|_\infty \le \|v-v'\|_\infty.$$

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**

Solution. Let us start with P. This follows because P is a stochastic operator, i.e., that $P\,\mathbb{1} = \mathbb{1}$ and P is monotone and linear. Thus, with $c = \|v-v'\|_\infty$, from $-c\,\mathbb{1} \le v - v' \le c\,\mathbb{1}$, applying P we get $-c\,\mathbb{1} \le P(v-v') \le c\,\mathbb{1}$, which implies that $\|P(v-v')\|_\infty \le c$.

That M is a non-expansion follows by the following elementary argument: Let $q, q' : \mathcal{A} \to \mathbb{R}$ be arbitrary. We want to prove that

$$|\max_a q(a) - \max_b q'(b)| \le \|q-q'\|_\infty. \tag{9}$$

If this is proven, we can apply this result $q(s,\cdot)$ and $q'(s,\cdot)$ and then taking a maximum over s to get that for any q, q', $\|Mq - Mq'\|_\infty \le \|q-q'\|_\infty$. To prove (9) assume without the loss of generality that

$$\max_a q(a) - \max_b q'(b) \ge 0. \tag{10}$$

Take any $a \in \mathcal{A}$. Then,

$$|q(a) - q'(a)| \ge q(a) - q'(a) \ge q(a) - \max_b q'(b).$$

Taking the maximum of both sides over $a \in \mathcal{A}$ gives

$$\|q - q'\|_\infty \geq \max_a q(a) - \max_b q'(b) = |\max_a q(a) - \max_b q'(b)|\,,$$

where the equality follows from (10).

The proof of $\|M_\pi q - M_\pi q'\|_\infty \leq \|q - q'\|_\infty$ follows by using the definition of M_π, expanding $\|M_\pi q - M_\pi q'\|_\infty$ and applying a few inequalities as follows:

$$\begin{aligned}
\|M_\pi q - M_\pi q'\|_\infty &= \max_s \left| \sum_{a \in \mathcal{A}} \pi(a|s) \left[q(s,a) - q'(s,a)\right] \right| \\
&\leq \max_s \sum_{a \in \mathcal{A}} \pi(a|s) |q(s,a) - q'(s,a)| && \text{Since } \pi(a|s) \geq 0 \\
&\leq \max_s \sum_{a \in \mathcal{A}} \pi(a|s) \max_a |q(s,a) - q'(s,a)| \\
&= \max_s \sum_{a \in \mathcal{A}} \pi(a|s) \max_a |q(s,a) - q'(s,a)| \\
&= \max_s \max_a |q(s,a) - q'(s,a)| \\
&= \|q - q'\|_\infty
\end{aligned}$$

□

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**

Solution. Follows immediately from the previous solutions and the properties of norms. Defining $v = Mq$ and $v' = Mq'$ we have

$$\begin{aligned}
\|\tilde{T}q - \tilde{T}q'\|_\infty &= \|r + \gamma PMq - r - \gamma PMq'\|_\infty \\
&= \gamma \|Pv - Pv'\|_\infty \\
&\leq \gamma \|v - v'\|_\infty \\
&= \gamma \|Mq - Mq'\|_\infty \\
&\leq \gamma \|q - q'\|_\infty
\end{aligned}$$

□

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

Total: **8 points**

Solution. Let $v = Mq^*$. By the Fundamental Theorem of MDPs and since T is a contraction, v^* is the unique fixed-point of T. Hence, it suffices to show that $Tv = v$ also holds. By definition,

$$q^* = \tilde{T}q^* = r + \gamma Pv\,,$$

where the last equality used the definition of $\tilde{T}$ and v. Applying M on both sides, we get

$$v = Mq^* = M(r + \gamma Pv) = Tv\,,$$

where the last equality used (7). □

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

Total: **5 points**

Solution. By Question 12, $v^* = Mq^*$. By the definition of $\tilde{T}$ and q^*, we have $q^* = \tilde{T}q^* = r + \gamma PMq^* = r + \gamma P v^*$. □

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**

Solution. Let π be greedy with respect to q^*. Hence, $M_\pi q^* = Mq^*$. We have

$$M_\pi q^* = Mq^* .$$

By the previous question, $q^* = r + \gamma P v^*$. Plugging this in, we get

$$T_\pi v^* = M_\pi(r + \gamma P v^*) = M(r + \gamma P v^*) = Tv^* .$$

The result follows by the fundamental theorem of MDPs. □

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

Total: **10 points**

Solution. We have

$$\begin{aligned} 0 \leq v^* - v^\pi &\leq M_\pi q^* - v^\pi + \epsilon\,\mathbb{1} = (r_\pi + \gamma P_\pi v^*) - (r_\pi + \gamma P_\pi v^\pi) + \epsilon\,\mathbb{1} \\ &= \gamma P_\pi (v^* - v^\pi) + \epsilon\,\mathbb{1} , \end{aligned} \tag{11}$$

where the first inequality uses the assumption on π, the second uses that $q^* = r + \gamma P v^*$ and that M_π is linear and $M_\pi r = r_\pi$ and $M_\pi P = P_\pi$, by definition. The last equality follows by algebra. Let $\Delta = \|v^* - v^\pi\|_\infty$. Now, taking the absolute value of both sides in Eq. (11) and then the maximum of both sides with respect to the states, using that P_π is stochastic hence $\|P_\pi v\|_\infty \leq \|v\|_\infty$, we get that

$$\Delta \leq \gamma\Delta + \epsilon .$$

Solving for Δ then gives the result.

□

Question 16. Show that if $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is such that $\|q - q^*\|_\infty \leq \epsilon$ 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**

Solution. Following the hint, we want to show that $M_\pi q^* \geq v^* - 2\epsilon\,\mathbb{1}$. We have

$$M_\pi q^* \geq M_\pi(q - \epsilon\,\mathbb{1}) = M_\pi q - \epsilon\,\mathbb{1} = Mq - \epsilon\,\mathbb{1} \geq M(q^* - \epsilon\,\mathbb{1}) - \epsilon\,\mathbb{1} \geq Mq^* - 2\epsilon\,\mathbb{1} ,$$

where we used that M_π is linear, and $M(v + c\,\mathbb{1}) = Mv + c\,\mathbb{1}$ by its definition ($c \in \mathbb{R}$) and that $-\epsilon\,\mathbb{1} \leq q^* - q \leq \epsilon\,\mathbb{1}$. The result then follows from the claim made in Question 15. □

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**

Solution. Fix $s \in \mathcal{S}$. Let $\mathcal{A}(s,\epsilon) = \{a \in \mathcal{A} : q^*(s,a) \geq v^*(s) - \epsilon\}$ be the set of ϵ-optimal actions in state s. By our assumption, $\sum_{a\in\mathcal{A}(s,\epsilon)} \pi(a|s) \geq 1-\zeta$ and, conversely, $\sum_{a\in\mathcal{A}\setminus\mathcal{A}(s,\epsilon)} \pi(a|s) = 1 - \sum_{a\in\mathcal{A}(s,\epsilon)} \pi(a|s) \leq 1-(1-\zeta) = \zeta$. Thus,

$$
\begin{aligned}
(M_\pi q^*)(s) &= \sum_{a\in\mathcal{A}(s,\epsilon)} \pi(a|s) q^*(s,a) + \sum_{a\notin\mathcal{A}(s,\epsilon)} \pi(a,s) q^*(s,a) \\
&\geq \sum_{a\in\mathcal{A}(s,\epsilon)} \pi(a|s)(v^*(s)-\epsilon) - \|q^*\|_\infty \sum_{a\notin\mathcal{A}(s,\epsilon)} \pi(a,s) \\
&= (v^*(s)-\epsilon) \sum_{a\in\mathcal{A}(s,\epsilon)} \pi(a|s) - \|q^*\|_\infty \sum_{a\notin\mathcal{A}(s,\epsilon)} \pi(a,s) \\
&= (v^*(s)-\epsilon) \left(\sum_{a\in\mathcal{A}} \pi(a|s) - \sum_{a\notin\mathcal{A}(s,\epsilon)} \pi(a,s) \right) - \|q^*\|_\infty \sum_{a\notin\mathcal{A}(s,\epsilon)} \pi(a,s) \\
&= v^*(s) - \epsilon - (v^*(s) - \epsilon + \|q^*\|_\infty) \sum_{a\notin\mathcal{A}(s,\epsilon)} \pi(a,s) \\
&\geq v^*(s) - \epsilon - 2\|q^*\|_\infty \sum_{a\notin\mathcal{A}(s,\epsilon)} \pi(a,s) \\
&\geq v^*(s) - \epsilon - 2\|q^*\|_\infty \zeta \,.
\end{aligned}
$$

The result then follows from the claim made in Question 15. □

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