CMPUT 653: Theoretical Foundations of Reinforcement Learning, Winter 2022
Midterm

Instructions

Submissions You need to submit a zip file, named `midterm_<name>.zip` or `midterm_<name>.pdf` where `<name>` is your name. The zip file should include a report in PDF, typed up (we strongly encourage to use pdfLaTeX) and the code that we asked for. Write your name on your solution. I provide a template that you are encouraged to use. You have to submit the zip file on the eclass website of the course.
Collaboration and sources Work on your own. No consultation, etc. 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: February 25 at 11:55 pm

Undiscounted infinite horizon problems

Let $M = (\mathcal{S}, \mathcal{A}, P, r)$ be a finite MDP as usual, but this time consider the infinite horizon undiscounted total reward criterion. In this setting, the value of policy π (memoryless or not) is

$$v^\pi(s) = \mathbb{E}_s^\pi \left[\sum_{t=0}^{\infty} r_{A_t}(S_t) \right] .$$

To guarantee that this value exist we make the following assumption on the MDP M:

Assumption 1 (All policies proper)**.** Assume that the MDP M has a state $s^\star$ such that the following hold:

1. For all actions $a \in \mathcal{A}$, $P_a(s^\star, s^\star) = 1$ (and thus, $P_a(s^\star, s') = 0$ for any $s' \neq s^\star$ state of the MDP);
2. For all actions $a \in \mathcal{A}$, $r_a(s^\star) = 0$;
3. The rewards are all nonnegative;
4. For any policy π of the MDP (memoryless or not), and for any $s \in \mathcal{S}$, $\sum_{t \geq 0} \mathbb{P}_s^\pi(S_t \neq s^\star) < \infty$.

In this section we assume that Assumption 1 holds even if this is not explicitly mentioned.

Question 1. Show that the value of any policy π can indeed be "well-defined" in the following sense: Let $(\Omega, \mathcal{F})$ be the measurable space that holds the random variables $(S_t, A_t)_{t \geq 0}$.

1. If we take $R = \sum_{t=0}^{\infty} r_{A_t}(S_t)$, this is well-defined as an *extended real random variable* from the measurable space $(\Omega, \mathcal{F})$ to $(\bar{\mathbb{R}}, \mathbb{B}(\bar{\mathbb{R}}))$ where $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ is the set of *extended reals* and $\mathbb{B}(\bar{\mathbb{R}})$ is the "natural" Borel σ-algebra over $\bar{\mathbb{R}}$ defined using $\mathbb{B}(\bar{\mathbb{R}}) = \sigma(\{[-\infty, x] : x \in \bar{\mathbb{R}}\})$ (i.e., the smallest σ-algebra generated by the set system in the argument of σ).

   5 points

2. For any policy π and state $s \in \mathcal{S}$, under $\mathbb{P}_s^\pi$, the expectation of R exists and is finite.

   20 points

Hint: For Part 1, recall the closure properties of the collection of extended real random variables (e.r.r.v.). Start your argument with showing that $r_{A_t}(S_t)$ is a random variable and build up things from there. For Part 2, recall that the expected value of a nonnegative e.r.r.v is equal to the limit of expected values assigned to simple functions below it provided that the limit of these simple functions converges to the e.r.r.v. For Part 2, see Prop 2.3.2 and for Part 1 see Prop 2.1.5 in (for example) this book here.[1]

Total: **25 points**

Solution. Part 1: By definition, S_t, A_t are random variables. By a slight abuse of notation, let the map $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be defined by $r(s,a) = r_a(s)$. Taking the discrete topology on $\mathcal{S} \times \mathcal{A}$ makes r a continuous map (any map from a discrete topological space to any other topological space is continuous). Hence, $r_{A_t}(S_t)$ is a random variable. Since the sum of finitely many random variables is again a random variable, it follows that for any $t \geq 0$, $R_t = \sum_{s=0}^{t} r_{A_t}(S_t)$ is also a random variable. Finally, $R = \sup_{t \geq 0} R_t$, since R_t is a nondecreasing sequence by our assumptions on the rewards. Since R is the supremum of a countable collection of random variables, it is also a random variable.

Part 2: Fix policy π and state $s \in \mathcal{S}$. Since R is a nonnegative extended real random variable, its expectation is well-defined. Furthermore, the expected value can be obtained by taking *any* sequence of simple functions f_n from $(\Omega, \mathcal{F})$ to the set of reals that approaches R from below and calculating $\lim_{n\to\infty} E_s^\pi[f_n]$. We choose $f_n = R_n$. To argue that R_n is a simple function write it as

$$R_n = \sum_{t=0}^{n} \sum_{s \neq s^\star} \sum_{a} \mathbb{I}_{\{U_{t,s,a}\}} r_a(s)\,, \tag{1}$$

where for $t \geq 0$, $(s,a) \in \mathcal{S} \times \mathcal{A}$, $U_{t,s,a} = \{S_t = s, A_t = a\}$. Here, we restrict the sum to $s \neq s^\star$: Despite this, the equality holds since $r_a(s^\star) = 0$ for any action $a \in \mathcal{A}$. This restriction is useful for the next step.

Now, as noted before, since the rewards are nonnegative, $R_n \leq R$ and thus $\sup_n R_n = \lim_{n\to\infty} R_n = R$. Hence,

$$\mathbb{E}_s^\pi[R] = \lim_{n\to\infty} \mathbb{E}_s^\pi[R_n]\,.$$

Thus, to show that $\mathbb{E}_s^\pi[R]$ is finite, it suffices to show that the limit on the right-hand side is finite. From (1),

$$\begin{aligned}
\mathbb{E}_s^\pi[R_n] &= \sum_{t=0}^{n} \sum_{s \neq s^\star} \sum_{a} r_a(s) \mathbb{P}_s^\pi(S_t = s, A_t = a) \\
&\leq \max_{s,a} r_a(s) \sum_{t=0}^{n} \sum_{s \neq s^\star} \mathbb{P}_s^\pi(S_t = s) \\
&= \max_{s,a} r_a(s) \sum_{t=0}^{n} \mathbb{P}_s^\pi(S_t \neq s^\star) \\
&\leq \max_{s,a} r_a(s) \sum_{t=0}^{\infty} \mathbb{P}_s^\pi(S_t \neq s^\star) < +\infty\,,
\end{aligned}$$

where the last inequality used that, by assumption, $\sum_{t=0}^{\infty} \mathbb{P}_s^\pi(S_t \neq s^\star) < \infty$. □

The last part of the previous problem allows us to define the value of π in state s using the usual formula

$$v^\pi(s) = \mathbb{E}_s^\pi[R]$$

and note that regardless of π and s, these values are always finite.

For a memoryless policy π and $s, s' \neq s^\star$, define $P_\pi(s,s') = \sum_{a \in \mathcal{A}} \pi(a|s) P_a(s,s')$, i.e., the usual way. We can also view P_π, as usual, an $(\mathrm{S}-1) \times (\mathrm{S}-1)$ matrix by identifying $\mathcal{S}$ with $\{1, \ldots, \mathrm{S}\}$, $s^\star = \mathrm{S}$.

[1] Krishna B. Athreya and Soumendra N. Lahiri. Measure Theory and Probability Theory. Springer, 2006.

Question 2 (Transition matrices)**.** Show that for any $s, s' \in \mathcal{S}$, $s, s' \neq s^\star$, and $t \geq 1$, $(P_\pi^t)_{s,s'} = \mathbb{P}_s^\pi(S_t = s')$.

Total: **10 points**

Solution. The solution is the same as the solution of Q2 on Assignment 1 with a few changes, which are marked by red. 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 $P_\pi(S_t, s')$ when $S_t \neq s^\star$. When $S_t = s^\star$, $P_{A_t}(S_t, s') = 0$. Plugging this in, using the law of total expectations,

$$\mathbb{E}[P_\pi(S_t, s')] = \sum_{s \neq s^\star} P_\pi(s, s') \mathbb{P}(S_t = s) .$$

Putting everything together we see that

$$\mathbb{P}(S_{t+1} = s') = \sum_{s \neq s^\star} P_\pi(s, s') \mathbb{P}(S_t = s)$$

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,

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

□

Question 3. Prove that for any memoryless policy π, defining $r_\pi(s) = \sum_a \pi(a|s) r_a(s)$, as usual, we have $v^\pi = \sum_{t \geq 0} P_\pi^t r_\pi$, where when viewed as vectors, v^π and r_π are restricted to $s \neq s^\star$ (i.e., they are $(\mathrm{S}-1)$-dimensional).
Hint: You may want to reuse the result of the previous exercise.

Total: **10 points**

Solution. Let $s, s' \neq s^\star$, $\mathbb{P} = \mathbb{P}_s^\pi$. 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,

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

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

$$\mathbb{E}[\mathbb{E}[r_{A_t}(S_t)|H_t]] = \sum_{s'\neq s^\star} \mathbb{P}(S_t = s') r_\pi(s')\,,$$

where we used that for $s' = s^\star$, $r_\pi(s') = 0$, hence the sum can be restricted to $s \neq s^\star$. Since for $t = 0$, $\mathbb{P}(S_t = s') = 1$ iff $s' = s$, this together with the result of the previous problem gives that

$$v^\pi(s) = \sum_{t\geq 0}\sum_{s'\neq s^\star} P_\pi^t(s,s') r_\pi(s')$$

(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

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

□

Question 4 (Policy evaluation fixed-point equation)**.** Show that for $s \neq s^\star$, v^π satisfies

$$v^\pi(s) = r_\pi(s) + \sum_{s'\neq s^\star} P_\pi(s,s') v^\pi(s')\,.$$

Total: **2 points**

Solution. By the previous problem $v^\pi = r_\pi + \sum_{t\geq 1} P_\pi^t r_\pi = r_\pi + P_\pi(\sum_{t\geq 0} P_\pi^t r_\pi) = r_\pi + P_\pi v^\pi$. □

Define now the $w(s)$ as the total expected reward incurred under π when it is started from s and *in each time step the reward incurred is one* until $s^\star$ is reached (that is, $r_a(s)$ is replaced by 1 for $s \neq s^\star$, while the zero rewards are kept at $s^\star$). By our previous result, w is well-defined. Furthermore,

$$w(s) \geq 1\,, \qquad s \neq s^\star$$

as for $s \neq s^\star$, in the zeroth period, a reward of one is incurred and in all subsequent periods the rewards incurred are nonnegative.

Introduce now the weighted norm, $\|\cdot\|_w$: For $x \in \mathbb{R}^{\mathrm{S}-1}$,

$$\|x\|_w = \max_{s\in[\mathrm{S}-1]} \frac{|x_s|}{w(s)}\,.$$

When the dependence on π is important, we will use w_π.

Question 5 (Contractions)**.** Show that P_π is a contraction under $\|\cdot\|_w$, that is, there exists $0 \leq \rho < 1$ such that for any $x, y \in \mathbb{R}^{\mathrm{S}-1}$,

$$\|P_\pi x - P_\pi y\|_w \leq \rho \|x - y\|_w\,.$$

Total: **15 points**

Solution. Since P_π is linear, $P_\pi x - P_\pi y = P_\pi(x-y)$. Hence, it suffices to show that for any $u \in \mathbb{R}^{\mathrm{S}-1}$,

$$\|P_\pi u\|_w \le \rho \|u\|_w \,.$$

Fix any $u \in \mathbb{R}^{\mathrm{S}-1}$ and $i \in [\mathrm{S}-1]$. Then, using that $w \ge 1$ and hence is positive,

$$\left|\frac{(P_\pi u)(i)}{w(i)}\right| \le \frac{1}{w(i)} \sum_{j=1}^{\mathrm{S}-1} P_\pi(i,j) w(j) \left|\frac{u(j)}{w(j)}\right| \le \|u\|_w \frac{\sum_{j=1}^{\mathrm{S}-1} P_\pi(i,j) w(j)}{w(i)} \,. \tag{2}$$

Now, recall that by the solution to Question 4,

$$w(i) = 1 + \sum_{j=1}^{\mathrm{S}-1} P_\pi(i,j) w(j) \,.$$

Hence,

$$\frac{\sum_{j=1}^{\mathrm{S}-1} P_\pi(i,j) w(j)}{w(i)} = \frac{w(i)-1}{w(i)} \le \max_{1 \le i \le \mathrm{S}-1} \frac{w(i)-1}{w(i)} =: \rho \,.$$

and thanks to $0 \ge w(i) - 1 < w(i)$ and since S is finite, $0 \le \rho < 1$. Plugging this into (2), we get

$$\|P_\pi u\|_w = \max_i \left|\frac{(P_\pi u)(i)}{w(i)}\right| \le \rho \|u\|_w$$

finishing the proof. □

We can define occupancy measures as before: For $s \ne s^\star$, policy π and initial state distribution μ defined over $s^\star \notin \mathcal{S}' := \{1, \dots, \mathrm{S}-1\}$,

$$\nu_\mu^\pi(s,a) = \sum_{t=0}^{\infty} \mathbb{P}_\mu^\pi(S_t = s, A_t = a).$$

Clearly, this is well-defined under our standing assumption (by Question 1). Noting that rewards from $s^\star$ are all zero, we have

$$v^\pi(\mu) = \langle \nu_\mu^\pi, r \rangle \,.$$

Question 6. Show that for any policy π and distribution $\mu \in \mathcal{M}_1(\mathcal{S}')$ there is a memoryless policy π' such that $\nu_\mu^\pi = \nu_\mu^{\pi'}$.

Total: **10 points**

Solution. The solution is the almost verbatim copy of the solution to Question 8 on Assignment 1. Discount factors need to be dropped. Other changes in the solution are indicated by red. For arbitrary μ, π, $s \in \mathcal{S}'$, 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,

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

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

$$\tilde{\nu}_\mu^\pi = \mu \sum_{t \geq 0} P_\pi^t \,. \tag{3}$$

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: For $s \in \mathcal{S}'$,

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

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

$$\tilde{\nu}_\mu^\pi = \mu + \tilde{\nu}_\mu^\pi P_{\pi'}, \tag{4}$$

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 occupancy measure of π' over the states. Indeed, by Eq. (3), $\tilde{\nu}_\mu^{\pi'}$ is indeed a solution and by the solution to Question 5, since P_π' is a contraction, by Banach's fixed point theorem, it is also a unique solution. Thus, $\tilde{\nu}_\mu^\pi = \tilde{\nu}_\mu^{\pi'}$ and thus, for $s \neq s^\star$,

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

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 (4) holds. For this, we have

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

which is equivalent to (4). Here, the last equality follows from the definition of $P_{\pi'}$. □

Define $v^*(s) = \sup_\pi v^\pi(s)$ and define $T : \mathbb{R}^{\mathrm{S}-1} \to \mathbb{R}^{\mathrm{S}-1}$ by $(Tv)(s) = \max_a r_a(s) + \langle P_a(s), v\rangle$, $s \neq s^\star$. For a memoryless policy, we also let $T_\pi v = r_\pi + P_\pi v$ (using vector notation). Greediness is defined as usual: π is greedy w.r.t. $v \in \mathbb{R}^{\mathrm{S}-1}$, if $T_\pi v = Tv$.

Question 7 (The Fundamental Theorem for Undiscounted Infinite-Horizon MDPs)**.** Show that the fundamental theorem still holds:

1. The optimal value function v^* is well-defined (i.e., finite);

20 points

2. Any policy that is greedy with respect to v^* is optimal: $v^\pi = v^*$;

3. It holds that $v^* = Tv^*$.

10 points

Total: **30 points**

Solution. By Question 6, for any policy π and a start state distribution $\mu \in \mathcal{M}_1(\{1, \dots, \mathrm{S}-1\})$, there exists a memoryless policy π' such that

$$\nu_\mu^{\pi'} = \nu_\mu^\pi .$$

We copy the proof given in Lecture 2, with modifications, shown, again in red. The proof would be easy if we only considered memoryless policies when defining v^*. In particular, letting ML stand for the set of memoryless policies of the given MDP, define

$$\tilde{v}^*(s) = \sup_{\pi \in \mathrm{ML}} v^\pi(s) \quad \text{for all } s \in \mathcal{S} \,.$$

Because the supremum is over a smaller set, $\tilde{v}^* \le v^*$.

As we shall see soon, it is not hard to show the theorem just with v^* replaced everywhere with $\tilde{v}^*$. That is:

1. $\tilde{v}^*$ is well-defined;
2. Any policy π that is greedy with respect to $\tilde{v}^*$ satisfies $v^\pi = \tilde{v}^*$;
3. It holds that $\tilde{v}^* = T\tilde{v}^*$.

This is what we will show in Part 1 of the proof, while in Part 2 we will show that $\tilde{v}^* \ge v^*$ and thus v^* is also well-defined and $\tilde{v}^* = v^*$. Clearly, the two parts together establish the desired result.

Part 1: We start by establishing that $\tilde{v}^*$ is well-defined. For this, fix $s \in \mathcal{S}$. We want to show that $\tilde{v}^*(s) < \infty$. Let $(\pi_k)_k$ be a sequence of memoryless policies such that $\lim_{k\to\infty} v^{\pi_k}(s) = \tilde{v}^*(s)$ (which could be infinite). By possibly considering a subsequence, we may assume that $(\pi_k)_k$ itself is convergent. This is because we can view $\pi_k \in \Delta_1(\mathcal{A})^{\mathcal{S}}$, where Δ_1 is the set of probability vectors over $\mathcal{A}$ and thus π_k takes values in a compact subset of a Euclidean space, hence, it has a convergent subsequence. Now, let $\pi(a|s) = \lim_{k\to\infty} \pi_k(a|s)$, $(s,a) \in \mathcal{S} \times \mathcal{A}$ (the pointwise limit of π_k). Then, by Question 1, $v^\pi(s) < +\infty$. We now claim that $v^{\pi_k}(s) \to v^\pi(s)$ as $k \to \infty$. Indeed, $v^{\pi_k} - v^\pi = (I - P_{\pi_k})^{-1}[T_{\pi_k} v^\pi - v^\pi]$ and $T_{\pi_k} v^\pi - v^\pi = T_{\pi_k} v^\pi - T_\pi v^\pi = r_{\pi_k} - r_\pi + (P_{\pi_k} - P_\pi) v^\pi$. Let $w = w_\pi$. Then, $\|v^{\pi_k} - v^\pi\|_w = \|(I - P_{\pi_k})^{-1}\|_w \, (\|r_{\pi_k} - r_\pi\|_w + \|P_{\pi_k} - P_\pi\|_w \|v^\pi\|_w) \to 0$ as $k \to \infty$ since, by a calculation similar to that given in the solution of Question 5, one can show that for k large enough, $\|P_{\pi_k}\|_w \le \frac{1+\rho}{2} < 1$ where ρ is the contraction coefficient for P_π defined in the solution of that problem, and hence for k large enough, $\|(I - P_{\pi_k})^{-1}\|_w = \|\sum_{t\ge 0} (P_{\pi_k})^t\|_w \le \frac{1}{1-\frac{1+\rho}{2}} = \frac{2}{1-\rho} < \infty$ and $\|r_{\pi_k} - r_\pi\|_w \to 0$ and $\|P_{\pi_k} - P_\pi\|_w \to 0$ by the continuity of $\|\cdot\|_w$ while $\|v^\pi\|_w < \infty$ because v^π is finite valued and there are finitely many states. Hence, $\tilde{v}^*(s) = \lim_{k\to\infty} v^{\pi_k}(s) = v^\pi(s) < +\infty$.

The idea of the proof then is to first show that

$$\tilde{v}^* \le T\tilde{v}^* \tag{5}$$

and then show that for any greedy policy π, $v^\pi \ge \tilde{v}^*$.

The displayed equation follows by noticing that $v^\pi \le \tilde{v}^*$ holds for all memoryless policies π by definition. Applying T_π on both sides, using $v^\pi = T_\pi v^\pi$, we get $v^\pi \le T_\pi \tilde{v}^*$. Taking the supremum of both sides over π and noticing that $Tv = \sup_{\pi \in \mathrm{ML}} T_\pi v$ for any v, together with the definition of $\tilde{v}^*$ gives (5).

Now, take any memoryless policy π that is greedy w.r.t. $\tilde{v}^*$. Thus, $T_\pi \tilde{v}^* = T\tilde{v}^*$.

Combined with (5), we get

$$T_\pi \tilde{v}^* \ge \tilde{v}^* \,. \tag{6}$$

Applying T_π on both sides and noticing that T_π keeps the inequality intact (i.e., for any u, v such that $u \le v$ we get $T_\pi u \le T_\pi v$), we get

$$T_\pi^2 \tilde{v}^* \ge T_\pi \tilde{v}^* \ge \tilde{v}^* \,,$$

where the last inequality follows from (6). With the same reasoning we get that for any $k \ge 0$,

$$T_\pi^k \tilde{v}^* \ge T_\pi^{k-1} \tilde{v}^* \ge \dots \ge \tilde{v}^* \,,$$

Now, by the solution to Question 5, T_π is easily seen to be a weighted norm-contraction with an appropriate weighted norm defined in that problem and thus the fixed-point iteration $T_\pi^k \tilde{v}^*$ converges to v^π. Hence, taking the limit above, we get

$$v^\pi \ge \tilde{v}^*.$$

This, together with $v^\pi \le \tilde{v}^*$ shows that $v^\pi = \tilde{v}^*$. Finally, $T\tilde{v}^* = T_\pi \tilde{v}^* = T_\pi v^\pi = v^\pi = \tilde{v}^*$.
Part 2: It remains to be shown that $\tilde{v}^* = v^*$. Let Π be the set of all policies. Because $\mathrm{ML} \subset \Pi$, $\tilde{v}^* \le v^*$. Thus, it remains to show that

$$v^* \le \tilde{v}^* \,. \tag{7}$$

To show this, we will use that by Question 6, for any state-distribution $\mu \in \mathcal{M}_1(\mathcal{S}')$ and policy π (memoryless or not) we can find a memoryless policy, which we will call for now $\mathrm{ML}(\pi)$, such that $\nu_\mu^\pi = \nu_\mu^{\mathrm{ML}}$. Fix a state $s \in \mathcal{S}'$. Applying this result with $\mu = \delta_s$ with $s \in \mathcal{S}'$, we get

$$\begin{aligned} v^\pi(s) &= \langle \nu_s^\pi, r \rangle \\ &= \langle \nu_s^{\mathrm{ML}(\pi)}, r \rangle \\ &\le \sup_{\pi' \in \mathrm{ML}} \langle \nu_s^{\pi'}, r \rangle \\ &= \sup_{\pi' \in \mathrm{ML}} v^{\pi'}(s) = \tilde{v}^*(s) \,. \end{aligned}$$

Taking the supremum of both sides over π, we get $v^*(s) = \sup_{\pi \in \Pi} v^\pi(s) \le \tilde{v}^*(s)$. Since $s \in \mathcal{S}'$ was arbitrary and for $s = s^\star$, $v^*(s^\star) = v^\pi(s^\star)$ for any policy π, we get $v^* \le \tilde{v}^*$, finishing the proof.

□

Question 8. Imagine that Assumption 1 is changed such that all immediate rewards are nonpositive (at $s^\star$ the rewards are still zero). What do you need to change in your answer to the previous questions? Just give a short summary of the changes.

Total: **3 points**

Solution. Recall that expectations of nonpositive random variables are defined through taking their negation. Hence, we need to consider $-R$, but this brings us back to the nonnegative case. Nothing else changes. □

Question 9. Imagine that Assumption 1 is changed such that there is no sign restriction on the rewards, they can be positive, or negative. Something will go wrong with the claims made in Question 1. Explain what.

Total: **3 points**

Solution. A simple example is when there are two actions, $\mathcal{A} = \{1, 2\}$ and for some state $s \ne s^\star$, $r_1(s) = +1$ and $r_2(s) = -1$. Then, R is not well-defined on the event $\{S_0 = s, A_0 = 1, S_1 = s, A_1 = 2, S_2 = s, A_2 = 1, \dots\}$, that is, when the actions are alternating between action one and action two. □

Approximate Policy Iteration

Question 10. In the context of the analysis of approximate policy iteration analysis it was suggested that the following identity holds:

$$P_{\pi'} - P_{\pi^*} + \gamma P_{\pi'}(I - \gamma P_{\pi'})^{-1}(P_{\pi'} - P_{\pi}) = P_{\pi'}(I - \gamma P_{\pi'})^{-1}(I - \gamma P_{\pi}) - P_{\pi^*} \,.$$

Show that this identity holds, actually, regardless the choice of the memoryless policies π, π' and π^*.

Total: **10 points**

Solution. For an arbitrary memoryless policy π introduce the notation $A_\pi = (I - \gamma P_\pi)$. We have

$$\begin{aligned}
&P_{\pi'} - P_{\pi^*} + \gamma P_{\pi'}(I - \gamma P_{\pi'})^{-1}(P_{\pi'} - P_{\pi}) \\
&\quad = P_{\pi'} - P_{\pi^*} + \gamma P_{\pi'} A_{\pi'}(P_{\pi'} - P_{\pi}) \\
&\quad = P_{\pi'} - P_{\pi^*} + \gamma P_{\pi'} A_{\pi'}(P_{\pi'} - P_{\pi}) A_\pi A_\pi^{-1} \\
&\quad = P_{\pi'} - P_{\pi^*} + P_{\pi'}(A_{\pi'} - A_\pi) A_\pi^{-1} \\
&\quad = P_{\pi'} - P_{\pi^*} + P_{\pi'} A_{\pi'} A_\pi^{-1} - P_{\pi'} A_\pi A_\pi^{-1} \\
&\quad = P_{\pi'} - P_{\pi^*} + P_{\pi'} A_{\pi'} A_\pi^{-1} - P_{\pi'} \\
&\quad = P_{\pi'} A_{\pi'} A_\pi^{-1} - P_{\pi^*} \,.
\end{aligned}$$

Here, the third equality comes from that

$$A_\pi^{-1} - A_{\pi'}^{-1} = (I - \gamma P_\pi) - (I - \gamma P_{\pi'}) = \gamma(P_{\pi'} - P_\pi) \,,$$

and thus, multiplying from the right with A_π and multiplying from the left with $A_{\pi'}$ we get

$$A_{\pi'} - A_\pi = \gamma A_{\pi'}(P_{\pi'} - P_\pi) A_\pi \,.$$

An alternate solution starts with noting that for any memoryless policy π,

$$\gamma P_\pi A_\pi = \sum_{i \ge 1} (\gamma P_\pi)^i = A_\pi - I \,.$$

Hence,

$$\begin{aligned}
P_{\pi'} + \gamma P_{\pi'} A_{\pi'}(P_{\pi'} - P_\pi) &= P_{\pi'} + (A_{\pi'} - I)(P_{\pi'} - P_\pi) \\
&= P_{\pi'} + A_{\pi'}(P_{\pi'} - P_\pi) - (P_{\pi'} - P_\pi) \\
&= A_{\pi'}(P_{\pi'} - P_\pi) + P_\pi \,.
\end{aligned}$$

Subtracting P_{π^*} from both sides gives the result. □

Question 11. Prove the following. Assume that the rewards lie in the $[0, 1]$ interval. Let $(\pi_k)_{k \ge 0}$ be a sequence of memoryless policies and $(q_k)_{k \ge 0}$ be a sequence of functions over the set of state-action pairs such that for $k \ge 1$, π_k is greedy with respect to q_{k-1}. Further, let $\varepsilon_k = \max_{0 \le i \le k} \|q^{\pi_i} - q_i\|_\infty$. Then, for any $k \ge 1$,

$$\|q^* - q^{\pi_k}\|_\infty \le \frac{\gamma^k}{1-\gamma} + \frac{2\gamma}{(1-\gamma)^2}\varepsilon_{k-1} \,,$$

and policy π_{k+1} is δ-optimal where

$$\delta \le \frac{2}{1-\gamma}\left(\frac{\gamma^k}{1-\gamma} + \frac{2}{(1-\gamma)^2}\varepsilon_k\right) \,.$$

How does this result compare to the Approximate Policy Iteration Corollary from Lecture 8 notes?
Hint: You can use the following geometric progress lemma for action-value functions without proof.

$$\|q^* - q^{\pi_k}\|_\infty \le \gamma \|q^* - q^{\pi_{k-1}}\|_\infty + \frac{2\gamma}{1-\gamma}\|q^{\pi_{k-1}} - q_{k-1}\|_\infty \,.$$

Total: **15 points**

Solution. From the geometric progress lemma we have

$$\|q^* - q^{\pi_k}\|_\infty \le \gamma \|q^* - q^{\pi_{k-1}}\|_\infty + \frac{2\gamma}{1-\gamma}\|q^{\pi_{k-1}} - q_{k-1}\|_\infty \,.$$

Iterating this gives

$$\begin{aligned}\|q^* - q^{\pi_k}\|_\infty &\le \gamma^k \|q^* - q^{\pi_0}\|_\infty + \frac{2\gamma}{(1-\gamma)^2}\max_{0\le i\le k-1}\|q^{\pi_i} - q_i\|_\infty \\ &\le \frac{\gamma^k}{1-\gamma} + \frac{2\gamma}{(1-\gamma)^2}\varepsilon_{k-1}\,,\end{aligned}$$

where in the last inequality we used the definition of ε_{k-1} and that both q^* and q^{π_0} take values in $[0, 1/(1-\gamma)]$, hence $\|q^* - q^{\pi_0}\|_\infty \le 1/(1-\gamma)$.

To get the second result note that policy π_{k+1} is greedy with respect to q_k. Hence, we first bound the distance between q_k and q^*:

$$\begin{aligned}\|q^* - q_k\|_\infty &\le \|q^* - q^{\pi_k}\|_\infty + \|q^{\pi_k} - q_k\|_\infty \\ &\le \frac{\gamma^k}{1-\gamma} + \frac{2\gamma}{(1-\gamma)^2}\varepsilon_{k-1} + \varepsilon_k \\ &\le \frac{\gamma^k}{1-\gamma} + \frac{2}{(1-\gamma)^2}\varepsilon_k\,,\end{aligned}$$

where the last inequality used that $\varepsilon_k \ge \varepsilon_{k-1}$ and $\frac{2\gamma}{(1-\gamma)^2} + 1 = \frac{2\gamma+1-2\gamma+\gamma^2}{(1-\gamma)^2} \le \frac{2}{1-\gamma}$.

Now, by the "policy error bound I." from Lecture 6, π_{k+1} is δ-optimal with

$$\delta \le \frac{2\|q_k - q^*\|_\infty}{1-\gamma} \le \frac{2}{1-\gamma}\left(\frac{\gamma^k}{1-\gamma} + \frac{2}{(1-\gamma)^2}\varepsilon_k\right)\,.$$

□

Total for all questions: 133. Of this, 23 are bonus marks (i.e., 110 marks worth 100% on this problem set).