CMPUT 605: Theoretical Foundations of Reinforcement Learning, Winter 2023
Midterm

Instructions

Submissions You need to submit a single PDF file, named `midterm_<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. 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 26 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.
Note: You may find it useful to resuse results from previous assignments and the lecture notes. If you believe a solution to any of the questions below is very similar to a previous assignment solution or proof in the lecture notes, you do not need to rewrite the entire solution. It is sufficient to indicate only what would need to be changed for the solution to hold in the setting defined above (i.e. under Assumption 1). When referring to an assignment solution, please refer to the solution PDF shared with you after the assignment deadline (named assignmentX_soln.pdf where X is the assignment number), not your personal solutions to the assignments.

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**

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 $\{1, \ldots, \mathrm{S}\}$, $s^\star = \mathrm{S}$.

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**

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**

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**

---

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

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,

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

since 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**

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

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

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**

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**

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**

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**

Approximate Policy Iteration

Question 10. Prove the Theorem(Approximate Policy Iteration) from lecture notes 8. Assume that the rewards lie in the $[0, 1]$ interval. Let $(\pi_k)_{k\geq 0}$, $(\varepsilon_k)_k$ be such that

$$Tv^{\pi_k} = T_{\pi_{k+1}} v^{\pi_k} + \varepsilon_k$$

holds for all $k \geq 0$ Then, for any $k \geq 1$,

$$\|v^* - v^{\pi_k}\|_\infty \leq \frac{\gamma^k}{1-\gamma} + \frac{1}{(1-\gamma)^2} \max_{0\leq s\leq k-1} \|\varepsilon_s\|_\infty.$$

Hint: You should make use of Lemma(Geometric progress lemma with approximate policy improvement) from lecture notes 8.

Total: **10 points**

Policy Gradients and Stationary Points

Question 11. In this question, we will show that there are no suboptimal stationary points in policy search, as long as there are no suboptimal stationary points on the corresponding policy improvement objective. For each $\theta \in \mathbb{R}^d$, we let $\pi_\theta : \mathcal{S} \to \mathcal{M}_1(\mathcal{A})$ be a memoryless policy and $\Pi = \{\pi_\theta : \theta \in \mathbb{R}^d\}$ the set of parametrized policies. As usual, π^* denotes the optimal policy, i.e. $v^{\pi^*} = v^*$. Note that we do not assume explicitly that $\pi^* \in \Pi$. For an initial state distribution $\mu \in \mathcal{M}_1(\mathcal{S})$, the return is

$$J(\pi) = \mu^\top v^\pi \tag{1}$$

We assume that $\theta \mapsto J(\pi_\theta)$ is differentiable. Recall that $\theta_0 \in \mathbb{R}^d$ is a stationary point of $J(\pi_\theta)$ if

$$\frac{d}{dx} J(\pi_x)|_{x=\theta_0} = \mathbf{0} \in \mathbb{R}^d\,. \tag{2}$$

Let $\theta_0 \in \mathbb{R}^d$ and assume that the following conditions hold:

i) The policy gradient theorem holds, i.e. for all $\theta \in \mathbb{R}^d$, $\frac{d}{dx} J(\pi_x)|_{x=\theta} = \tilde{\nu}_\mu^{\pi_\theta} \frac{d}{dx} M_{\pi_x} q^{\pi_\theta}|_{x=\theta}$, where $\tilde{\nu}_\mu^{\pi_\theta}$ is the discounted state occupancy measure for policy π_θ and initial state distribution μ.

ii) The policy class is closed under policy iteration, i.e. for all $\theta \in \mathbb{R}^d$, $\max_{\pi \in \Pi} \tilde{\nu}_\mu^{\pi_\theta} M_\pi q^{\pi_\theta} = \max_{\pi \in \mathrm{ML}} \tilde{\nu}_\mu^{\pi_\theta} M_\pi q^{\pi_\theta}$, where ML is the set of memoryless policies.

iii) For each $\theta \in \mathbb{R}^d$, the map $x \mapsto \tilde{\nu}_\mu^{\pi_\theta} M_{\pi_x} q^{\pi_\theta}$ has no suboptimal stationary points.

iv) The occupancy measure of π^* is absolutely continuous w.r.t. the occupancy measure of π_{θ_0}: $\tilde{\nu}_\mu^{\pi^*} \ll \tilde{\nu}_\mu^{\pi_{\theta_0}}$ (the definition of absolutely continuous is that $\tilde{\nu}_\mu^{\pi_{\theta_0}}(s) = 0 \implies \nu_\mu^{\pi^*}(s) = 0$).

In the following, we show that θ_0 is a stationary point of $J(\pi_\theta)$ if and only if $J(\pi_\theta) = J(\pi^*)$.

1. Show that $J(\pi_{\theta_0}) = J(\pi^*)$ implies that θ_0 is a stationary point.

5 points

2. For the other direction, assume that θ_0 is a stationary point of $J(\pi_\theta)$. Use assumptions i)-iii) to show that $\tilde{\nu}_\mu^{\pi_{\theta_0}} v^{\pi_{\theta_0}} = \tilde{\nu}_\mu^{\pi_{\theta_0}} T v^{\pi_{\theta_0}}$ (i.e. θ_0 satisfies an "average" Bellman optimality equation). Conclude the proof using the performance difference lemma and assumption iv).

25 points

Total: **30 points**

Total for all questions: 148. Of this, 28 are bonus marks (i.e., 120 marks worth 100% on this problem set).