Homework 0:
CS 4789/5789: Introduction to Reinforcement Learning

Cornell University

1 Policies *[0 points]*

Please read these policies. **Please answer the three questions below and include your answers marked in a "problem 1" in your solution set.** Homeworks which do not include these answers will not be graded.

Homework Scoring: All bonus questions on homeworks are worth a maximum of 5 points. Note bonus points can only be used to get you a maximum of a 100% on the assignment.

Gradescope submission: When submitting your HW, please tag your pages correctly as is requested in gradescope. Untagged homeworks will not be graded, until the tagging is fixed.

Readings: Read the notes and required material.

Submission format: Submit your report as a *single* pdf file. Please typeset your writing using LaTeX.

Collaboration: It is acceptable for you to discuss problems with other students; it is not acceptable for students to look at another students written answers. Each student must understand, write, and hand in their own answers.

Acknowledgments: If students find out solutions in published material, on the web, or from other textbooks, this must be acknowledged. If students find proofs in existing papers, it is ok to use these for guidance; students must acknowledge this, and students should first make an attempt at the answer on their own. All students must understand all the written steps that they write.

1.1 List of Collaborators

List the names of all people you have collaborated with and for which question(s).

1.2 List of Acknowledgements

If you find an assignment's answer or use a another source for help, acknowledge for which question and provide an appropriate citation (there is no penalty, provided you include the acknowledgement). If not, then write "none".

1.3 Certify that you have read the instructions

Write "I have read these policies" to certify this.

2 Certify that you have read the website *[0 points]*

Please read the course policies on the website (up until the Lecture Notes section) and write“I have read the course policies on the website”. It is your responsibility to understand and follow these policies. If further clarification is needed, please post to the discussion board.

3 Bellman Optimality (25 points)

In the class, we studied Bellman optimality for Value function, i.e., $V : \mathcal{S} \mapsto \mathbb{R}$. In this section, we will first derive similar Bellman optimality for Q functions, i.e., $Q : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$.

We focus on discounted infinite horizon MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \gamma, P, r\}$. Recall the definition of the optimal policy $\pi^\star$, and $V^\star, Q^\star$.

3.1 Bellman Optimality for $Q^\star$

Q (10 points): Prove the following equalities:

$$Q^\star(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q^\star(s',a'), \forall s, a.$$

3.2 Bellman Optimality for Q

Q (15 points): Consider a Q function that satisfies the following equalities:

$$Q(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q(s',a'), \forall s, a.$$

Prove that $Q(s,a) = Q^\star(s,a), \forall s, a$.

4 Distributions induced by Policies and MDPs (40 points)

Then, we will study some properties of the state (action) distributions induced by the policy. Again we focus on the discounted infinite horizon MDP.

Denote s_0 as some state from $\mathcal{S}$. Recall that $\mathbb{P}_h^\pi(s,a;s_0)$ is the probability of hitting state-action (s,a) at time step h by following policy π starting from s_0. Recall that we have average discounted state-action distribution defined as $d^\pi(s,a;s_0) = (1-\gamma)\sum_{h=0}^\infty \gamma^h \mathbb{P}_h^\pi(s,a;s_0)$

4.1 Value Function

Q (5 points) Prove that $V^\pi(s_0) = \frac{1}{1-\gamma}\sum_{s,a} d^\pi(s,a;s_0) r(s,a)$.

4.2 Valid Distribution

Q (5 points) Prove that $d^\pi(\cdot,\cdot;s_0)$ is a valid distribution. Namely prove that $d^\pi(s,a;s_0) \geq 0$ for all s, a, and $\sum_{s\in\mathcal{S}, a\in\mathcal{A}} d^\pi(s,a;s_0) = 1$.

4.3 Generalizing to Initial State Distribution

Recall that $d^\pi(s, a; s_0)$ is defined assuming that we always start from s_0 at $h = 0$. Now let us generalize this s_0 to a fixed state distribution μ_0 which is an arbitrary distribution over states $\mathcal{S}$. Let us denote $\mathbb{P}_h^\pi(s, a; \mu_0)$ as the probability of π hitting s, a at time step h with s_0 at $h = 0$ being sampled from μ_0. Denote $d_{\mu_0}^\pi(s, a) = (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s, a; \mu_0)$, i.e., the discounted average state-action distribution of π with μ_0 as the initial state distribution.

Q (10 points) Prove that:

$$d_{\mu_0}^\pi(s, a) = \mathbb{E}_{s_0 \sim \mu_0} d^\pi(s, a; s_0).$$

Hint: use Bayes rule: $P(x, y) = P(x)P(y|x)$.

4.4 Recursion in Distributions

Given $d_{\mu_0}^\pi(s, a)$, let us denote $d_{\mu_0}^\pi(s)$ as the state distribution, i.e., $d_{\mu_0}^\pi(s) = \sum_{a \in \mathcal{A}} d_{\mu_0}^\pi(s, a)$. Note that $d_{\mu_0}^\pi(s)$ can be understood as the (discounted) average probability of π hitting s with μ_0 as the initial state distribution.

Q (20 points) Prove the following equality:

$$d_{\mu_0}^\pi(s') = (1-\gamma)\mu_0(s') + \gamma \sum_{s,a \in \mathcal{S}\times\mathcal{A}} d_{\mu_0}^\pi(s, a) P(s'|s, a), \forall s' \in \mathcal{S}.$$

Hint: What is the probability of visiting s at time step h if I tell you the state-action distribution at time step $h-1$? Namely recall the graphical model and think about how to write $\mathbb{P}_h^\pi(s; \mu_0)$ in terms of $\mathbb{P}_{h-1}^\pi(\cdot, \cdot; \mu_0)$ using the Markov property.

5 The property of $(I - \gamma P)$ in Policy Evaluation (25 points)

Recall that in policy evaluation (given MPD and a policy π) lecture slides, we have $V^\pi = (I - \gamma P)^{-1} R$, where recall the definition of $R \in \mathbb{R}^{|\mathcal{S}|}$ and $P \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ from the lecture slides (note rows of P will be indexed by s, and that the row corresponding to s is $P(\cdot|s, \pi(s))$ and there are in total $|\mathcal{S}|$ many rows). The goal of this problem is to help you recall linear algebra.

We will prove that $(I - \gamma P)$ is full rank, thus invertible. To prove that a matrix $A \in \mathbb{R}^{d\times d}$ is full rank, we will show the following: A is full rank if for all $x \in \mathbb{R}^d$ with $x \neq 0$, we have $Ax \neq \mathbf{0}$, where $\mathbf{0}$ is a d-dim vector with zero everywhere. Essentially this means that the null space of A is empty (please recall these basic concepts like rank, null space, span, etc).

Q (10 points): Recall the definition of P. Prove the following equality: for any $x \in \mathbb{R}^{|\mathcal{S}|}$, we have:

$$\|Px\|_\infty \leq \|x\|_\infty.$$

Q (15 points): Prove the following for any non-zero vector $x \in \mathbb{R}^{|\mathcal{S}|}$,

$$\|(I - \gamma P)x\|_\infty > 0.$$

Hint, first try to prove that triangle inequality holds for ℓ_∞ norm, i.e., for any two vectors x, y we have $\|x - y\|_\infty \geq \|x\|_\infty - \|y\|_\infty$.

6 Basic Inequalities (10 points)

In this section, we are going to prove two inequalities that we used in the lecture for proving Bellman optimality.

Q (5 points): Consider a distribution $P \in \Delta(\mathcal{X})$, and any two functions $f : \mathcal{X} \mapsto \mathbb{R}$ and $g : \mathcal{X} \mapsto \mathbb{R}$. Prove the following inequality:

$$\left|\mathbb{E}_{x \sim P}\left[f(x)\right] - \mathbb{E}_{x \sim P}\left[g(x)\right]\right| \leq \mathbb{E}_{x \sim P}\left|f(x) - g(x)\right|.$$

Q (5 points): Consider two functions $f : \mathcal{X} \mapsto \mathbb{R}$ and $g : \mathcal{X} \mapsto \mathbb{R}$ where for simplicity we assume $\mathcal{X}$ is a discrete set. Prove the following:

$$\left|\max_{x \in \mathcal{X}} f(x) - \max_{x \in \mathcal{X}} g(x)\right| \leq \max_{x \in \mathcal{X}} \left|f(x) - g(x)\right|.$$

The above two proofs together with the steps that we went through during the lecture, formally conclude our proof for Bellman Optimality. You will see that these two inequalities will be used again in the upcoming lectures.