CMPUT 653: Theoretical Foundations of Reinforcement Learning, Winter 2021
Homework #5*

Instructions

Submissions You need to submit a zip file, named `p5*_<name>.zip` or `p5*_<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. 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: April 6 at 11:55 pm

Sufficient condition for policy gradients to exist

In Lecture 16 it is stated that a sufficient condition for the differentiability of $x \mapsto J(\pi_x)$ at $x = \theta_0$ in a finite MDP is that for $x \mapsto \pi_x(a|s)$ is continuously differentiable at $x = \theta_0$ for any (s, a) pair. Our purpose here is to show that this is correct.

We start with a more general sufficient condition that allows for infinite state and action spaces and would be applicable, for example, for LQR problems with the discounted total reward criterion. The next question will be concerned with the (simpler) finite case.

Question 1. Let $M = (\mathcal{S}, \mathcal{A}, P, r)$ be an MDP so that for any memoryless policy v^π exists and is bounded. For $x \in \mathbb{R}^d$ let π_x be a memoryless policy. Let $\mu \in \mathcal{M}_1(\mathcal{S})$ and for a memoryless policy let $J(\pi) = \mu v^\pi$. For $\mu' \in \mathcal{M}_1(\mathcal{S})$ and bounded $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, let $F(\mu', x, q) = \mu' M_{\pi_x} q$. Assume that for any fixed μ', q, $F(\mu', \cdot, q)$ has continuous partial derivatives at $x = \theta_0 \in \mathbb{R}^d$ and that

$$(\mu', q) \mapsto \sum_{i=1}^{d} \frac{\partial}{\partial x_i} F(\mu', x, q)|_{x=\theta_0} e_i$$

is continuous in a neighborhood of $((1-\gamma)\tilde{\nu}_\mu^{\pi_{\theta_0}}, q^{\pi_{\theta_0}})$ (here e_i is the ith vector in the standard Euclidean basis). Then the following hold:

1. $x \mapsto J(\pi_x)$ is differentiable;
2. the conditions of the policy gradient theorem are met.

Total: **50 points**

Solution. We first show the second part. In particular, we show that *(a)* $\theta \mapsto f'_{\pi_\theta}(\theta_0)$ exists and is continuous in a neighborhood of θ_0 and *(b)* $g'_{\pi_{\theta_0}}(\theta_0)$ exists, where

$$f_\pi(x) = \tilde{\nu}_\mu^\pi M_{\pi_x} q^\pi \,,$$
$$g_\pi(x) = \tilde{\nu}_\mu^{\pi_x} v^\pi \,.$$

Then, the policy gradient theorem itself shows that Part 1 is true.

Let us start with Claim (a). Fix a memoryless policy π. Since q^π is bounded and $(1-\gamma)\nu_\mu^\pi \in \mathcal{M}_1(\mathcal{S})$, by our assumption the partial derivatives of $f_\pi(x)$ exist and are continuous. Hence, f_π is differentiable and

$$\frac{d}{dx}\tilde{\nu}_\mu^\pi M_{\pi_x} q^\pi = \sum_{i=1}^d \frac{\partial}{\partial x_i}\tilde{\nu}_\mu^\pi M_{\pi_x} q^\pi e_i \,.$$

Since by assumption the right-hand side is continuous in a neighborhood

For θ in a neighborhood of θ_0, we need that f_{π_θ} is differentiable at $x = \theta_0$. Fix θ.

Claim (a) follows since by our assumption $x \mapsto \tilde{\nu}_\mu^{\pi_{\theta_0}} M_{\pi_x} q^{\pi_{\theta_0}}$ has continuous partial derivatives in a neighborhood of θ_0, hence it is differentiable in a neighborhood of θ_0.

Cs: really?

For brevity let $v = v^{\pi_{\theta_0}}$. We have

$$J(\pi_x) =$$

□

Total for all questions: 50. Of this, 100 are bonus marks (i.e., 100 marks worth 100% on this problem set).