Northeastern University
CS 4180/5180 – Reinforcement Learning and Decision Making
Fall 2019, Chris Amato

Policy Gradients Assignment

Name: _____

Problem	Points
Linear Function Approximators	/20
Missing pseudocode	/20
Semi-Gradient One-Step Expected SARSA	/20
Generalize REINFORCE	/20
Eligibility Vector for Softmax Policy	/20
Eligibility Vector for Gaussian Policy	/20
Total	/120

Instructions

- Don't cheat, write clearly, and staple your sheets!
- Questions about this assignment should be posted on Piazza for all to see.
- Contribute to the Piazza discussions, without providing direct answers.
- Graded out of 120 points for graduate students and 110 for undergrads.

(20 pts.) Linear Function Approximators

Exercise 9.1 in the SB textbook. Show that tabular methods such as presented in Part I of this book are a special case of linear function approx. What would the feature vectors be?

Exercise 9.3 in the SB textbook. What n and $c_{i,j}$ produce the feature vectors $\boldsymbol{x}(s) = (1, s_1, s_2, s_1 s_2, s_1^2, s_2^2, s_1 s_2^2, s_1^2 s_2, s_1^2 s_2^2)^\top$?

(20 pts.) Missing pseudocode

Exercise 10.1 in the SB textbook. We have not explicitly considered or given pseudocode for any Monte Carlo methods in this chapter. What would they be like? Why is it reasonable not to give pseudocode for them? How would they perform on the Mountain Car task?

(20 pts.) Semi-Gradient One-Step Expected SARSA

Exercise 10.2 in the SB textbook. Give pseudocode for semi-gradient one-step *Expected* Sarsa for control.

(20 pts.) Generalize REINFORCE

Exercise 13.2 in the SB textbook. Generalize the box on page 199, the policy gradient theorem (13.5), the proof of the policy gradient theorem (page 325), and the steps leading to the REINFORCE update equation (13.8), so that (13.8) ends up with a factor of γ^t and thus aligns with the general algorithm given in the pseudocode.

(20 pts.) Eligibility Vector for Softmax Policy

Exercise 13.3 in the SB textbook. In Section 13.1 we considered policy parameterizations using the softmax in action preferences (13.2) with linear action preferences (13.3). For this parameterization, prove that the eligibility vector is

$$\nabla \ln \pi(a \mid s, \boldsymbol{\theta}) = \boldsymbol{x}(s,a) - \sum_b \pi(b \mid s, \boldsymbol{\theta}) \boldsymbol{x}(s,b)\,,$$

using the definitions and elementary calculus.

(20 pts.) Eligibility Vector for Gaussian Policy

Exercise 13.4 in the SB textbook. Show that for the gaussian policy parameterization (13.19) the eligibility vector has the following two parts:

$$\nabla \ln \pi(a \mid, \boldsymbol{\theta}_\mu) = \frac{\nabla \pi(a \mid s, \boldsymbol{\theta}_\mu)}{\pi(a \mid s, \boldsymbol{\theta})} = \frac{1}{\sigma(s, \boldsymbol{\theta})^2} \left(a - \mu(s, \boldsymbol{\theta})\right) \boldsymbol{x}_\mu(s) \text{ and}$$

$$\nabla \ln \pi(a \mid, \boldsymbol{\theta}_\sigma) = \frac{\nabla \pi(a \mid s, \boldsymbol{\theta}_\sigma)}{\pi(a \mid s, \boldsymbol{\theta})} = \left(\frac{\left(a - \mu(s, \boldsymbol{\theta})\right)^2}{\sigma(s, \boldsymbol{\theta})^2} - 1 \right) \boldsymbol{x}_\sigma(s) \,.$$