4180/5180: Reinforcement Learning and Sequential Decision Making (Fall 2022) Christopher Amato
Northeastern University Dec. 9th, 2022

Exercise 9: Policy Gradient (PG)

Please remember the following policies:

- Exercise due at **11:59 PM EST Dec. 9, 2022**.
- Submissions should be made electronically on Canvas. Please ensure that your solutions for both the written and programming parts are present. You can upload multiple files in a single submission, or you can zip them into a single file. You can make as many submissions as you wish, but only the latest one will be considered.
- For **Written** questions, solutions may be handwritten or typeset. If you write your answers by hand and submit images/scans of them, please please ensure legibility and order them correctly in a single PDF file.
- The PDF file should also include the figures from the **Plot** questions.
- For both **Plot** and **Code** questions, submit your source code in Jupyter Notebook (.ipynb file) along with reasonable comments of your implementation. Please make sure the code runs correctly.
- You are welcome to discuss these problems with other students in the class, but you must understand and write up the solution and code yourself. Also, you *must* list the names of all those (if any) with whom you discussed your answers at the top of your PDF solutions page.
- Each exercise may be handed in up to two days late (24-hour period), penalized by 10% per day late. Submissions later than two days will not be accepted.
- Contact the teaching staff if there are medical or other extenuating circumstances that we should be aware of.
- **Notations: RL2e is short for the reinforcement learning book 2nd edition. x.x means the Exercise x.x in the book.**

1. **2 points.** (RL2e 13.2) *Generalize REINFORCE*

   Written: 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.

2. **2 points.** (RL2e 13.3) *Eligibility Vector for Softmax Policy*

   Written: In Section 13.1 we considered policy parameterizations using the soft-max in action preferences (13.2) with linear action preferences (13.3). For this parameterization, prove that the eligibility vector is

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

   using the definitions and elementary calculus.

3. **3 points.** (RL2e 13.4) *Eligibility Vector for Gaussian Policy*

   Written: Show that for the gaussian policy parameterization (13.19) the eligibility vector has the following two parts:

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

4. **4 points** *Four Rooms with PG*

   (a) **Code, Plot:** Implement the REINFORCE algorithm (one page 328) using neural networks and test on our favorite environment, Four Rooms, plot your learning curve.

(b) **Code, Plot:** Implement and plot learning curve of REINFORCE with baseline (on page 330) using neural networks.

(c) **Written:** Compare the results of the REINFORCE algorithm and the REINFORCE with baseline algorithm.

We provide the scffolding code in the notebook.