4180/5180: Reinforcement Learning and Sequential Decision Making (Spring 2025) Robert Platt
Northeastern University Apr. 3rd, 2024

Exercise 7: Policy Gradient (PG)

Please remember the following policies:

- This exercise has **extra credit**.
- Exercise due at **11:59 PM EST Apr. 3rd, 2024**.
- 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 should be typeset.
- 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. **1 point.** (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.** *REINFORCE with Baseline*

   Written: Consider the standard REINFORCE algorithm, where the policy gradient estimate is given by:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} R(\tau)\nabla_\theta \log \pi_\theta(a_t|s_t)\right].$$

   In practice, we often introduce a state-dependent baseline function $b(s_t)$ to reduce variance, resulting in:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} (R(\tau) - b(s_t))\nabla_\theta \log \pi_\theta(a_t|s_t)\right].$$

   (a) Prove that adding a baseline to REINFORCE is still unbiased and the variance is lower.

   (b) What values could replace the $R(\tau)$ in REINFORCE?

3. **2 points.** *Off-policy PG*

   Written: In reinforcement learning, we often want to learn an optimal target policy $\pi_\theta(a|s)$ parameterized by θ, while collecting data using a different behavior policy $\beta(a|s)$. The standard policy gradient objective for on-policy learning is:

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$$

   where τ represents a trajectory $\{s_0, a_0, r_0, ..., s_T, a_T, r_T\}$ and $R(\tau)$ is the total return of the trajectory. Consider the problem of optimizing a target policy π_θ using data collected from a behavior policy β. Derive the off-policy policy gradient theorem by completing the following steps:

(a) Explain why the standard policy gradient cannot be directly estimated when using trajectories sampled from β.

(b) Apply importance sampling to express the policy gradient objective in terms of expectations over trajectories from the behavior policy β to derive off-policy PG.

4. **5 points with 5 points Extra Credit** *Atari with PG*

We provide the scffolding code in the notebook.

(a) **Code, Plot:** **3 points.** Implement DDPG and evaluate it on cartpole.

(b) **Code, Plot:** **2 points.** Implement TD3 and evaluate it on cartpole.

(c) **Code, Plot:** **Extra Credit 2 points.** Implement SAC and evaluate it on cartpole.

(d) **Code, Plot:** **Extra Credit 2 points.** Implement PPO and evaluate it on cartpole.

(e) **Written:** **Extra Credit 1 point.** Compare the results and show your findings.