Lab 7

Policy-Based Reinforcement Learning

Alison Wen, Wei Hung

Contents

Background	A2C, PPO, GAE
Lab Description	Task 1: A2C on Pendulum Task 2: PPO on Pendulum Task 3: PPO on Walker
Model & Packages	Classes and required packages
Grading Policy	Report + Code + Video
Submission Policy	There will be penalty using the wrong file names!!

Background

Policy-Based RL

A2C: Advantage Actor Critic

- Actor: $\boldsymbol{\pi}_\theta(s|a)$

$$J_{\text{actor}}(\theta) = \mathbb{E}_{s\sim d_\mu^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)}\left[\log \pi_\theta(a|s)\cdot A^{\pi_\theta}(s,a)\right] \approx \frac{1}{|B|}\sum_{(s,a,r,s')\in B} \log \pi_\theta(a|s)\hat{A}_w(s,a)$$

- Critic: $V_w(s)$

$$L_{\text{critic}}(w) = \mathbb{E}_{(s,a,r,s')}\left[\left(r+\gamma V_w(s') - V_w(s)\right)^2\right] \approx \frac{1}{|B|}\sum_{(s,a,rs')} \left(r+\gamma V_w(s') - V_w(s)\right)^2$$

- Advantage

$$\hat{A}_w(s,a) = r + \gamma V_w(s') - V_w(s)$$

Algorithm 1 Advantage Actor-Critic (A2C)

Initialize actor network $\pi_\theta(a|s)$ and critic network $V_w(s)$
for each episode **do**
    Initialize state s_0
    for $t = 0$ to T **do**
        Sample action $a_t \sim \pi_\theta(a_t|s_t)$
        Execute a_t, observe r_t, s_{t+1}
        Compute TD-error: $\delta_t = r_t + \gamma V_w(s_{t+1}) - V_w(s_t)$
        Update critic: $w \leftarrow w - \alpha_c \nabla_w \delta_t^2$
        Update actor: $\theta \leftarrow \theta + \alpha_a \nabla_\theta \log \pi_\theta(a_t|s_t) \cdot \delta_t$
        $s_t \leftarrow s_{t+1}$
    end for
end for

PPO-Clip With GAE

- **Clipped Surrogate Actor Objective**

$$L_{\text{clip}}(\theta) = \mathbb{E}_{(s,a,r,s')}\left[\min\left(\rho_{s,a}(\theta)A^{\text{GAE}}(s,a),\ \text{clip}(\rho_{s,a}(\theta), 1-\epsilon, 1+\epsilon)A^{\text{GAE}}(s,a)\right)\right],$$
$$\approx \frac{1}{|B|}\sum_{(s,a,r,s')\in B} \min\left(\rho_{s,a}(\theta)A^{\text{GAE}}(s,a),\ \text{clip}(\rho_{s,a}(\theta), 1-\epsilon, 1+\epsilon)A^{\text{GAE}}(s,a)\right)$$

- **Critic Loss:**

$$L_{\text{critic}}(w) = \mathbb{E}_{(s,a,r,s')}\left[\left(r + \gamma V_w(s') - V_w(s)\right)^2\right] \approx \frac{1}{|B|}\sum_{(s,a,rs')}\left(r + \gamma V_w(s') - V_w(s)\right)^2$$

- **Overall Objective**

$$J_{\text{PPO}} = J_{\text{clip}}(\theta) - c_1 L_{\text{critic}}(\phi) + c_2 \mathbb{E}_{s\sim d_\mu^{\pi_\theta}}\left[H(\pi_\theta(\cdot|s))\right]$$

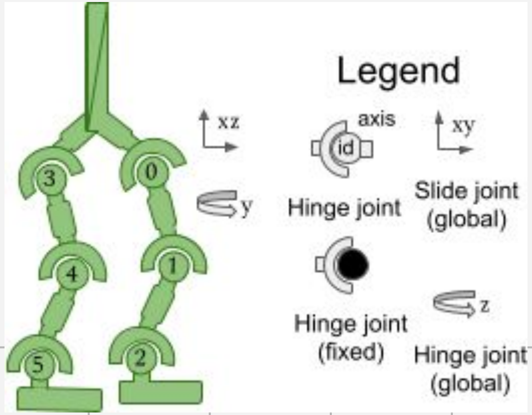
Lab Description

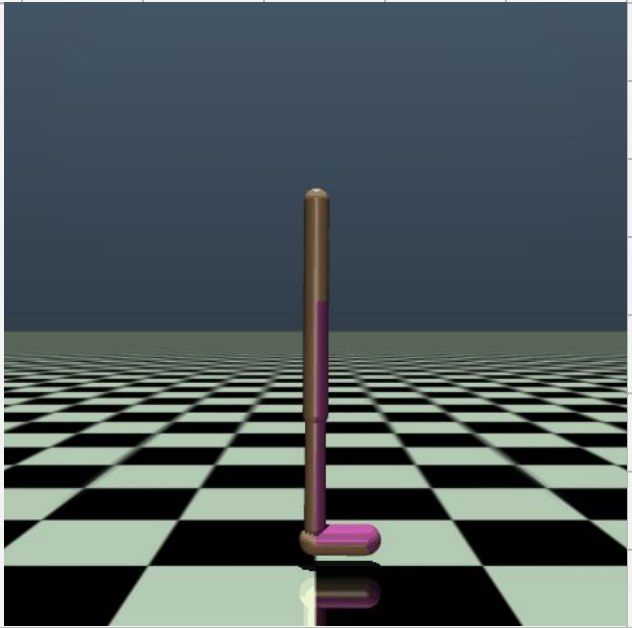
Environment: Pendulum

- Goal: swing it into an upright position
- State: $\boldsymbol{\theta} \in [-2, 2]$, $\boldsymbol{\omega} \in [-1, 1]$
- Action: $\boldsymbol{\tau} \in [-2, 2]$
- Observation Space: [x, y, $\boldsymbol{\omega}$]
- Reward:
  $R = -(\theta + 0.1\frac{d\theta}{dt} + 0.001\tau^2)$

Environment: Walker

- Goal: Walk in the forward direction
- Observation Space
- reward = healthy_reward bonus + forward_reward - ctrl_cost
- Action Space:
  - 6 joint torques





Tasks

- Task1: A2C on Pendulum
- Task2: PPO with GAE on Pendulum
- Task3: PPO on Walker2d

Grading Policy

Report

- Introduction (5%): Please provide a high-level introduction to your report. You can mention the most important findings and the overall organization of this report.
- Your implementation (20%): Please briefly explain your implementation for Tasks 1-3. Specifically, please describe:
  - How do you obtain stochastic policy gradient and TD error for A2C?
  - How do you implement the clipped objective in PPO?
  - How do you obtain the estimator of GAE?
  - How do you collect samples from the environment?
  - How do you enforce exploration (despite that both A2C and PPO are on-policy RL methods)?
  - Explain how you use Weight \& Bias to track model performance and the loss values (including actor loss, critic loss, and the entropy).

Report

- Analysis and discussions (25%)
  - Plot the training curves (evaluation score versus environment steps) for Task 1, Task 2, and Task 3 separately
  - Compare the sample efficiency and training stability of A2C and PPO.
  - Perform an empirical study on the key parameters, such as clipping parameter and entropy coefficient
  - Additional analysis on other training strategies (Bonus)

Demo Video

- Total Duration: 5–6 minutes
- Language: English (unless pre-approved by TAs)
  - 🔹 Source Code (~2 minutes): Describe your implementation
  - 🔹 Model Performance (~3 minutes): Demonstrate your obtained models

⚠️ Model snapshots will NOT be graded if no valid demo video is provided.

Model Snapshots

- **Task 1 & Task 2**

$$\text{Score percentage} = \left(1 - \frac{\max\{0, X - 200\text{k}\}}{800\text{k}}\right) \times 15\%$$

- **Task 3**

Environment steps needed (Reaching score 2500 on Walker2d)	1M	1.5M	2M	2.5M	3M	>3M
Score Percentage	20%	16%	12%	10%	8%	5%

Submission Policy

Directory Structure

```
LAB7_StudentID_YourName.zip
|-- LAB7_StudentID_YourName_Code/        <- Source code folder
|   |-- ppo_walker.py                          <- Your code files
|   |-- (any other .py files)
|-- LAB7_StudentID_YourName.pdf          <- Technical report (single PDF)
|-- LAB7_StudentID_YourName.mp4          <- Demo video (5 - 6 minutes)
|-- LAB7_StudentID_task1_a2c_pendulum.pt     <- Task 1 model snapshot
|-- LAB7_StudentID_task2_ppo_pendulum.pt         <- Task 2 model snapshot
|-- LAB7_StudentID_task3_ppo_1m.pt    <- Task 3 snapshot (step = 1M)
|-- LAB7_StudentID_task3_ppo_1p5m.pt    <- Task 3 snapshot (step = 1.5M)
|-- ...
|-- LAB7_StudentID_task3_ppo_3m.pt   <- Task 3 snapshot (step = 3M)
```

Thanks for Your Attention