Homework 1 Template

Use this template to record your answers for Homework 1. Add your answers using LaTeXand then save your document as a PDF to upload to Gradescope. You are required to use this template to submit your answers. **You should not alter this template in any way** other than to insert your solutions. You must submit all 8 pages of this template to Gradescope. Do not remove the instructions page(s). Altering this template or including your solutions outside of the provided boxes can result in your assignment being graded incorrectly. You may lose points if you do not follow these instructions.

Instructions to upload code have been provided in the handout.

Instructions for Specific Problem Types

On this homework, you must fill in the blank for each problem; please make sure your final answer is fully included in the given space. **Do not change the size of the box provided.** For short answer questions you should **not** include your work in your solution. Only provide an explanation or proof if specifically asked. Otherwise, your assignment may not be graded correctly, and points may be deducted from your assignment.

Fill in the blank: What is the course number?

10-703

Problem 0: Collaborators

Enter your team's names and Andrew IDs in the boxes below. If you do not do this, you may lose points on your assignment.

Name 1:	Boxiang Fu	Andrew ID 1:	boxiangf
Name 2:		Andrew ID 2:	
Name 3:		Andrew ID 3:	

Problem 1: REINFORCE (48 pts)

1.1 Reinforce plot (10 pts)

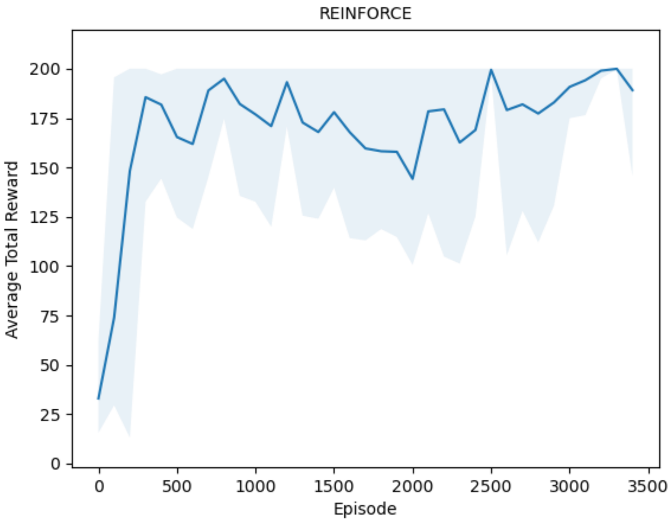

Figure 1: REINFORCE with default hyper-parameters

1.2 Reinforce with baseline plot (10 pts)

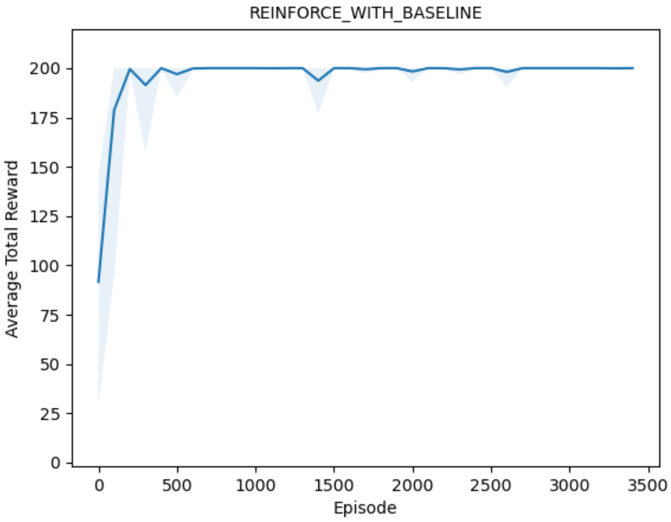

Figure 2: REINFORCE with baseline with default hyper-parameters

1.3 N-step A2C (20 pts)

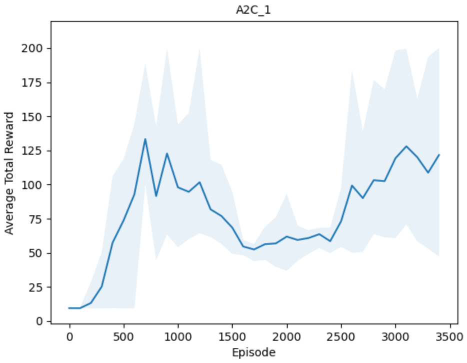

Figure 3: A2C with $N = 1$ and default hyper-parameters

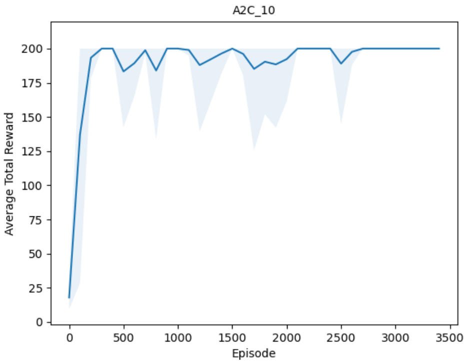

Figure 4: A2C with $N = 10$ and default hyper-parameters

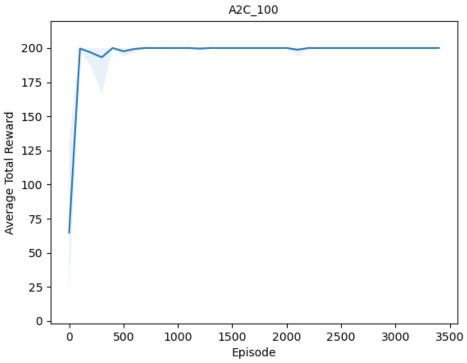

Figure 5: A2C with $N = 100$ and default hyper-parameters

1.4 N-step A2C & REINFORCE with baseline (4 pts)

The N-step A2C algorithm becomes equivalent to REINFORCE when we set the bootstrapping horizon to the episode end (i.e. $N = T - t$ or even $N \geq T - t$ to ensure $V_{end} = 0$) and remove the critic network V_ω or set it to zero and never update it. Intuitively, this acts as if we remove the critic and don't do bootstrapping.

The N-step A2C algorithm becomes equivalent to REINFORCE with baseline when we set the bootstrapping horizon to the episode end (i.e. $N = T - t$ or even $N \geq T - t$ to ensure $V_{end} = 0$) but keep the critic network V_ω and use it for the baseline. Intuitively, this acts as if we keep the baseline but don't do bootstrapping.

1.5 REINFORCE with & without baseline (4 pts)

Yes, adding a good state-dependent (but action-independent) baseline would generally improve performance. This is seen by comparing the plots in 1.1 and 1.2. It can be seen that adding a baseline allows the algorithm to converge faster and with less variability. The reason for this is because the baseline keeps the gradient estimator unbiased, but lowers its variance across episodes (see min/max bars in 1.1 and 1.2). It does this by assigning actions an advantage compared to the other actions, and updates the actor network based on this advantage, regardless of whether the state it is in is good or bad. However, a baseline that is bad (e.g. noisy, mis-scaled, etc) can actually decrease performance due to increased variance, so care must be taken in choosing a baseline.

2 Question Answering (12 pts)

1.

False.
Q-learning is off-policy, it converges to the optimal Q function independent of policy. SARSA is on-policy, it converges to the optimal Q function depending on the policy being followed. The two Q functions are equivalent only under GLIE (Greedy in the Limit with Infinite Exploration) conditions.

2.

False.
Q-learning needs sufficient exploration to allow the computed Q function to

converge to the optimal function. Normally, a epsilon-greedy method is used, where an epsilon chance a random action is chosen to explore. Just a greedy policy will limit the state-action pairs visited and result in a suboptimal policy.

3.

False.
By definition the optimal policy π^* maximizes value for every state. That is $v_{\pi^*}(s) \geq v_\pi(s)$ for all s. Therefore, the value $v_{\pi^*}(s)$ cannot be lower than $v_\pi(s)$ for some state s.
Reference: https://rltheory.github.io/lecture-notes/planning-in-mdps/lec2/

4.

False.
We can parametrize the output layer of the actor network to output discrete categorical actions (e.g. softmax such as in problem 1.3). Therefore, discrete actions can use actor-critic methods.

5.

False.
Actor-critic methods typically use stochastic policies for exploration. It samples trajectories by deploying the current policy using probabilities from a distribution. It does not typically use epsilon-greedy for exploration.

6.

Only the first statement is correct.
Switching to a_1 for state s will provide us with a policy better than π since other values stay the same for other states but is strictly better for state s. This also means statement 2 is incorrect since the switch will not decrease the value for other states when we switch the action at state s only. Finally, we do not know the global optimal policy as other action switches under different states may be required to reach the optimal policy.

Feedback

Feedback: You can help the course staff improve the course for future semesters by providing feedback. You will receive a point of you provide actionable feedback. What was the most confusing part of this homework, and what would have made it less confusing?

One of the confusing parts of the assignment was on how to combine PyTorch and Gym to be used together. Recitation went through tutorials on PyTorch and Gym separately, but did not touch on how to incorporate Gym into a PyTorch training pipeline. I think an improvement could be to have a recitation tutorial on how to use outputs from Gym and incorporate them into a PyTorch pipeline.

Collaboration: Detail the work division amongst your group below.

I did not join a group for this assignment as I had a weekend to spare to work on the assignment.

Time Spent: How many hours did you spend working on this assignment? Your answer will not affect your grade. Please average your answer over all the members of your team.

Alone	20
With teammates	0
With other classmates	0
At office hours	0