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

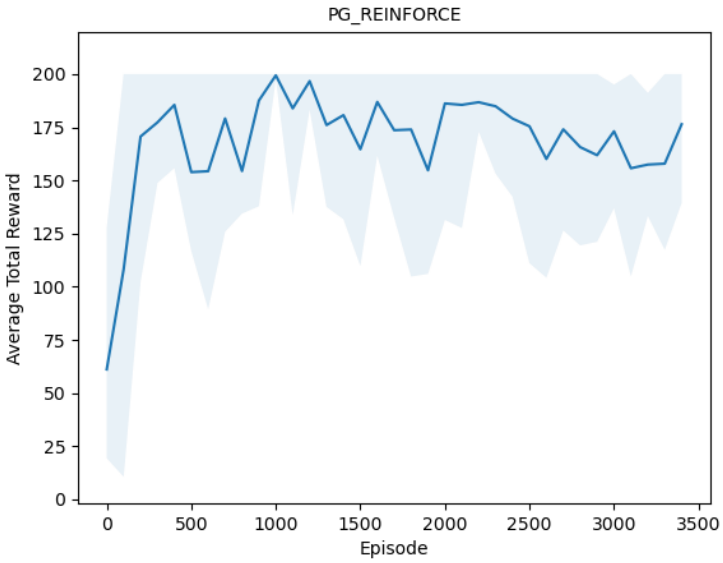
Problem 0: Collaborators

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.

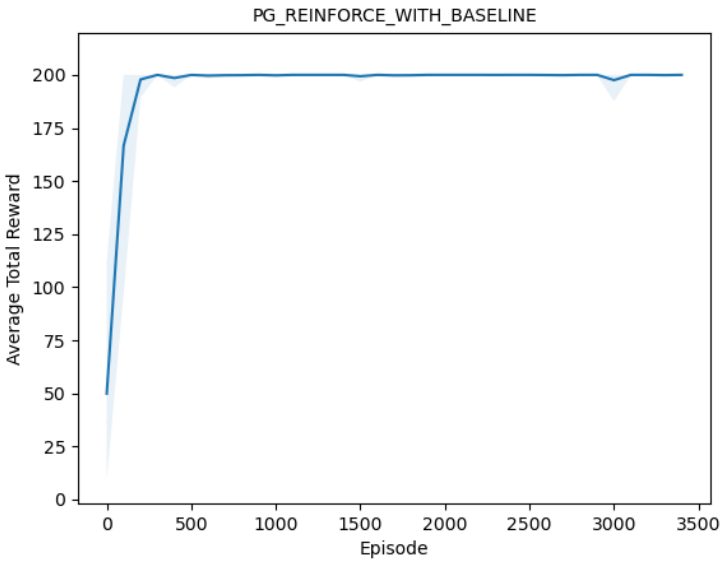
Name 1:	Mike Anoruo	Andrew ID 1:	manoruo
Name 2:		Andrew ID 2:	
Name 3:		Andrew ID 3:	

Problem 1: REINFORCE (48 pts)

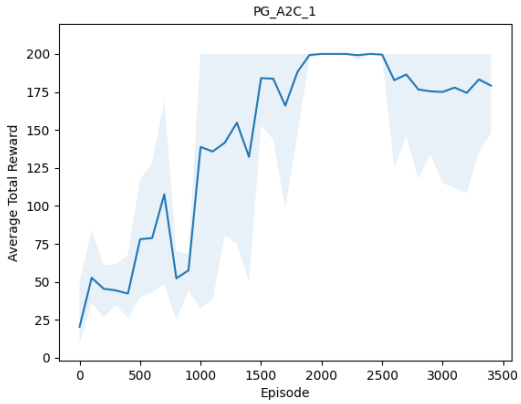
1.1 Reinforce plot (10 pts)

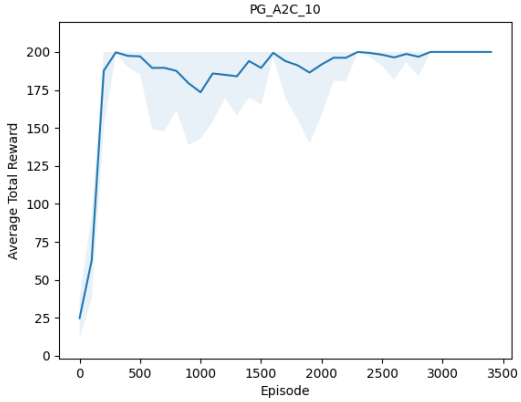


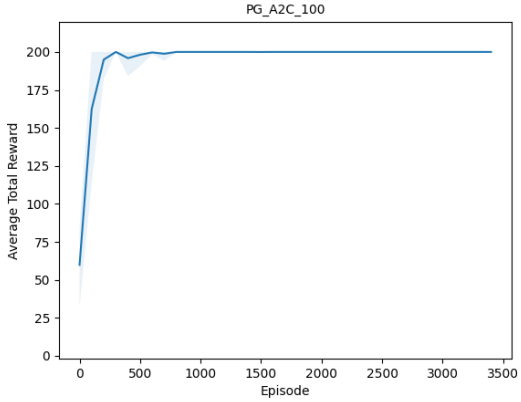
1.2 Reinforce with baseline plot (10 pts)



1.3 N-step A2C (20 pts)







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

N-step A2C, REINFORCE, and REINFORCE with baseline are all policy gradient methods that update the policy to increase the probability of actions leading to higher returns.

- **N-step A2C → REINFORCE with baseline:** when $n = T$, i.e., the return is computed over the entire episode (full-horizon), so no bootstrapping is used.
- **N-step A2C or REINFORCE with baseline → REINFORCE:** if the baseline $V(s_t)$ is zero (or not used), then both reduce to plain REINFORCE.

1.5 REINFORCE with & without baseline (4 pts)

Adding a baseline generally improves performance by reducing the variance of the policy gradient estimates, which allows the model to converge faster. The baseline acts like a "standard" or reference point for expected return from a state. Actions that lead to rewards higher than this standard are reinforced, while actions that fall below it are not. Without a baseline, REINFORCE tends to reinforce all actions in a trajectory that eventually lead to a reward, even if some of them didn't necessarily contribute to it. By using a baseline, the algorithm puts more "blame" or "credit" on actions that truly affect the outcome, reducing noise and improving learning efficiency.

2 Question Answering (12 pts)

1.

2.

False. Q-learning requires exploration during training to discover optimal actions in all states. If it only follows the greedy policy over the current Q function, it

may get stuck in suboptimal trajectories and fail to maximize long-term rewards.

3.

False. By definition, $v^{\pi^*}(s)$ is the value under the optimal policy, which is always at least as large as the value under any other policy. No policy can have a higher value at a state than the optimal policy.

4.

False. Actor-critic methods can optimize policies over discrete actions using a parameterized policy, such as a softmax output to represent action probabilities.

5.

False. Actor-critic methods typically use stochastic policies (e.g., softmax) to explore, rather than epsilon-greedy, which is primarily used in Q-learning.

6.

- Switching to a_1 for state s will provide us with a policy better than π. **(Only correct answer)**

Changing the action at state s to the higher-Q action a_1 improves the policy at that state. We cannot guarantee that other states' values will decrease, nor that

the resulting policy is globally optimal if other states are still suboptimal.

7.

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?

I had some confusion with the difference between A2C and REINFORCE with baseline. From the lecture, it seemed like this REINFORCE with baseline was the same as actor critic since we werent subtracting a constant value. I wasnt sure though since the homework didnt call it this. Also some template code for the size of the graphics would be good. Wasn't sure how big to make the images without making them too small or too big.

Collaboration: Detail the work division amongst your group below.

Worked on this assignment individually.

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	16
With teammates	0
With other classmates	0
At office hours	0