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

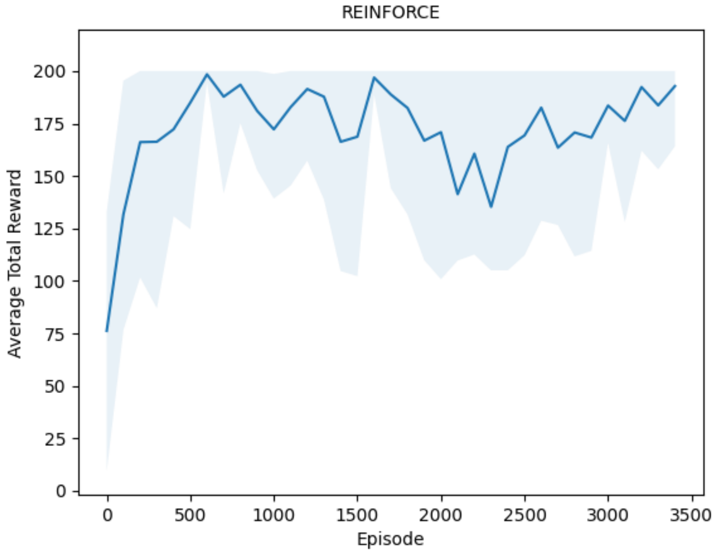
Problem 0: Collaborators

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.

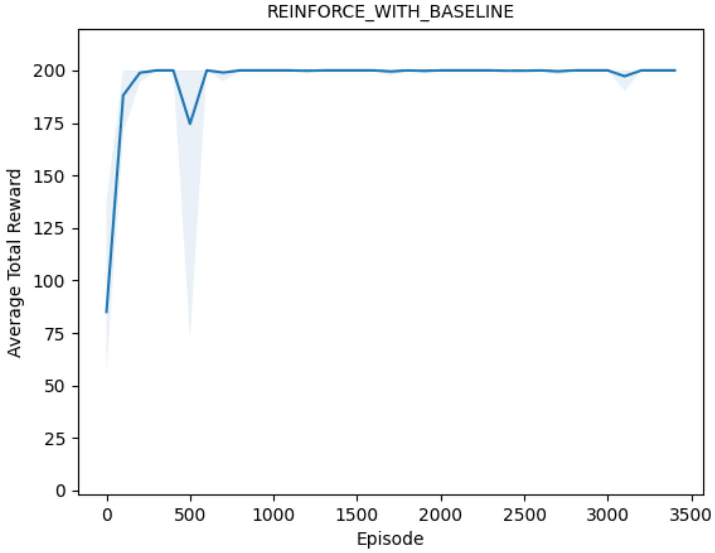
Name 1:	Gabriel Olin	Andrew ID 1:	golin
Name 2:	Andrew Jong	Andrew ID 2:	ajong
Name 3:		Andrew ID 3:	

Problem 1: REINFORCE (48 pts)

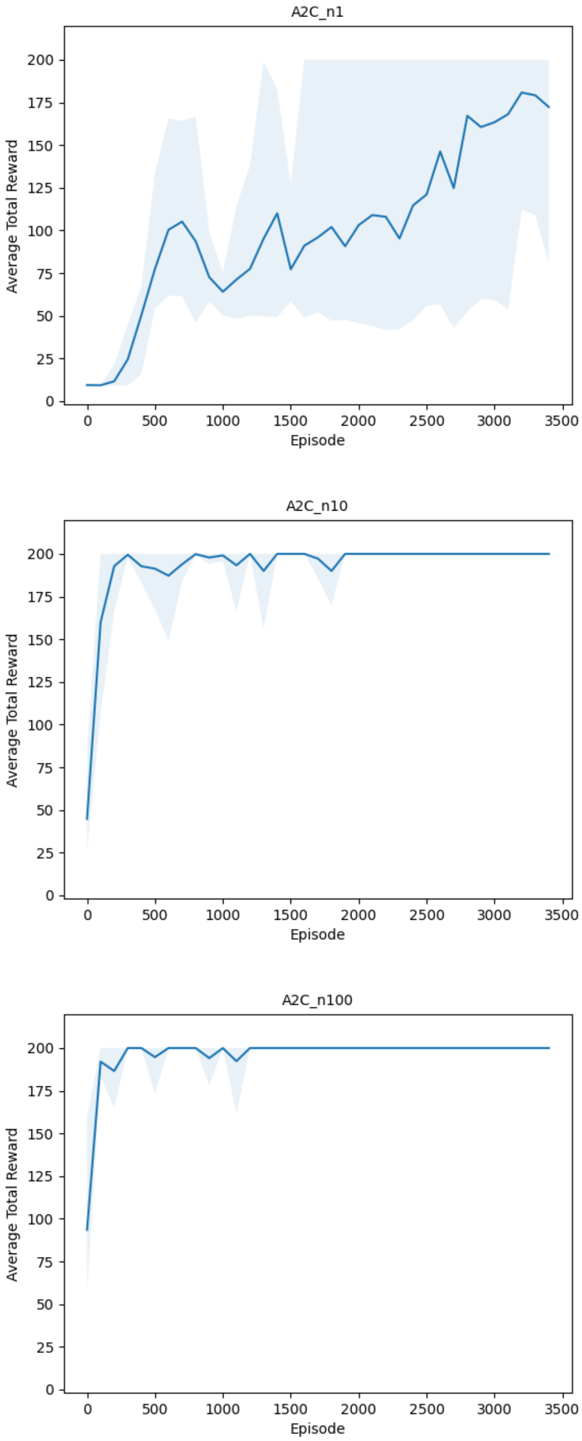
1.1 Reinforce plot (10 pts)



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)

REINFORCE uses the full monte carlo rewards from each episode to estimate the on-policy expected returns. The policy gradient is computed using the total reward from each trajectory. REINFORCE with Baseline is the same as REINFORCE but it subtracts a baseline from the rewards to reduce variance. N-step A2C (Advantage Actor-Critic) uses an n-step bootstrapped return as the target, and the baseline is the value function. N-step A2C becomes REINFORCE with Baseline when N is set to the episode length. The advantage is then the total return minus the value estimate, matching REINFORCE with baseline. N-Step A2C becomes vanilla REINFORCE when N is the episode length AND the value function for all states is zero.

1.5 REINFORCE with & without baseline (4 pts)

Adding a baseline to REINFORCE improves performance as the reward curve increases faster and stays close to the maximum reward for the environment with less fluctuation. This is because the baseline reduces variance in the policy gradient estimate without introducing bias, because it encodes which actions are better than average, rather than being influenced by randomness. This results in more stable training and less noisy policy updates.

2 Question Answering (12 pts)

1.

False. Q-learning is off policy and learns the optimal Q function for the MDP as value iteration converges, due to taking argmax over actions when updating Q-values. SARSA learns the Q function for its current policy (e.g. epsilon-greedy), and only converges to the Q function of that policy.

2.

False. Even though Q-learning uses argmax over actions to do the TD-update of the Q function, it still needs some exploratory policy to visit all state-action pairs. Otherwise, it won't converge to the optimal policy because it won't explore better state-actions.

3.

False. The optimal value function is defined as the maximum expected return over all possible policies. So, if it is less than some other policy pi for some state s, then it cannot be the optimal policy

4.

False. Actor-critic can be used for discrete action spaces. The actor network outputs a vector of probabilities for taking each each action. The value network outputs a scalar estimating the expected future returns from a given state.

5.

False. Actor-critic doesn't use epsilon-greedy to explore. As an on-policy algorithm, the exploration comes from the stochastic nature of the actor. That is, it will choose different actions with some probability, and then update the policy to take better actions with higher probability

6.

a. True. Switching to a1 will improve the policy because by definition of the q function, the expected reward achieved by following pi after taking a1 is greater than after taking 2 and then following pi.
b. False. Acting more optimally in state s can only improve the value of states that are visited before it under the new policy. For states that come after it (even if it cycles through s again), the value will not decrease.
c. False. We only know the Q function for the current policy. Acting more optimally in one state does not guarantee we have found the global 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?

Collaboration: Detail the work division amongst your group below.

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	
With teammates	
With other classmates	
At office hours	