**Dear all students,**

**Follow the instructions carefully, please:**

* You can find the following three questions below. You have **120 minutes** to answer the questions and **10 minutes of upload** time (not extendable more).
* For the questions that you need to answer in the text, add your answer in this world file and upload it into Moodle. For the questions that you need to write code for, upload your Python file into Moodle. Rename both files using your English name or student ID.
* You can use Slides and class code examples.
* No need to mention any communication during the exam is acceptable.
  + No open chat apps in background phones, only in the pocket all the time, and using the internet is completely prohibited (you must disconnect your personal laptop or computer from the network during the exam).

**Good Luck.**

--------------------------------------------------------------------------------------------------

English Name: CHEN, HSIN-YU

Chinese name: 陳昕佑

Student ID:61375017H

--------------------------------------------------------------------------------------------------

**Questions:**

1. A) What are the differences in reward function in model-free and model-based RL algorithms? B) Explain clearly why Temporal Differences are in the Bootstrapping class.

Answer:

1. Model-Free Learning: A method where the agent does not rely on a model of the environment’s dynamics.(Exploration)

Model-Based Learning: A method that uses a model of the environment to predict outcomes of actions and plan accordingly. (Exploitation)

The Model-free method doesn’t know the whole environment, it has to explore.

1. The Bootstrapping means the process of updating estimates based on other learned estimates rather than waiting for the final outcome. And TDs is a method that learns by updating value estimates based on differences between predictions at consecutive time steps. That’s why TDs are in the Bootstrapping class.
2. A) In value function approximation approaches, if we calculate the Q-value for a given action, we might have Multiple forward passes through the network challenge; how can we solve this challenge? B) Regarding Policy optimization approaches, is the critic network output a policy to run on the agent? Explain?

Answer:

1. Single Forward Pass with All Actions (DQN Implementation):

Instead of calculating Q(s,a) separately for each action a, modern DQN implementations structure the output of the neural network to return Q-values for all actions simultaneously. This allows the agent to select the best action (argmax a Q(s,a)) or compute the loss for a specific action using just one forward pass.

1. No, the network gives the propability of each policy, than do the best one to run on the agent.
2. Considering the Q-Learning code provided for the OpenAI Gym environment in class. Based on the 'FrozenLake-v1' environment from Gym, update the given code by merging eligibility traces implementation with Q-learning and showing the performance plot as total reward/episode convergence.

Answer:

I wanted to use the TD(lambda) to merg eligibility traces implementation with Q-learning, but I didn’t make sure that if the state and action can be the same.

Upload your Python code only (don't compress it, and there are no screenshots of the code!). You can only show screenshots of plot results here in the Word file.