**Dear all students,**

**Follow the instructions carefully, please:**

* Kindly you can find the following 4 questions you have whole **120 minutes** to answer the questions and **10 minutes of upload** time (not extendable).
* 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, upload your python file into Moodle. Rename both files to your English name or student ID.
* You can use Slides and Class code examples.
* You are not allowed to use internet, and communication during the exam is not allowed.
  + No open chat apps in background
  + Phones only in the pocket all the time
  + Disconnect your personal laptop or computer from the network during the exam

**Good Luck.**

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

English Name: Jun Yu Shen

Chinese name: 沈峻宇

Student ID: 61075029H

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

**Questions:**

1. A) What is the benefit and disadvantage of using the TD(0)? B) How the TD(λ) backward view is working? Explain motivation and how it can increase the performance?

Answer:

TD learning is model-free and learn directly from raw experience without a model of the environment’s dynamics, also can update estimates based in part on other learned estimates, without waiting for a final outcome.

After each step in the episode it makes updates to all prior steps.ET assigns credit to states that are both visited frequently as well as visited recently with respect to the terminal state. At each moment we look at the current TD error and assign it backward to each prior state according to the state's eligibility trace at that time.

Because the traces rate indicates the degree to which each state is eligible for undergoing learning changes should a reinforcing event occur.

1. A) How we can convert or change DQL off-policy algorithm to one on-policy algorithm? B) Why exactly policy gradient algorithms are good option for continues action space while DQL is not?

Answer:

If for some of actions or random:

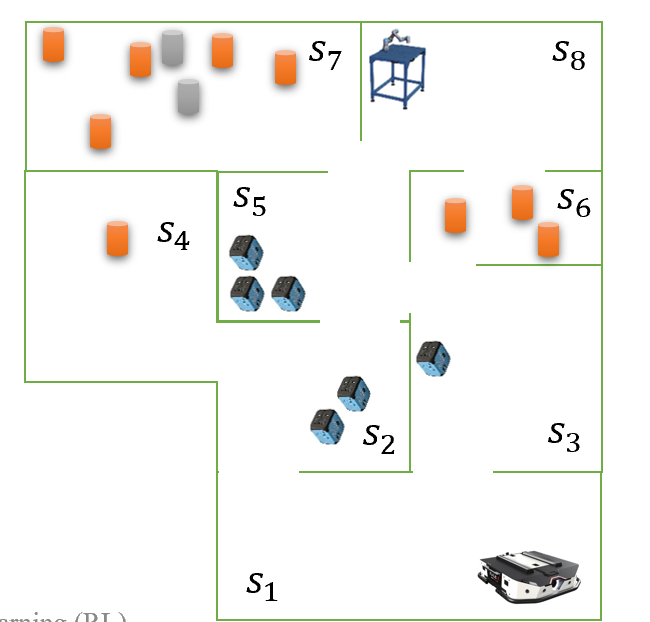
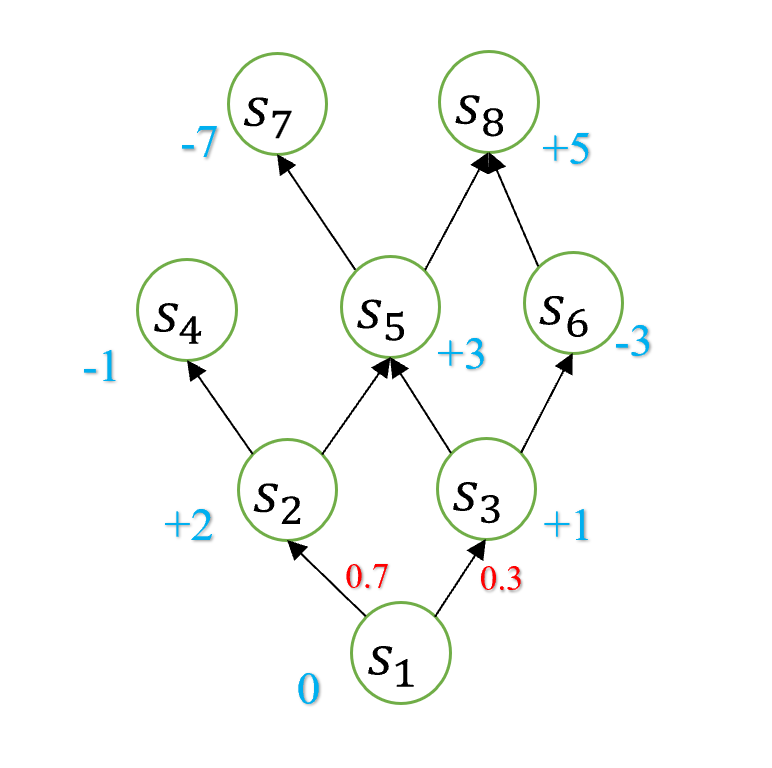
We have off-policy DQL.

If for all actions:

We have on-policy DQL.

Because DQL is using Q-Learning. Value-based approaches are computationally very expensive if we want to use them in the continuous space.

1. Show Q-learning algorithm results for two episodes on following environment (no code). (Initial state and parameters that you use state them clear; all arbitrary). You need to only show update values for Q-value (no need to draw whole tables).



Answer:

= 0.9

= 0.8

Q(S2, S5) = Q(S2, S5) + 0.9 [R(S2, S5) + 0.8 \* [Q(S5, S7), Q(S5, S8)] – Q(S2, S5)] = 0 + 0.9 \* [3 + 0.8 \* 0 - 0 ] = 2.7

Q(S3, S6) = Q(S3, S6) + 0.9[R(S3, S6) + 0.8 \* [Q(S6, S8)] – Q(S3, S6)] = 0 + 0.9 \* [-3 + 0.8 \* 0 - 0 ] = -2.4

1. Merge Eligibility Trace algorithm with Actor-Critic algorithm to update network weights (show only the algorithm and flow must be correct, no need to code).

Answer: