Deep Reinforcement Learning

Homework #1

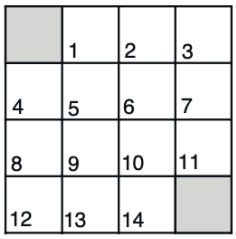
Due: 2021/03/29 (Mon.) 23:59

Change Log

1. 2021.03.19: Clarify the axes order of the inputs, upload to Google Drive.
2. 2021.03.16: Extend deadline by one week. Enhanced the description and provided explicit sample I/O.
3. 2021.03.09: No need to demo, TAs will grade your report directly.

Problem Description

1. Random Policy Evaluation (“`<Student_ID>_hw1_1.py`”)
   Consider the following environment:

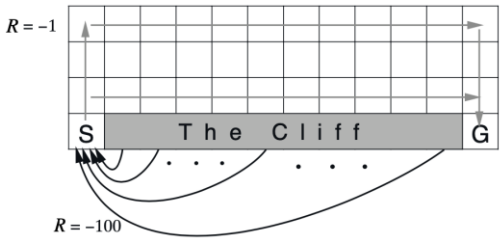


   a. Normal states: {1-14}, e.g., 
   b. Terminal states: grey states, e.g., 
   c. Action space: {→, ↑ ,←, ↓ }
   d. Rewards: -1 for all transitions
   e. If the agent hits the wall (e.g., move left at state 4), it stays at the same state in the next timestep
   f. Random Policy: $\pi(a|s) = 0.25, \forall a, s$

   Calculate the state value (v_π) of the environment:

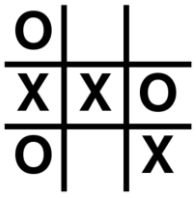
   a. Output the learned (converged) v_π for state 1-14 (order matters). The output should be rounded to the second decimal place.
   b. Calculate the correct results of v_π with $\gamma = 0.9$, and store the result in “`<Student_ID>_hw1_1_data_gamma_0.9`”
   c. Calculate the correct results of v_π with $\gamma = 0.1$, and store the result in “`<Student_ID>_hw1_1_data_gamma_0.1`”
   d. Explain your code and results. How does the iteration stop? How does the discount factor γ affect the results?
   e. Sample Output: (numbers separated by spaces, no trailing characters) See “`hw1-1_sample_output`”.

2. Compare Q-Learning with SARSA (“<Student_ID>_hw1_2.py”)
   Design an environment to compare Q-learning with SARSA. Think about the Cliff Walking example, what reward setting could lead to such difference?

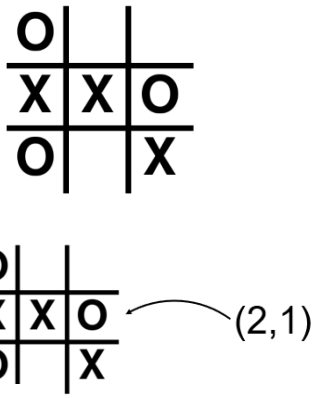


   a. Designed a different environment and plot the learning curves of both Q-learning and SARSA on the new environment.
   b. The two algorithms should learn different policies and have different learning curves in your designed environment.
   c. Elaborate on your environment design. How do you design the reward of your new environment? What are the behaviors of these algorithms? Explain the behaviors of these algorithms.

3. Tic-tac-toe (3x3) (“<Student_ID>_hw1_3_<train|test>.py”)
   Learn a policy for playing Tic-tac-toe:

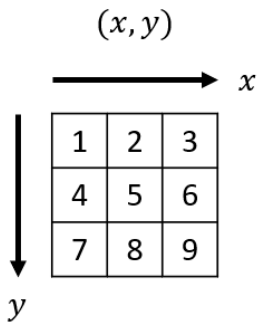


   a. State space: 3x3 integer array, e.g., the array $[[-1,0,0],[1,1,-1],[-1,0,1]]$ corresponds to:
      i. 1 represents ‘X’
      ii. -1 represents ‘O’
      iii. 0 represents empty space



   b. Action space: integer coordinates (x, y), e.g., The coordinate represents the position for your move.
   c. Goal: win the game.
   d. Use Q-learning, SARSA or other tabular method (without function approximation)
   e. You may store your learned results in an external file “./hw1_3_data” (readonly, max 10 MB), and access it with your program (during testing).
   f. Explain your training procedure and how you implemented the environment.

g. Program I/O:
    i. The input will be 10 integers separated by spaces. The first integer indicates your player $\{1, -1\}$, the reset of them corresponds to the current board.
    ii. The input state is a valid state, begin with either O or X.
    iii. The output should be 2 integers followed by a newline specifying your move.
    iv. Your program should support multiple I/Os until EOF.
    v. The input integers are structured with the follow order:



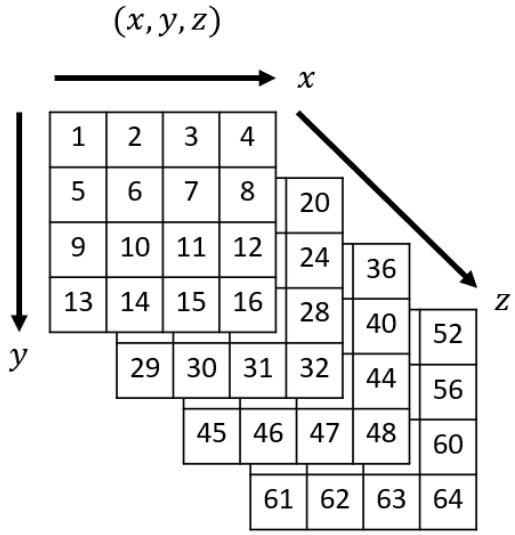
(The cell number corresponds to the i^{th} input)

h. Sample I/O:
See "hw1-3_sample_input", "hw1-3_sample_output"

4. 3D Tic-tac-toe (4x4x4) ("<Student_ID>_hw1_4_<train|test>.py")
    a. State space: 4x4x4 integer array.
    b. Action space: integer coordinates (x, y, z)
    c. Goal: win the game.
    d. Implement you policy with Monte Carlo Tree Search (MCTS)
    e. Other improvements are allowed. The detailed rules are specified in the next section.
    f. Program I/O:
        i. The input will be 65 integers separated by spaces. The first integer indicates your player $\{1, -1\}$, the rest of them corresponds to the current board.
        ii. The input state is a valid state, begin with either O or X.
        iii. The output should be 3 integers followed by a newline specifying your move.
        iv. You can keep overriding your move during the time limit by output new coordinates. Only the last successful output move is selected.
        v. Make sure to flush your output to make sure we can observe your action if your program is killed due to timeout. (e.g.,

```
print("0 0 0", flush=True))
```

vi. The input integers are structured with the follow order:



(The cell number corresponds to the i^{th} input)

g. Sample I/O:
See "`hw1-4_sample_input`", "`hw1-4_sample_output`"

h. You may store your learned results in an external file "`./hw1_4_data`" (readonly, max 50 MB), and access it with your program (during testing).

i. Explain your training / testing procedure and how you implemented the environment.

Detailed Rules for 3D Tic-tac-toe

1. If your program outputs invalid moves, you lose and the game ends immediately.
2. Time limit for each move is 5 seconds, and the memory limit is 4 GB. (Note that the 5 second duration may vary depending on different processors.)
3. You are allowed to access an external file (readonly, max 50 MB) for storing your learned policy. You can read the file at the following path: "`./hw_1_4_data`".
4. You can keep overriding your move during the time limit. Only the last successful output move is selected.
5. You are allowed to use the following Python packages for 2D & 3D Tic-tac-toe:
   a. numpy, scipy, gym, pandas, tensorflow, pytorch (non-GPU accelerated)
   b. You are allowed to use Python default installed packages.
   (e.g., sys, time, pickle, random, etc.)
   c. If you need to use other packages, state your reasons and post on iLMS.
6. You are not allowed to use the following:
   a. Other External API/libraries (e.g. System calls)

b. Inline assembly (e.g. asm)
c. Vectorization instructions (e.g. SSE, AVX)
d. Multi-threading (e.g. multiprocessing package)
e. GPU acceleration (e.g. CUDA code)
f. Networking (e.g. Sockets, MPI)

Program Submission

1. For each problem, please use Python to implement with a single source file.
2. Your files must be named as:
   a. "`<Student_ID>_hw1_1.py`"
   b. "`<Student_ID>_hw1_1_data_gamma_0.9`"
   c. "`<Student_ID>_hw1_1_data_gamma_0.1`"
   d. "`<Student_ID>_hw1_2.py`"
   e. "`<Student_ID>_hw1_3_train.py`"
   f. "`<Student_ID>_hw1_3_test.py`"
   g. "`<Student_ID>_hw1_3_data`"
   h. "`<Student_ID>_hw1_4_train.py`"
   i. "`<Student_ID>_hw1_4_test.py`"
   j. "`<Student_ID>_hw1_4_data`"
   k. and please make sure that all characters of the filename are in lower case. For example, if your student id is `108062000`, the name of your program file should be `108062000_hw1_1.py` and so on.
1. Your program will be ran in a GNU/Linux environment with Python 3.8:
   `python <Student_ID>_hw1_4.py`
2. 0 points will be given to Plagiarism. NEVER SHOW YOUR CODE to others and you must write your code by yourself. If the codes are similar to other people and you can't explain your code properly, you will be identified as plagiarism.
3. 0 points will be given if you violate the rules above.
4. If you use modularized / OOP code and want to use multiple files to keep your code structured, please email us beforehand.

Report

1. Elaborate on how you design your tic-tac-toe. The report is graded directly. So, make sure you have included enough details and figures to help TAs grade your report.
2. TAs will not refer to your code when grading your report, so make sure you have taken a screenshot of the important code snippets.
3. The report filename must be "`<Student_ID>_hw1_report.pdf`" and please

make sure that all characters of the filename are in lower case.

Grading Policy

1. The project accounts for 15 points (tentative) of your total grade.
2. You must submit both your source code and report. Remember the submission rules mentioned above, or you will be punished on your grade. Late submission rules are specified in the Lecture 1 Slides.
3. Compress all your files directly (do not compress the folder containing your files) and upload to this Google Form before the deadline. (Total 11 files)
4. The TA code will not be released. Your code will be tested against them after the submission deadline.

- Random Policy Evaluation
  - Converged Value Function (5%)
  - Report (Discussion) (15%)
- Compare Q-Learning with SARSA
  - Designed Environment and Learning Curves (10%)
  - Report (Discussion) (10%)
- Tic-tac-toe
  - Never Lose the Random Agent (5%)
  - Never Lose the TA Agent (5%)
  - Report (Discussion) (10%)
- 3D Tic-tac-toe
  - Compete with Random Agent (10%)
  - Compete with Your Classmates (10%)
  - Report (Discussion) (20%)