CSE546 - Reinforcement Learning

Assignment - 1 Checkpoint

Venkata Sai Divya Pallineni
Department of Computer Science and Engineering
University at Buffalo, Buffalo, NY 14260
vpalline@buffalo.edu

1 Defining RL Environments

1.1 Deterministic Environment

A 5x5 Grid-World is created and when the Agent is in State (s) and performing an Action (a) the probability that it will reach state ($s\prime$) is either 0 or 1.

- **State (S) :-** Defined 25 states from [0][0] to [4][4] of which Agent's
  . Initial position = [0][4]
  . Terminal position = [4][0]

- **Action (A) :-** The agent can take four actions in a single time step by changing the x,y coordinates.
  . $A = \{Up, Down, Right, Left\}$

- **Transition Probability (P):-** When the Agent is in State (s) and performing an Action (a) the probability that it will reach state ($s\prime$) is either 0 or 1.
  . $P(s\prime, r|s, a) = \{0, 1\}$

- **Reward (R):-** 5 rewards of which for Gold chest(+10), Food (+25), Devil (-5), Dragon (-10) , End Goal (+50) and zero(0) is awarded.
  . $R = \{-10, -5, 0, +5, +10, +50\}$

- **Discount factor (γ):-** The agentś goal is to maximize the expected sum of rewards without any discounting.
  . $\gamma = 1$

- **Main Objective:-** The agentś main goal is to reach End Goal or terminal position which is state [4][0] to collect maximum reward of +50

Preprint. Under review.

1.2 Stochastic Environment

A 5x5 Grid-World is created and when the Agent is in State (s) and performing an Action (a) there is a randomness or uncertainty in which state the Agent will end-up.

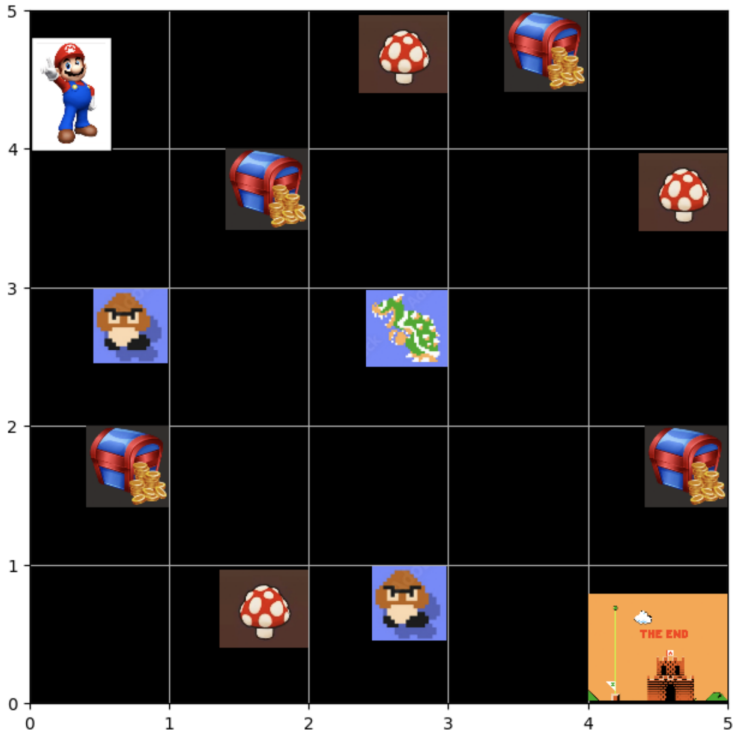
- **State (S) :-** Defined 25 states from [0][0] to [4][4] of which Agent's
  . Initial position = Random or Uncertain
  . Terminal position = [4][0]
- **Action (A) :-** The agent can take four actions in a single time step by changing the x,y coordinates.
  . $$A = \{Up, Down, Right, Left\}$$
- **Transition Probability (P):-** When the Agent is in State (s) and performing an Action (a) the probability that it will reach state ($s\prime$) is either 0 or 1.
  . $$P(s\prime, r|s, a) = \{0, 1\}$$
- **Reward (R):-** 5 rewards of which for Gold chest(+10), Food (+25), Devil (-5), Dragon (-10) , End Goal (+50) and zero(0) is awarded.
  . $$R = \{-10, -5, 0, +5, +10, +50\}$$
- **Discount factor (γ):-** The agentś goal is to maximize the expected sum of rewards without any discounting.
  . $$\gamma = 1$$
- **Main Objective:-** The agentś main goal is to reach End Goal or terminal position which is state [4][0] to collect maximum reward of +50

2 Visualizations

2.1 Deterministic Environment

2.1.1 Displaying Deterministic Environment with Agent in its initial position

```
# Initial State Position of the Environment
env = Mario_Game_Deterministic_Environment(16)
env.reset()
env.render()
```



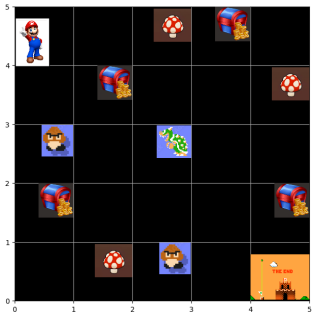
2.1.2 Cumulative rewards collected during 16 time steps

In [95]:

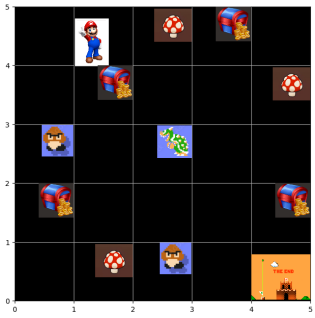
```
#Running the Mario GridWorld Game in Deterministic Environment
done = False
Actions={3:"Left",1:"Right",2:"Up",0:"Down"}
print("=============================================================================================================")
while not done:
    action = random.randint(0,3)
    reward, done, info = env.step(action)
    env.render()
    print("Timestep: {}".format(env.timeStep)+"\t\t\t Performing Action: "+Actions[int(action)])
    print(info)
    print("=============================================================================================================")
```

```
=============================================================================================================
Timestep: 1                     Performing Action: Left
Current Agent Position: [0,4] ; Current State Reward: 0 ; Total Cumulative Reward: 0
=============================================================================================================
Timestep: 2                     Performing Action: Right
Current Agent Position: [1,4] ; Current State Reward: 0 ; Total Cumulative Reward: 0
=============================================================================================================
Timestep: 3                     Performing Action: Down
Current Agent Position: [1,3] ; Current State Reward: 10 ; Total Cumulative Reward: 10
=============================================================================================================
Timestep: 4                     Performing Action: Right
Current Agent Position: [2,3] ; Current State Reward: 0 ; Total Cumulative Reward: 10
=============================================================================================================
Timestep: 5                     Performing Action: Down
Current Agent Position: [2,2] ; Current State Reward: -10 ; Total Cumulative Reward: 0
=============================================================================================================
Timestep: 6                     Performing Action: Down
Current Agent Position: [2,1] ; Current State Reward: 0 ; Total Cumulative Reward: 0
=============================================================================================================
Timestep: 7                     Performing Action: Left
Current Agent Position: [1,1] ; Current State Reward: 0 ; Total Cumulative Reward: 0
=============================================================================================================
Timestep: 8                     Performing Action: Down
Current Agent Position: [1,0] ; Current State Reward: 25 ; Total Cumulative Reward: 25
=============================================================================================================
Timestep: 9                     Performing Action: Up
Current Agent Position: [1,1] ; Current State Reward: 0 ; Total Cumulative Reward: 25
=============================================================================================================
Timestep: 10                    Performing Action: Up
Current Agent Position: [1,2] ; Current State Reward: 0 ; Total Cumulative Reward: 25
=============================================================================================================
Timestep: 11                    Performing Action: Down
Current Agent Position: [1,1] ; Current State Reward: 0 ; Total Cumulative Reward: 25
=============================================================================================================
Timestep: 12                    Performing Action: Down
Current Agent Position: [1,0] ; Current State Reward: 25 ; Total Cumulative Reward: 50
=============================================================================================================
Timestep: 13                    Performing Action: Down
Current Agent Position: [1,0] ; Current State Reward: 25 ; Total Cumulative Reward: 75
=============================================================================================================
Timestep: 14                    Performing Action: Right
Current Agent Position: [2,0] ; Current State Reward: -5 ; Total Cumulative Reward: 70
=============================================================================================================
Timestep: 15                    Performing Action: Right
Current Agent Position: [3,0] ; Current State Reward: 0 ; Total Cumulative Reward: 70
=============================================================================================================
Timestep: 16                    Performing Action: Right
Current Agent Position: [4,0] ; Current State Reward: 50 ; Total Cumulative Reward: 120
=============================================================================================================
```

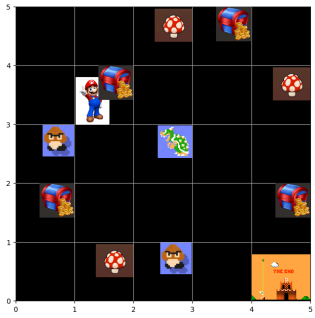
2.1.3 Visualization- Running environment for 16 time steps



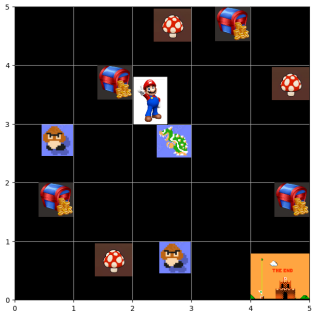
(a) Ts:1 Action:Left +0



(b) Ts:2 Action:Right +0

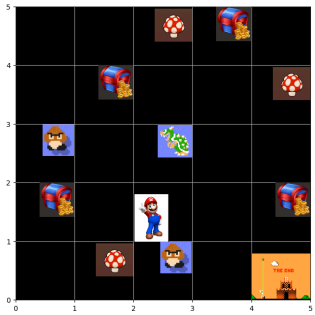


(c) Ts:3 Action:Down +10



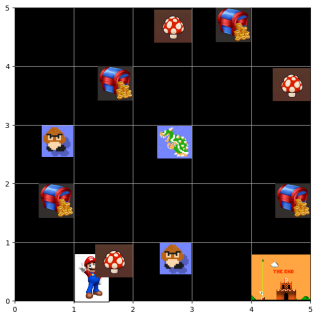
(d) Ts:4 Action:Right +10

(e) Ts:5 Action:Down +0

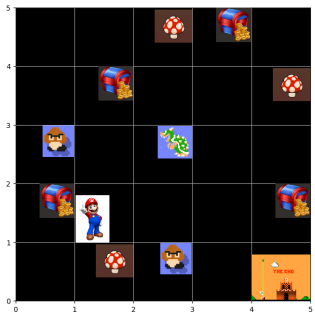


(f) Ts:6 Action:Down +0

(g) Ts:7 Action:Left +0



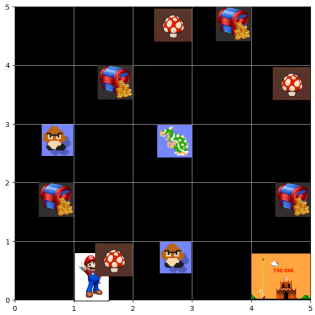
(h) Ts:8 Action:Down +25



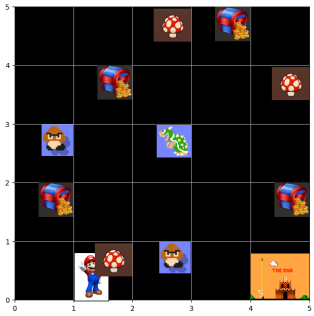
(i) Ts:9 Action:Up +25

(j) Ts:10 Action:Up +25

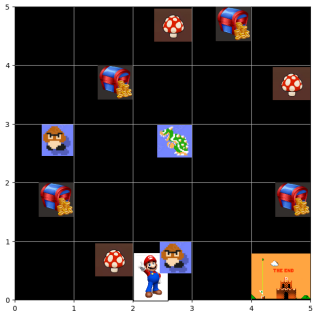
(k) Ts:11 Action:Down +25



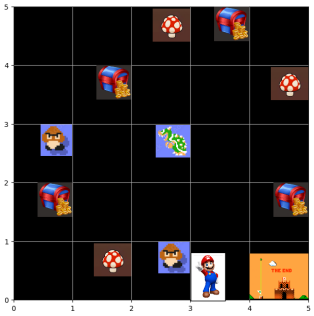
(l) Ts:12 Action:Down +50



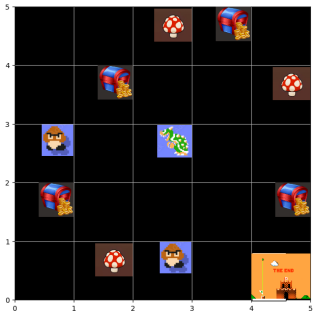
(m) Ts:13 Action:Down +75



(n) Ts:14 Action:Right +70



(o) Ts:15 Action:Right +70

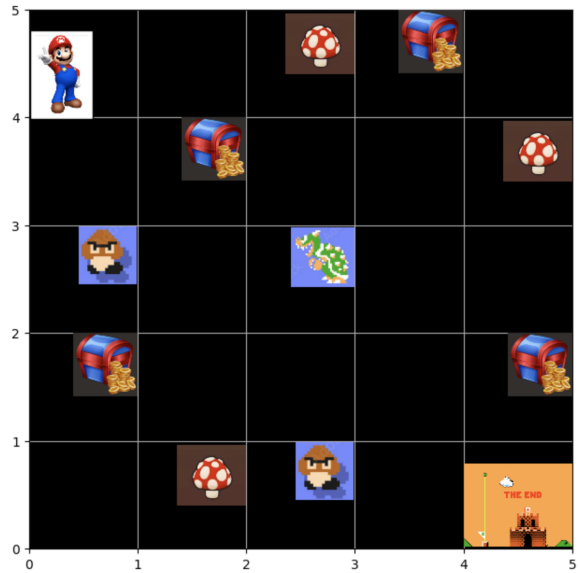


(p) Ts:16 Action:Right +120

2.2 Stochastic Environment

2.2.1 Displaying Stochastic Environment with Agent in its initial position

```
# Initial State Position of the Environment
env = Mario_Game_Stochastic_Environment(16)
env.reset()
env.render()
```

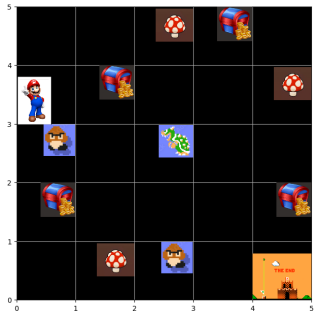


2.2.2 Cumulative rewards collected during 16 time steps

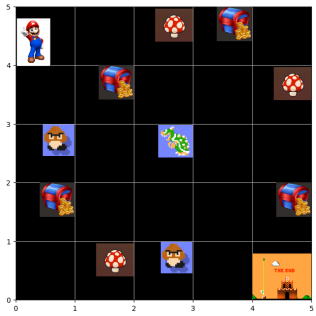
```
#Running the Mario GridWorld Game in Stochastic Environment
done = False
print("\n================================================================================================")
while not done:
    action = random.randint(0,3)
    reward, done, info = env.step(action)
    env.render()

    print("Timestep: {}".format(env.timeStep)+"\t\t\t Performing Action: {}\n".format(action))
    print(info)
    print("\n================================================================================================")
```

```
================================================================================================
Timestep: 1                     Performing Action: 0

Current Agent Position: [0,3] ; Current Reward: 0 ; Total Cumulative Reward: 0

================================================================================================
Timestep: 2                     Performing Action: 1

Current Agent Position: [0,4] ; Current Reward: 0 ; Total Cumulative Reward: 0

================================================================================================
Timestep: 3                     Performing Action: 3

Current Agent Position: [1,4] ; Current Reward: 0 ; Total Cumulative Reward: 0

================================================================================================
Timestep: 4                     Performing Action: 3

Current Agent Position: [2,4] ; Current Reward: 25 ; Total Cumulative Reward: 25

================================================================================================
Timestep: 5                     Performing Action: 1

Current Agent Position: [2,4] ; Current Reward: -1 ; Total Cumulative Reward: 24

================================================================================================
Timestep: 6                     Performing Action: 3

Current Agent Position: [4,4] ; Current Reward: 0 ; Total Cumulative Reward: 24

================================================================================================
Timestep: 7                     Performing Action: 3

Current Agent Position: [4,4] ; Current Reward: -1 ; Total Cumulative Reward: 23

================================================================================================
Timestep: 8                     Performing Action: 1

Current Agent Position: [4,4] ; Current Reward: -1 ; Total Cumulative Reward: 22

================================================================================================
Timestep: 9                     Performing Action: 0

Current Agent Position: [4,3] ; Current Reward: 25 ; Total Cumulative Reward: 47

================================================================================================
Timestep: 10                    Performing Action: 3

Current Agent Position: [4,3] ; Current Reward: -1 ; Total Cumulative Reward: 46

================================================================================================
Timestep: 11                    Performing Action: 3

Current Agent Position: [4,3] ; Current Reward: -1 ; Total Cumulative Reward: 45

================================================================================================
Timestep: 12                    Performing Action: 2

Current Agent Position: [3,3] ; Current Reward: 0 ; Total Cumulative Reward: 45

================================================================================================
Timestep: 13                    Performing Action: 1

Current Agent Position: [3,4] ; Current Reward: 10 ; Total Cumulative Reward: 55

================================================================================================
Timestep: 14                    Performing Action: 2

Current Agent Position: [0,0] ; Current Reward: 0 ; Total Cumulative Reward: 55

================================================================================================
Timestep: 15                    Performing Action: 1

Current Agent Position: [0,1] ; Current Reward: 10 ; Total Cumulative Reward: 65

================================================================================================
Timestep: 16                    Performing Action: 0

Current Agent Position: [0,0] ; Current Reward: 0 ; Total Cumulative Reward: 65

================================================================================================
```

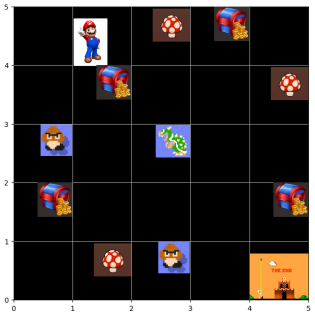
2.2.3 Visualization- Running environment for 16 time steps



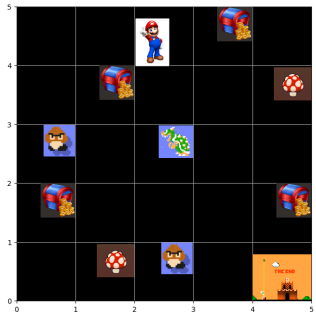
(q) Ts:1 R:0



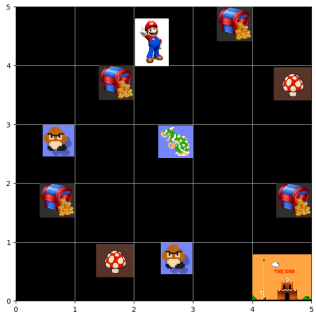
(r) Ts:2 R:0



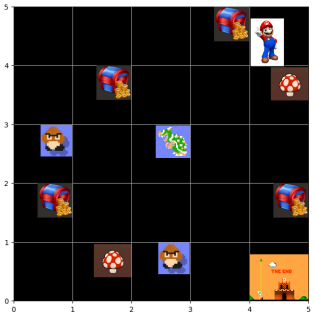
(s) Ts:3 R:0



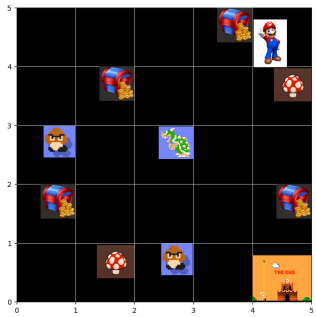
(t) Ts:4 R:+25



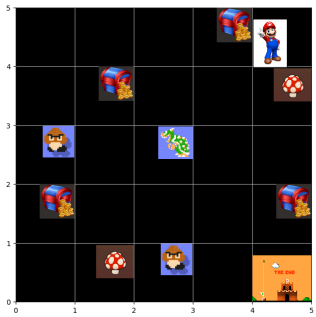
(u) Ts:5 R:+24



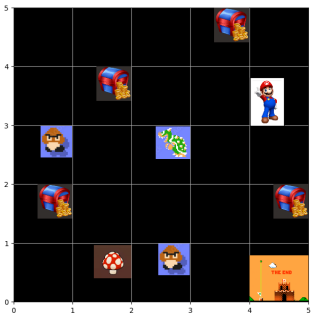
(v) Ts:6 R:+24



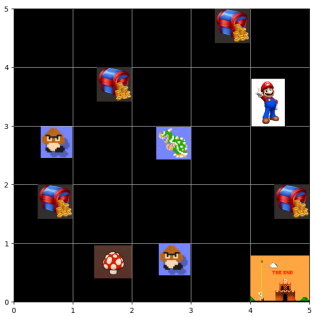
(w) Ts:7 R:+23



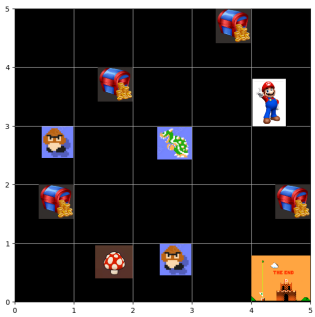
(x) Ts:8 R:+22



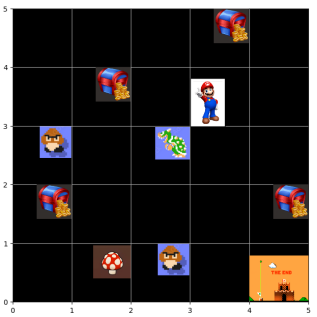
(y) Ts:9 R:+47



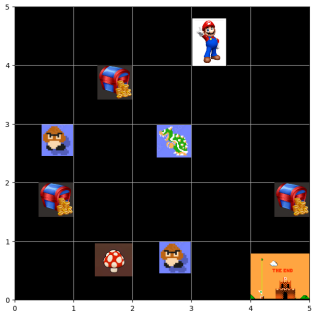
(z) Ts:10 R:+46



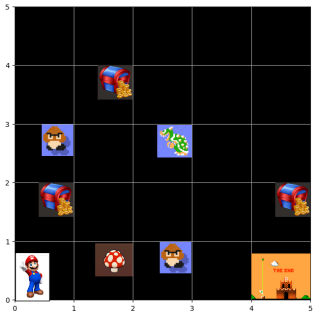
() Ts:11 R:+45



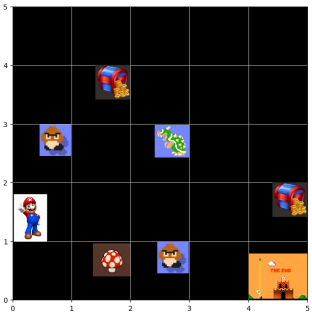
() Ts:12 R:+45



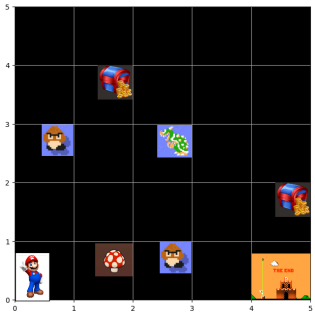
() Ts:13 R:+55



() Ts:14 R:+55



() Ts:15 R:+65



() Ts:16 R:+65

3 How did you define the stochastic environment?

In the current setup the Stochastic environment is designed in a way that the movement of the agent is random and uncertain using probability distribution. Also the immediate rewards disappear as the agent collects them.

4 What is the difference between the deterministic and stochastic environments?

1. Deterministic: When the Agent is in State (s) and performing an Action (a) the probability that it will reach state ($s\prime$) is either 0 or 1.

. $$P(s\prime, r|s, a) = \{0, 1\}$$

2. Stochastic: When the Agent is in State (s) and performing an Action (a) there is a randomness or uncertainty in which state the Agent will end-up.

. $$P(s\prime, r|s, a) = \{0, 1\}$$

5 Safety in AI

While defining the environment certain measures where taken in order to avoid agent running into error states while performing actions. For example, np.clip is used in environment definition to avoid agent going beyond the grid limit during any time-step. Taking care of such error states is very important because in real-world deployments it could lead to fatal incidents.

6 Bonus Tasks

6.1 Git Expert [2 points]:

`https://github.com/DivyaPallineni/CSE546_vpalline`

6.2 CCR Submission:

Submitted Jupyter notebook that is executed on CCR

6.3 Grid-World Scenario Visualization:

As instructed visulizations are captured.

7 References

Resources shared in Piazza

1. `spring_23_rl_lec_3_random_agent.ipynb`
2. `visualizing_rl_environments_and_representing_the_results`

Report is prepared using NIPS template