Artificial Intelligence and Machine Learning

Assignment 03

Student : Kuldeep Jatav

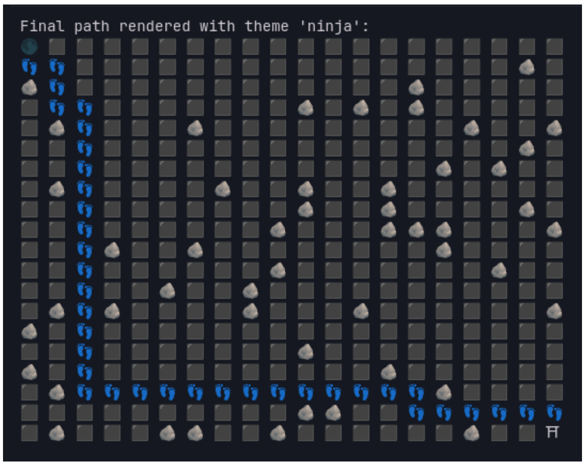
SR Number: 23684

INDIAN INSTITUTE OF SCIENCE

भारतीय विज्ञान संस्थान

Q Learning

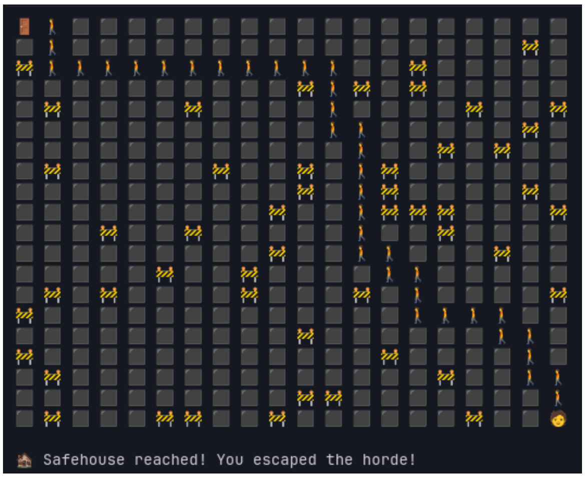
- Completed the `QLAssignment.ipynb` file refer `23684_Assignment3.ipynb` attached with this report.
- I trained the agent on two scenarios, one where traps and boots are disabled and another where they are enabled.
  - **BFS:** The agent is able to reach the goal state in 38 steps using BFS algorithm.



```
Final path rendered with theme 'ninja':
```

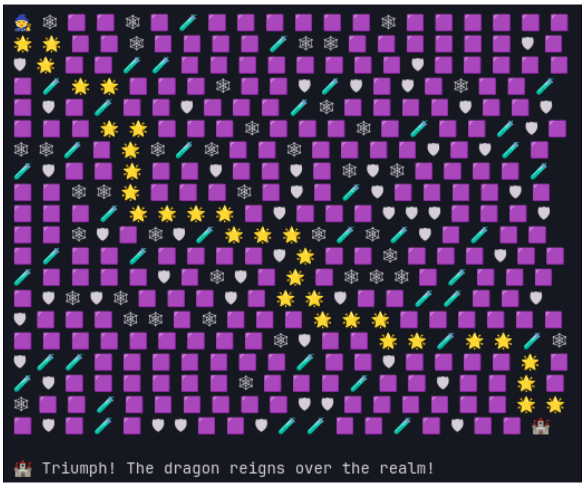
Figure 1: BFS

  - **Traps & Boots:** In both disabled and enabled case the agent took 38 (same as BFS) steps to reach the goal state(see the figure below). for the configurations 1(mentioned in next part of question).



```
Safehouse reached! You escaped the horde!
```

Figure 2: Traps and Boots Disabled



```
Triumph! The dragon reigns over the realm!
```

Figure 3: Traps and Boots Enabled

- **Different reward configurations:**

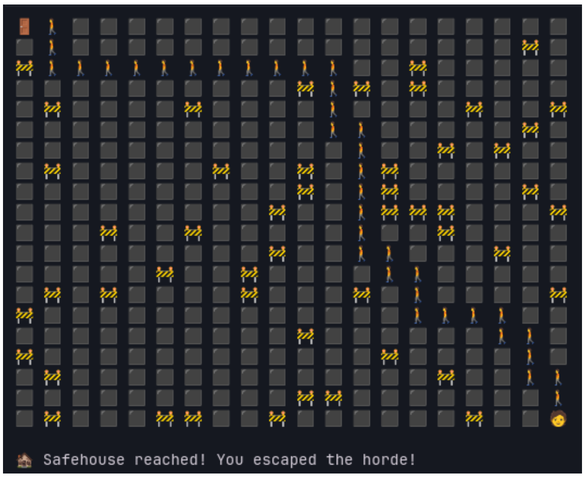
Configuration 1

  - `REWARD_GOAL = 10000`
  - `REWARD_TRAP = -500`
  - `REWARD_OBSTACLE = -100`
  - `REWARD_REVISIT = -200`
  - `REWARD_ENEMY = -2000`
  - `REWARD_STEP = -5`
  - `REWARD_BOOST = 200`

Configuration 2

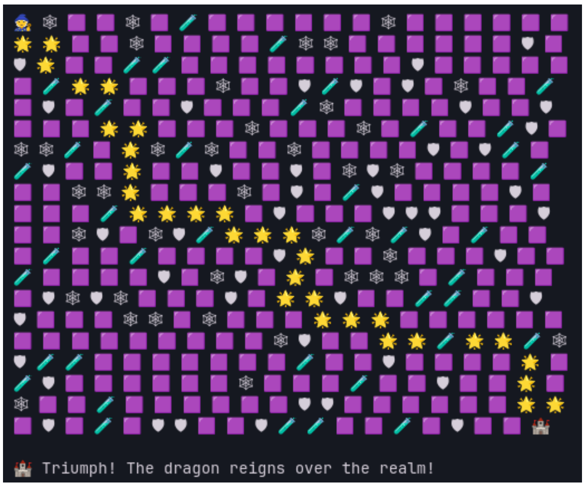
  - `REWARD_GOAL = 5000`
  - `REWARD_TRAP = -500`
  - `REWARD_OBSTACLE = -100`
  - `REWARD_REVISIT = -300`
  - `REWARD_ENEMY = -1500`
  - `REWARD_STEP = -10`
  - `REWARD_BOOST = 250`

Images for configurations 1



```
Safehouse reached! You escaped the horde!
```

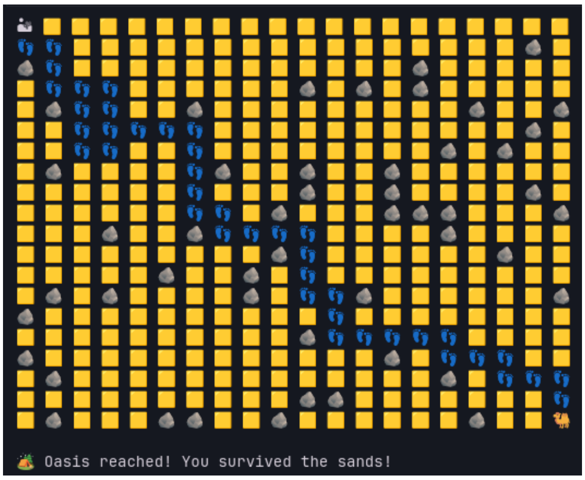
Figure 4: Traps and Boots Disabled



```
Triumph! The dragon reigns over the realm!
```

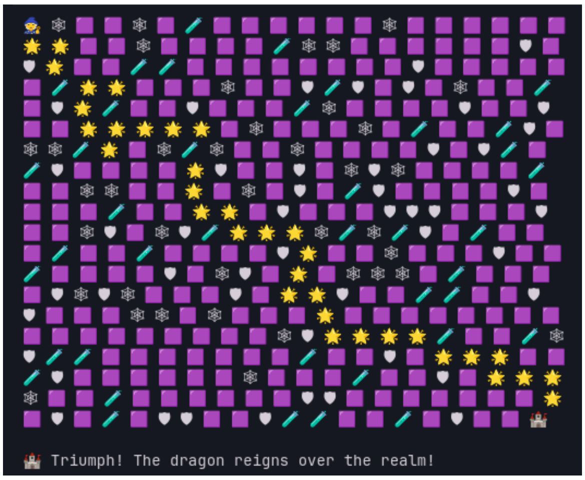
Figure 5: Traps and Boots Enabled

Images for configurations 2



```
Oasis reached! You survived the sands!
```

Figure 6: Traps and Boots Disabled



```
Triumph! The dragon reigns over the realm!
```

Figure 7: Traps and Boots Enabled

- **Manual Q-value Calculation:** For the initial 5 steps, the Q-values are calculated as follows:
  - Learning rate (α) = 0.4
  - Discount factor (γ) = 0.95
  - Rewards:
    - REWARD_STEP = -5
    - REWARD_GOAL = 10000
    - REWARD_OBSTACLE = -100

  We use the Q-learning update rule:

$$Q(s,a) = Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right) \tag{1}$$

Step 1. (0,0) → (1,0)

– Current Q-values for $(0,0)$:

Action →	Up	Down	Left	Right
Q-value →	-767.152172	-719.833527	-801.478881	-631.947255

– Chosen action: Down (max Q-value = -631.947255)

- Reward (r): -5 (valid move to $(1,0)$)
- Next state (s'): $(1,0)$
- $\max Q(s',a') = -621.173244$ (from Q-table at $(1,0)$)
- Update:

$$\begin{aligned} Q_{\text{new}} &= -631.947255 + 0.4\,[-5 + 0.95 \times -621.173244 - (-631.947255)] \\ &= -631.947255 + 0.4\,[-5 - 589.1140828 + 631.947255] \\ &= -631.947255 + 0.4 \times 37.8331722 \\ &= -631.947255 + 15.13326888 \\ &= -616.81398612 \end{aligned}$$

Step 2. (1,0) → (1,1)

- Updated Q-values for $(1,0)$:

Action →	Up	Down	Left	Right
Q-value →	-899.367045	-621.173244	-725.177182	-706.806879

- Chosen action: Right (max Q-value = -621.173244)
- Reward (r): -5 (valid move to $(1,1)$)
- Next state (s'): $(1,1)$
- $\max Q(s',a') = -650.457508$ (from Q-table at $(1,1)$)
- Update:

$$\begin{aligned} Q_{\text{new}} &= -621.173244 + 0.4\,[-5 + 0.95 \times -650.457508 - (-621.173244)] \\ &= -621.173244 + 0.4\,[-5 - 618.9346326 + 621.173244] \\ &= -621.173244 + 0.4 \times -2.9613886 \\ &= -621.173244 - 1.18455544 \\ &= -622.35779944 \end{aligned}$$

Step 3. (1,1) → (1,2)

- Updated Q-values for $(2,0)$:

Action →	Up	Down	Left	Right
Q-value →	-794.747436	-650.457508	-780.813579	-703.640530

- Chosen action: Right (max Q-value = -650.457508)
- Reward (r): -5 (valid move to $(2,1)$)
- Next state (s'): $(2,1)$
- $\max Q(s',a') = -634.087470$ (from Q-table at $(2,1)$)
- Update:

$$\begin{aligned} Q_{\text{new}} &= -650.457508 + 0.4\,[-5 + 0.95 \times -634.087470 - (-650.457508)] \\ &= -650.457508 + 0.4\,[-5 - 602.3830965 + 650.457508] \\ &= -650.457508 + 0.4 \times 43.0744115 \\ &= -650.457508 + 17.2297646 \\ &= -633.2277434 \end{aligned}$$

Step 4. (1,2) → (2,2) (Goal)

- Updated Q-values for $(2,1)$:

Action →	Up	Down	Left	Right
Q-value →	-825.636038	-700.874738	-791.611875	-634.087470

- Chosen action: Right (max Q-value = -634.087470)
- Reward (r): -5 (valid move to $(2,1)$)
- Next state (s'): $(2,1)$

- $\max Q(s', a') = -618.309086$ (from Q-table at $(2, 1)$)
- Update:

$$\begin{aligned} Q_{\text{new}} &= -634.087470 + 0.4 \left[-5 + 0.95 \times -618.309086 - (-634.087470)\right] \\ &= -634.087470 + 0.4 \left[-5 - 586.3936317 + 634.087470\right] \\ &= -634.087470 + 0.4 \times 42.6938383 \\ &= -634.087470 + 17.07753532 \\ &= -617.00993468 \end{aligned}$$

Step 5. (2,2) → (3,2)

- Updated Q-values for $(2, 2)$:

Action →	Up	Down	Left	Right
Q-value →	-729.267172	-661.403121	-866.272165	-618.309086

- Chosen action: Down (max Q-value = -618.309086)
- Reward (r): -5 (valid move to $(3, 2)$)
- Next state (s'): $(3, 2)$
- $\max Q(s', a') = -627.114874e$ (from Q-table at $(3, 2)$)
- Update:

$$\begin{aligned} Q_{\text{new}} &= -618.309086 + 0.4 \left[-5 + 0.95 \times -627.114874 - (-618.309086)\right] \\ &= -618.309086 + 0.4 \left[-5 - 595.7581303 + 618.309086\right] \\ &= -618.309086 + 0.4 \times 17.5509557 \\ &= -618.309086 + 7.02038228 \\ &= -611.28870372 \end{aligned}$$

Thank You