HW3: Solving MDPs

Atharv Kulkarni - u1322897

1 Part 1

```
--- gen_simple_world ---

rewards
1.00    -1.00   0.00    0.00
0.00    -1.00   0.00    0.00
0.00    0.00    0.00    0.00
0.00    0.00    0.00    0.00

visualize a random policy
.       .       v       >
<       .       v       ^
v       >       <       <
v       ^       <       ^

--- value iteration ---

Values from Value Iteration
0.00    0.00    0.42    0.44
0.77    0.00    0.45    0.48
0.71    0.59    0.55    0.51
0.66    0.62    0.58    0.54

Optimal Policy
.       .       >       v
^       .       >       v
^       v       <       <
^       <       <       <

--- policy evaluation ---

Optimal Policy
.       .       >       v
^       .       >       v
^       v       <       <
^       <       <       <

Values from Policy Iteration
0.00    0.00    0.42    0.44
0.77    0.00    0.45    0.48
0.71    0.59    0.55    0.51
0.66    0.62    0.58    0.54
```

```
--- gen_simple_world 2 ---

rewards
1.00    -1.00   0.00    5.00
0.00    -1.00   0.00    0.00
0.00    0.00    0.00    0.00
-1.00   0.00    2.00    0.00

visualize a random policy
.       v       >       .
^       ^       v       >
^       >       ^       ^
.       v       v       v

--- value iteration ---

Values from Value Iteration
0.00    13.05   13.68   0.00
13.41   16.71   18.24   18.19
13.84   18.76   20.72   19.76
0.00    20.94   21.40   21.18

Optimal Policy
.       >       v       .
v       v       v       v
^       >       v       v
.       >       v       <

--- policy evaluation ---

Optimal Policy
.       >       v       .
v       v       v       v
^       >       v       v
.       >       v       <

Values from Policy Iteration
0.00    13.05   13.68   0.00
13.41   16.71   18.24   18.19
13.84   18.76   20.72   19.76
0.00    20.94   21.40   21.18

--- gen_simple_world 3 ---

rewards
5.00    -5.00   -2.00   -2.00
-2.00   -2.00   -2.00   -2.00
-2.00   -2.00   -2.00   -100.00
-2.00   -2.00   -2.00   10.00
```

```
visualize a random policy
.       v       ^       >
^       ^       v       v
<       >       <       >
<       v       >       .

--- value iteration ---

Values from Value Iteration
0.00    3.68    -2.23   -4.37
4.01    0.80    -1.73   -5.59
0.86    0.70    -0.97   -1.94
0.86    3.98    8.22    0.00

Optimal Policy
.       <       <       <
^       <       <       ^
^       v       <       <
>       >       >       .

--- policy evaluation ---

Optimal Policy
.       <       <       <
^       <       <       ^
^       v       <       <
>       >       >       .

Values from Policy Iteration
0.00    3.68    -2.23   -4.37
4.01    0.80    -1.73   -5.59
0.86    0.70    -0.97   -1.94
0.86    3.98    8.22    0.00


--- gen_simple_world 4 ---

rewards
5.00    -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00
-1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00
-1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00
-1.00   -1.00   -1.00   -5.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00
-1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00
-1.00   -1.00   -1.00   -1.00   -1.00   -100.00 -1.00   -1.00   -1.00   -1.00
-1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00
-1.00   -1.00   2.00    -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00
-1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   5.00    -1.00
-1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   -1.00   10.00
```

```
visualize a random policy
.       v       <       v       v       v       <       >       >       <
v       ^       >       <       >       >       >       ^       v       v
^       ^       v       ^       >       <       <       ^       <       v
<       >       ^       ^       v       <       >       <       >       ^
>       ^       v       v       v       >       >       ^       >       >
<       >       <       ^       <       v       ^       >       >       <
<       >       ^       <       v       <       <       <       v       ^
v       v       >       >       ^       v       >       >       v       <
v       <       >       <       >       <       <       ^       ^       ^
v       v       v       ^       >       v       <       >       ^       .

--- value iteration ---

Values from Value Iteration
0.00    4.04    1.86    0.05    -1.47   -2.73   -3.61   -3.29   -2.87   -2.72
4.04    2.15    0.49    -0.98   -2.23   -3.18   -3.03   -2.25   -1.68   -1.54
1.86    0.49    -0.89   -2.16   -3.10   -3.04   -2.06   -0.98   -0.26   -0.17
0.05    -0.95   -1.98   -2.86   -3.26   -2.30   -0.82   0.53    1.42    1.42
-1.44   -2.05   -1.89   -2.36   -2.95   -2.75   0.67    2.30    3.42    3.23
-2.31   -1.46   -0.47   -1.13   -2.00   0.62    2.44    4.36    5.84    5.29
-1.47   -0.14   1.39    0.32    0.84    2.47    4.49    6.74    8.78    7.59
-0.53   1.35    0.43    1.62    2.33    4.37    6.74    9.44    12.46   10.13
-1.33   0.04    1.59    1.57    3.43    5.84    8.78    12.46   10.50   12.83
-2.13   -1.05   0.14    1.44    3.23    5.29    7.59    10.13   12.83   0.00

Optimal Policy
.       <       <       <       <       <       <       v       v       v
^       ^       <       <       <       <       v       v       v       v
^       ^       <       ^       <       >       v       v       v       v
^       ^       <       <       >       >       v       v       v       v
^       ^       v       v       v       ^       >       v       v       v
>       v       v       v       <       v       >       v       v       v
>       >       v       v       v       v       v       v       v       v
>       >       >       <       >       >       >       v       v       v
>       >       ^       >       >       >       >       >       >       <
>       >       >       >       >       >       >       >       ^       .

--- policy evaluation ---

Optimal Policy
.       <       <       <       <       <       <       v       v       v
^       ^       <       <       <       <       v       v       v       v
^       ^       <       ^       <       >       v       v       v       v
^       ^       <       <       >       >       v       v       v       v
^       ^       v       v       v       ^       >       v       v       v
>       v       v       v       <       v       >       v       v       v
>       >       v       v       v       v       v       v       v       v
>       >       >       <       >       >       >       >       v       v
>       >       ^       >       >       >       >       >       v       <
>       >       >       >       >       >       >       >       ^       .
```

```
Values from Policy Iteration
0.00    4.04    1.86    0.05    -1.47   -2.73   -3.61   -3.29   -2.87   -2.72
4.04    2.15    0.49    -0.98   -2.23   -3.18   -3.03   -2.25   -1.68   -1.54
1.86    0.49    -0.89   -2.16   -3.10   -3.04   -2.06   -0.98   -0.26   -0.17
0.05    -0.95   -1.98   -2.86   -3.26   -2.30   -0.82   0.53    1.42    1.42
-1.44   -2.05   -1.89   -2.36   -2.95   -2.75   0.67    2.30    3.42    3.23
-2.31   -1.46   -0.47   -1.13   -2.00   0.62    2.44    4.36    5.84    5.29
-1.47   -0.14   1.39    0.32    0.84    2.47    4.49    6.74    8.78    7.59
-0.52   1.35    0.43    1.62    2.33    4.37    6.74    9.44    12.46   10.13
-1.33   0.04    1.59    1.57    3.43    5.84    8.78    12.46   10.50   12.83
-2.13   -1.05   0.14    1.44    3.23    5.29    7.59    10.13   12.83   0.00
```

---

2 Part 2

2.1 Analysis of `gen_simple_world()`

2.1.1 Code Changes:

- $\gamma = 0.95$ (encourages long-term planning).
- **Noise:** 0.1 (introduces some randomness in movement).
- **Terminal states:** $[0, 1, 5]$.
- **Rewards:**

$$\begin{bmatrix} 1.00 & -1.00 & 0.00 & 0.00 \\ 0.00 & -1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$$

2.1.2 Model Behavior:

- The agent tries to reach the highest reward while avoiding penalties.
- The **optimal policy** is:

$$\begin{bmatrix} . & . & \rightarrow & \downarrow \\ \uparrow & . & \rightarrow & \downarrow \\ \uparrow & \downarrow & \leftarrow & \leftarrow \\ \uparrow & \leftarrow & \leftarrow & \leftarrow \end{bmatrix}$$

Explanation:

- The agent **moves right** initially since it avoids the -1 penalty at [1].
- It then moves downwards to collect the most reward while avoiding unnecessary risks.
- The policy is stable, meaning the agent converges to this behavior consistently.

—

2.2 Analysis of `gen_simple_world2()`

2.2.1 Code Changes:

- $\gamma = 0.95$ (encourages long-term planning).
- Increased **noise** to 0.2 (more randomness).
- Changed **terminal states** to $[0, 3, 12]$.
- Adjusted **rewards**:

$$\begin{bmatrix} 1.00 & -1.00 & 0.00 & 5.00 \\ 0.00 & -1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ -1.00 & 0.00 & 2.00 & 0.00 \end{bmatrix}$$

2.2.2 Model Behavior:

- The +5 reward at [3] attracts the agent strongly.
- The +2 reward at [14] also influences movement but to a lesser extent.
- The **optimal policy** is:

$$\begin{bmatrix} . & \rightarrow & \downarrow & . \\ \downarrow & \downarrow & \downarrow & \downarrow \\ \uparrow & \rightarrow & \downarrow & \downarrow \\ . & \rightarrow & \downarrow & \leftarrow \end{bmatrix}$$

Explanation:

- The agent moves right aggressively toward [3] to maximize rewards.
- It follows a downward path through [14] to collect the additional +2.
- Increased **noise** makes actions less deterministic, meaning the agent sometimes explores more than in `gen_simple_world()`.

—

2.3 Analysis of `gen_simple_world3()`

2.3.1 Code Changes:

- Reduced γ to 0.85 (short-term planning favored).
- Added a -100 penalty at [11] (huge obstacle).
- Defined **terminal states** as $[0, 15]$.
- Adjusted **rewards**:

$$\begin{bmatrix} 5.00 & -5.00 & -2.00 & -2.00 \\ -2.00 & -2.00 & -2.00 & -2.00 \\ -2.00 & -2.00 & -2.00 & -100.00 \\ -2.00 & -2.00 & -2.00 & 10.00 \end{bmatrix}$$

2.3.2 Model Behavior:

- The agent strongly avoids [11] due to the huge penalty.
- The goal state at [15] attracts movement.
- The **optimal policy** is:

$$\begin{bmatrix} . & \leftarrow & \leftarrow & \leftarrow \\ \uparrow & \leftarrow & \leftarrow & \uparrow \\ \uparrow & \downarrow & \leftarrow & \leftarrow \\ \rightarrow & \rightarrow & \rightarrow & . \end{bmatrix}$$

Explanation:

- The agent starts moving leftward in the first two rows, avoiding negative rewards.
- In the third row, the agent takes a downward step at $[2, 1]$ before continuing left, demonstrating an alternate risk-averse strategy.
- In the last row, the agent moves rightward towards the goal state at [15], avoiding the high penalty at [11].
- Since $\gamma = 0.85$, the agent prioritizes safer short-term rewards over long-term optimization.
- The agent **completely avoids state** [11], confirming that the -100 penalty effectively deters it from that path.

—

2.4 Analysis of `gen_simple_world4()` (10x10 Grid)

2.4.1 Code Changes:

- **Larger state space** (10×10 grid).
- Added -100 penalty at $[5, 5]$ (major obstacle).
- Increased **noise** to 0.15 (more randomness).
- Defined **terminal states** as $[0, 99]$.
- $\gamma = 0.9$ (encourages long-term planning).
- Adjusted **rewards**:

$$\begin{bmatrix} 5.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 \\ -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 \\ -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 \\ -1.00 & -1.00 & -1.00 & -5.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 \\ -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 \\ -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -100.00 & -1.00 & -1.00 & -1.00 & -1.00 \\ -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 \\ -1.00 & -1.00 & 2.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 \\ -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & 5.00 & -1.00 \\ -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & 10.00 \end{bmatrix}$$

2.4.2 Model Behavior:

- The agent avoids $[5, 5]$ at all costs.
- The goal at $[9, 9]$ is prioritized.
- The **optimal policy** is:

$$\begin{bmatrix}
. & \leftarrow & \leftarrow & \leftarrow & \leftarrow & \leftarrow & \leftarrow & \downarrow & \downarrow & \downarrow \\
\uparrow & \uparrow & \leftarrow & \leftarrow & \leftarrow & \leftarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\uparrow & \uparrow & \leftarrow & \uparrow & \leftarrow & \rightarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\uparrow & \uparrow & \leftarrow & \leftarrow & \rightarrow & \rightarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\uparrow & \uparrow & \downarrow & \downarrow & \downarrow & \uparrow & \rightarrow & \downarrow & \downarrow & \downarrow \\
\rightarrow & \downarrow & \downarrow & \downarrow & \leftarrow & \downarrow & \rightarrow & \downarrow & \downarrow & \downarrow \\
\rightarrow & \rightarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\rightarrow & \rightarrow & \rightarrow & \leftarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \downarrow & \downarrow \\
\rightarrow & \rightarrow & \uparrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \downarrow & \leftarrow \\
\rightarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \uparrow & .
\end{bmatrix}$$

Explanation:

- The agent strictly avoids $[5, 5]$ due to the severe -100 penalty, taking a longer but safer route.
- The goal at $[9, 9]$ is prioritized, with movement becoming more direct in the lower rows.
- The +5 reward at $[8, 8]$ and +2 at $[7, 2]$ influence movement but do not outweigh reaching $[9, 9]$.
- Early movement is leftward, mid-movement is downward, and final movement is rightward, optimizing risk avoidance.
- Higher noise (0.15) causes occasional detours, but the main strategy remains stable.
- $\gamma = 0.9$ ensures long-term rewards are considered, preventing loops in suboptimal areas.