HW3: MDPS

Authors: **Jainta Paul & Yeaseen Arafat**

Part 1:

Code repo: https://github.com/pauljainta/mdps

Output from the original:

```
(myenv) ➜  ~/Y_E_@_S_E_EN/solving-mdps/mdps git:(main) ✗ python3 homework3.py
Rewards in the grid:
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
Visualizing a random policy:
.       .       ^       >
<       .       <       ^
^       ^       <       ^
<       <       >       >
--- Starting Value Iteration ---
State values after Value Iteration:
1.00    -1.00   0.82    0.87
1.51    -1.00   0.89    0.93
1.39    1.15    1.06    1.00
1.29    1.21    1.12    1.05
Optimal Policy from Value Iteration:
.       .       >       v
^       .       >       v
^       v       <       <
^       <       <       <
--- Starting Policy Iteration ---
Optimal Policy from Policy Iteration:
.       .       >       v
^       .       >       v
^       v       <       <
^       <       <       <
State values from Policy Iteration:
1.00    -1.00   0.82    0.87
1.51    -1.00   0.89    0.93
1.39    1.15    1.06    1.00
1.29    1.21    1.12    1.05
```

Key observations from the original output:

- Policy drives agents toward the top-left terminal (+1) while avoiding negative terminals (-1)
- Value gradient shows highest values near the +1 terminal state, decreasing with distance
- Terminals at [0,1,5] create asymmetric navigation patterns
- Values stay positive in non-terminal states due to the accessibility of +1 reward

- Policy steers clear of negative terminals by creating safe corridors through zero-reward states

Part 2:

1. **gen_simple_world2 vs original:**

- **Overview of Changes in gen_simple_world2():**
  - **Rewards:** A high positive reward (+2) was added in one of the non-terminal states, and a high negative reward (-2) was added in another non-terminal state.
  - **Terminal States:** Adjusted terminal states include the top-left and bottom-right corner with rewards of +1 and -2, respectively.
  - **Noise:** Noise was increased from 0.1 to 0.2
  - The policy takes more direct routes due to the higher penalty risk from noise.
  - V-values show a transparent gradient toward a high-reward terminal.
- **Comparison with Original MDP (Part 1):**
  - **Original Policy Behavior**: Likely more uniform or balanced without extreme rewards or penalties, leading to more conservative strategies.
  - **Modified Policy Behavior**: More aggressive or risk-averse strategies tailored to the specific placement of rewards and penalties.
- **Does the Change in the Optimal Policy Make Sense?** Yes, the changes make sense as they logically align to maximize rewards while minimizing penalties, demonstrating the agent's adaptability to environmental cues.
- **Impact of Tweaks**: The tweaks in the MDP setup significantly impact the optimal policy, showcasing how sensitive optimal policies are to the reward structure and the location of terminal states within the MDP.

2. **gen_simple_world3 vs original:**

- **Overview of Changes in gen_simple_world3():**
  - **Rewards**: A high positive reward (+2) was added in the terminal state at the top right, while negative rewards (-1) were placed in two lower grid states, adding strategic complexities.
  - **Terminal States**: Adjusted terminal states now include the top-left corner with a reward of +1 and the top-right corner with a reward of +2, effectively making the top row the focal point for terminal rewards.
  - **Noise**: Noise was reduced to 0.05, allowing for more predictable and strategic movements towards reward-bearing states.
  - Policy routes agents around negative terminals while providing access to positive ones.
  - V-values show the barrier effect of the negative terminal and attraction to the positive one.
- **Comparison with Original MDP (Part 1):**

    - **Original Policy Behavior**: The original setup likely featured more uniformly distributed minor rewards and penalties, leading to more evenly spread strategies across the grid.
    - **Modified Policy Behavior**: The new setup encourages more calculated, direct routes towards high-reward states, especially evident in the avoidance of negative rewards and strategic advancement towards the top-right terminal reward.
- **Does the Change in the Optimal Policy Make Sense?** Yes, the adjustments in rewards and terminal states logically drive the agent to optimize pathways that maximize rewards and minimize encounters with penalties. This demonstrates an enhanced adaptability and strategic planning in response to the altered environmental cues.
- **Impact of Tweaks**: These tweaks significantly alter the optimal policy, underlining the high sensitivity of optimal policies to the specific placement of rewards and terminal states. The reduction in noise further accentuates this by allowing the agent to more reliably pursue the most rewarding paths without as much risk of unintended movements.

Both examples demonstrate how optimal policies sensibly adapt to changes in MDP parameters (noise, rewards, terminal locations) while maintaining reasonable risk-reward trade-offs.