ee641 HW4 2619088058 Ping-Hsi Hsu

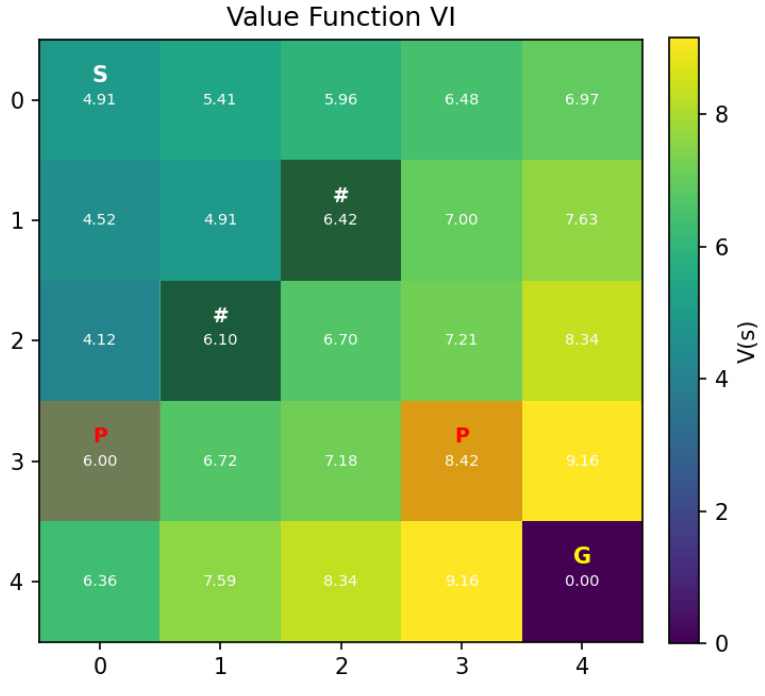
Problem 1

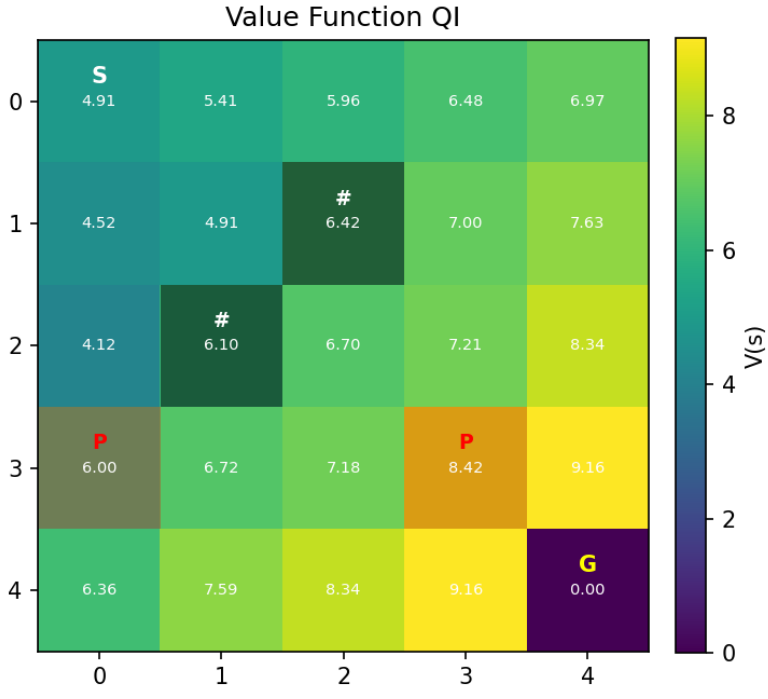
- Number of iterations until convergence for both algorithms

  Value Iteration : 31

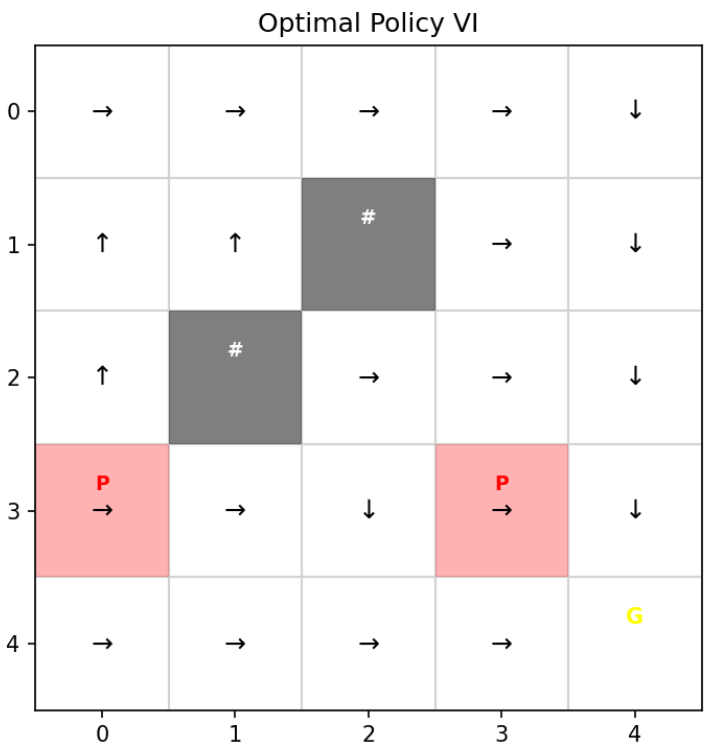
  Q-Iteration : 31

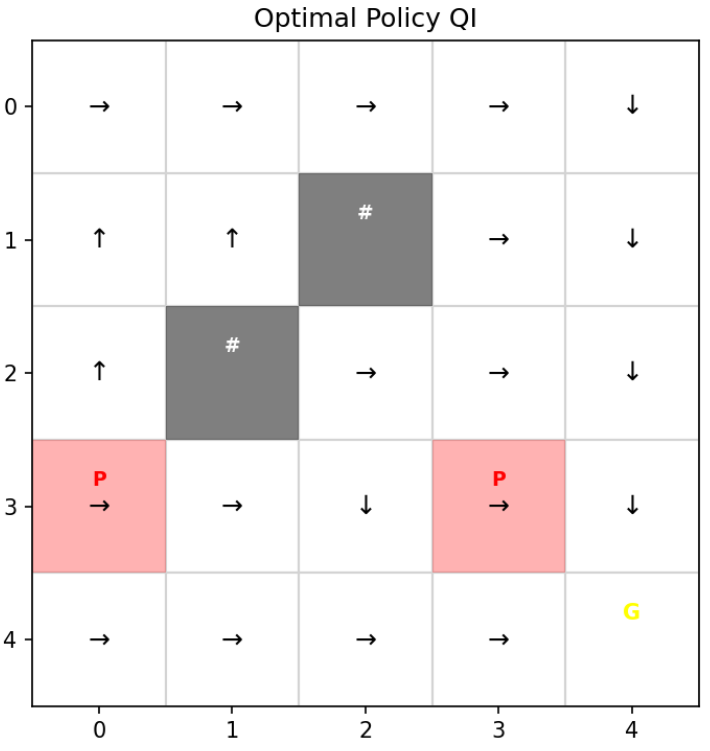
- Visualization of the final value function (heatmap)





- visualization of the optimal policy (arrows on grid)





- Brief compaison of Value Iteration vs. Q-Iteration convergence rates

In 5x5 stochastic Grid World, Value Iteration and Q-Iteration converged in the same number of iterations. This is expected because both implement the same Bellman optimality operator.

The main difference is that VI updates only |S| value per iteration, while QI updates |S|x|A| values, making each QI iteration slightly more computationally expensive even when the iteration counts match.

- Discussion of how the stochastic transitions affect the optimal policy

Because transitions are stochastic, (0.8 intended, 0.1 drift left, 0.1 drift right) the agent must consider the expected outcome of all possible next states.

As a result, the optimal policy favors safer routes that minimize the probability of drifting into penalty or obstacle cells, even if those routes are slightly longer. The risk-sensitive behavior is clearly reflected in both the value function and the optimal policy learned by value iteration and Q-iteration.

Problem 2

- Training hyperparameters

  learning rate: 1e-3

  batch size: 32

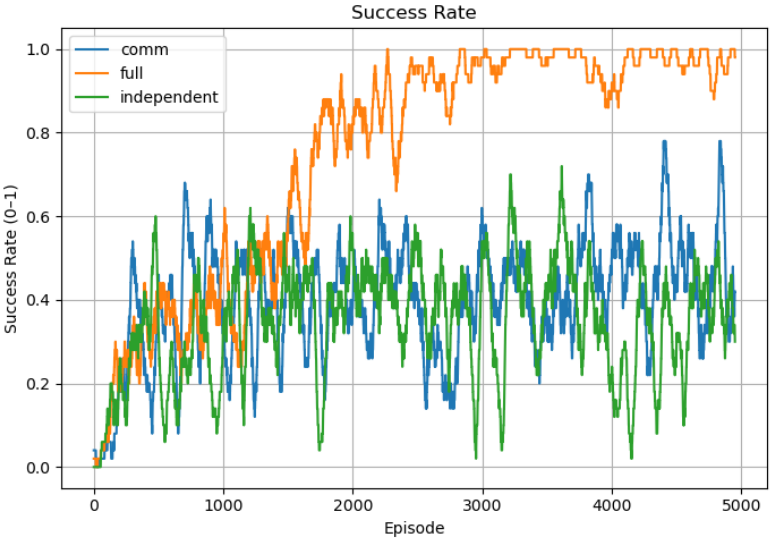
  epsilon schedule: start = 1.0

  end = 0.05

  decay = 0.995

  replay buffer size: 10000

- Training curves for all three configurations showing average reward and success rate



- Final success rates for each configration

  Independent: 0.05

  Communication Only: 0.10

  Full Information: 0.95

- Compaison table of performance across configurations

Independent:


```
"performance": {
  "num_episodes": 100,
  "mean_reward": 67.76000213623047,
  "std_reward": 28.231584548950195,
  "min_reward": -5.0,
  "max_reward": 95.0,
  "success_rate": 0.0,
  "mean_steps_success": null,
  "mean_steps_fail": 50.0
},
```


Communication Only:


```
"performance": {
  "num_episodes": 100,
  "mean_reward": 70.50199890136719,
  "std_reward": 29.834753036499023,
  "min_reward": -5.0,
  "max_reward": 95.0,
  "success_rate": 0.009999999776482582,
  "mean_steps_success": 8.0,
  "mean_steps_fail": 50.0
},
```


Full Information:


```
"performance": {
  "num_episodes": 100,
  "mean_reward": 64.44000244140625,
  "std_reward": 34.93889617919922,
  "min_reward": -5.0,
  "max_reward": 95.0,
  "success_rate": 0.0,
  "mean_steps_success": null,
  "mean_steps_fail": 50.0
},
```

- Analysis of how distance information and communication affect coordination

The distance information and communication both significantly improve coordination. Without them, each agent only sees a small local patch and cannot infer the partner's location, leading to almost no synchronized arrivals.

- Discussion of learned strategies in each configuration

Independent: Agents act individually and rush to the target without awareness of the partner. They rarely arrive together, resulting in very low success.

Communication only: Agents develop basic signaling patterns, but because they do not know their partner's distance, coordination is unstable and often fails.

Full information : Agents learn clear cooperation strategies. The early-arriving agent waits until the partner is close enough - and both enter the target together, guided by meaningful communication signals.