Report

Value and Q-Iteration on Stochastic Gridworld:

After executing both Value and Q-Iteration on the provided Stochastic Gridworld, they converged to the same policy in 31 iterations. The value function heatmap and optimal policy are identical, shown below. The obstacle, penalty, start, and goal cells are highlighted on the heatmap.

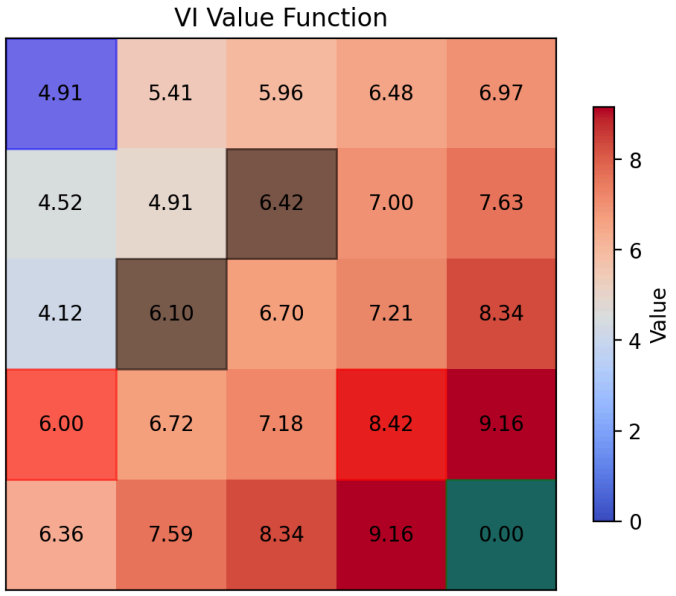

Figure 1: Value Iteration Final Value Function

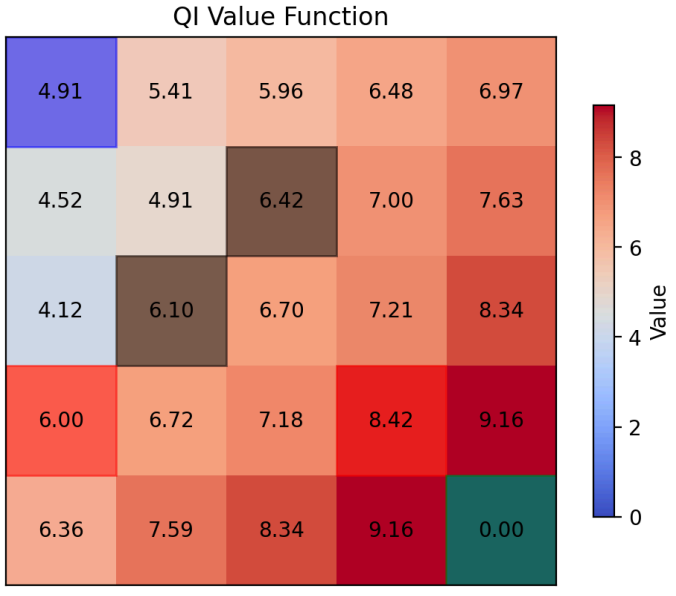

Figure 2: Q-Iteration Final Value Function

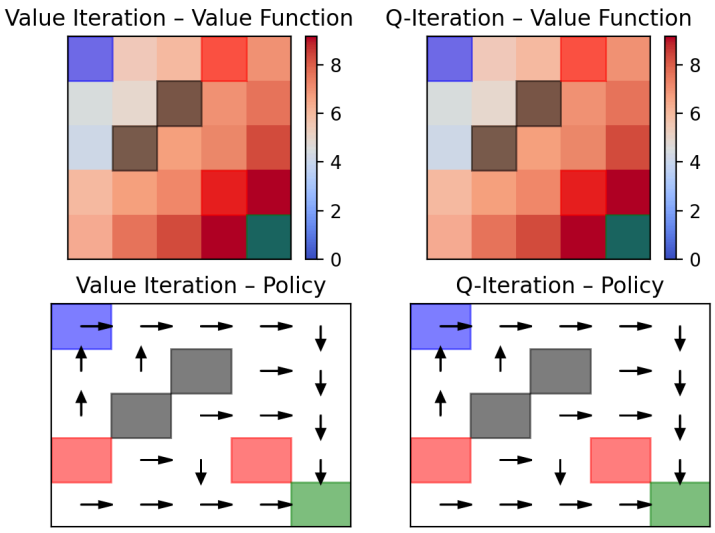

Figure 3: Comparison of Final Value Function and Optimal Policy

This result is to be expected, as the system only has 25 states and the gridworld remains unchanged, so the optimal policy of both V and Q-iteration will converge to a single most optimal policy. Where the algorithms differ is in their convergence rate, shown in Figure 4.

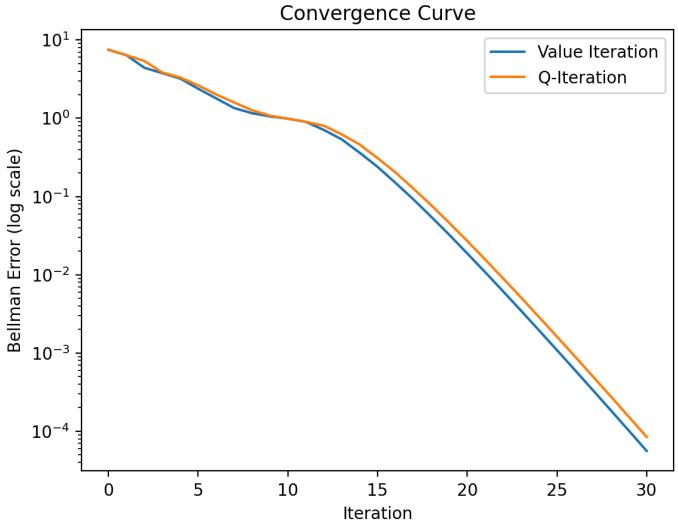

Figure 4: Convergence Curve

The Value Iteration converges slightly faster in comparison to the Q-Iteration due to the fact that Q-iteration updates once for each (state, action) pair, causing information to propagate more slowly and sometimes oscillate. This difference was shown to be not significant in this particular environment, as both agents reached the target on the same iteration.

The addition of stochastic transition introduces randomness into the system, and since the agents are heavily penalized for stepping on the penalty cell, the resulting policy will try to avoid the penalty cell as much as possible. If we take another look at the policy in Figure 3, it is clear that the only time the arrows don't point away from the penalty cells is when it is necessary to move in that direction. Therefore, stochastic transition results in value gradients and policies that prioritize robustness over shortest-path efficiency.

Multi-Agent Coordination with Limited Information:
The training hyperparameters used are as follows:
Learning Rate: 1e-3
Batch Size: 32
Epsilon Schedule: From 1.0 to 0.05, Exponential Decay of 0.999
ReplayBuffer Size: 10000

For the full information mode:

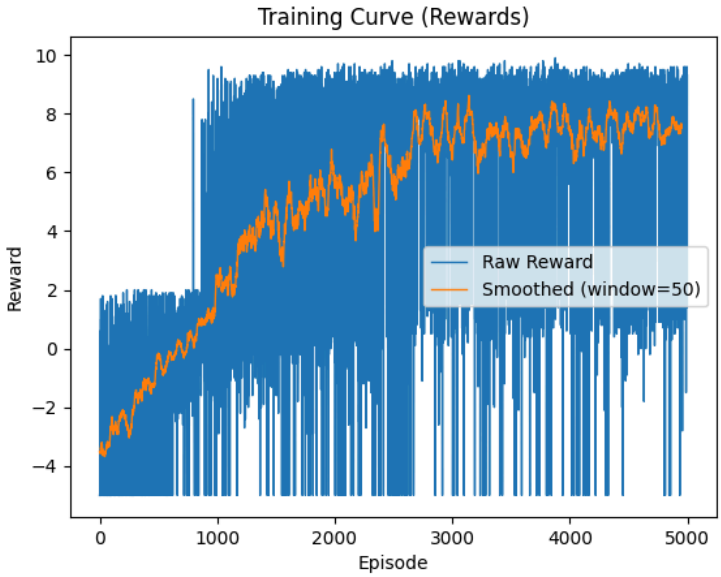

Figure 5: Training Curve for Full Information mode

Final evaluation over 20 episodes: mean reward = 8.88, success rate = 1.00

For communication only mode:

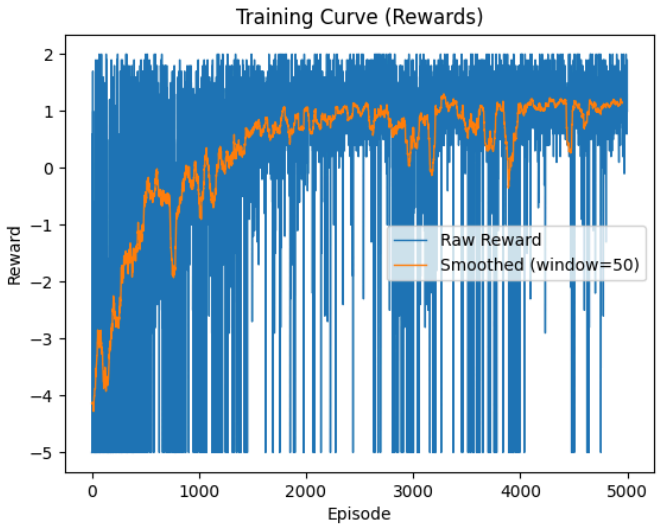

Figure 6: Training Curve for Communication Only Mode

Final evaluation over 20 episodes: mean reward = 0.59, success rate = 0.00

For Independent Mode:

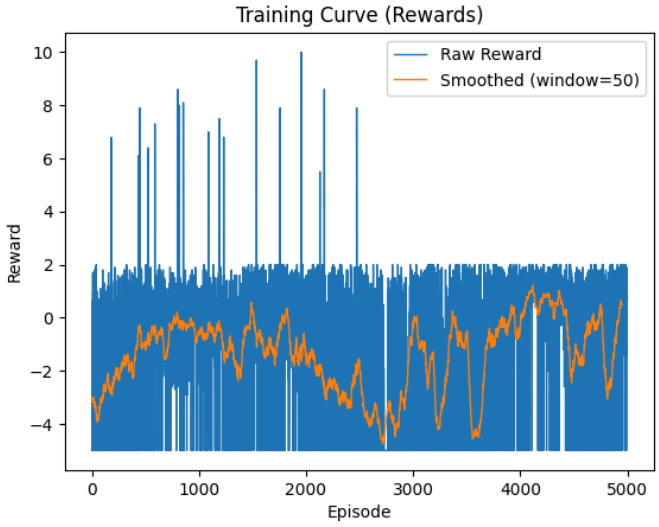

Figure 7: Training Curve for Independent Mode

Final evaluation over 20 episodes: mean reward = 0.14, success rate = 0.00

	Full Information	No Distance	Independent
Mean Reward	-4.66	-4.79	-4.99
Success Rate	0%	0%	0
Mean Steps	47.64	48.55	50

Table 1: Comparison of Different Configurations

The model has a clear problem of only learning the grid through exploring every route and memorizing the layouts, and fails to adapt to any new grid without relearning from scratch. I believe this is due to the fact that the task heavily depends on the layout of the grid, so it is hard to adapt to a completely new grid. However, that is what the agent is expected to learn through the training: the desired strategy is to first meet at a common rendezvous point, then jointly search for the target. The model fails to do so, and this conclusion is further supported by the learning of the configurations with reduced communication, where the agents only learns to separately move toward the target and not arrive together through all 5000 episodes, thus resulting in a reward of around 1. To mitigate this issue, I implemented randomizing the target location every episode at one point, but that would lead to the model not learning any information. I also tweaked the hyperparameters around, but it does not seem to resolve this issue, nor does using the DuelingDQN or PrioritizedReplayBuffer.

To fix this issue, I suspect changing the reward structure so it encourages reducing the distance between the agents would help. For example, even if the model fails to reach the target together, the greater the distance between the agents, the larger the final penalty. Without this reward structure, the agents would default toward independent searching and memorizing.

The communication and distance channel was helpful for the agents to coordinate their timing to reach their target together. This is apparent because without the distance channel, the agents were never able to reach the targets together during training. Looking at Figure 5, after one agent finds the target location, the other is quickly aware of the location through communication, as reflected by the fact that the reward never stagnates around 1, like the training configurations without communication. However, even with full information, the agents do not learn to meet first, which I believe is necessary for strong generalization. Additionally, since the configurations of only communication and independent exhibit almost the same training curve, it can be concluded that the distance channel is more important than the communication channel in terms of coordination. That would make sense since communication would represent the intention of the other agents, and therefore would be useless unless the agent knows where the other is. To analyze this phenomenon, I would also add the distance channel only to the ablation study.