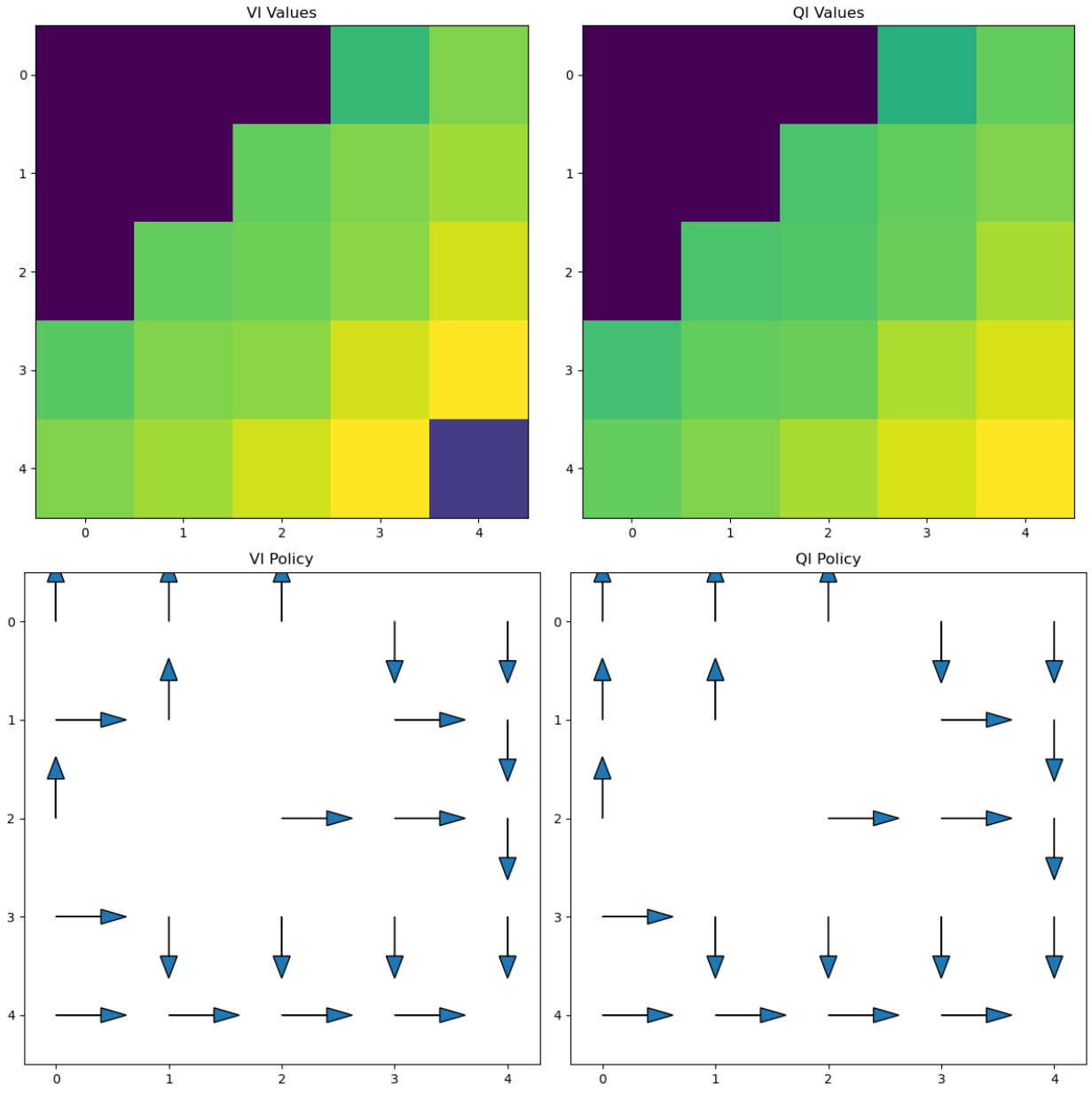
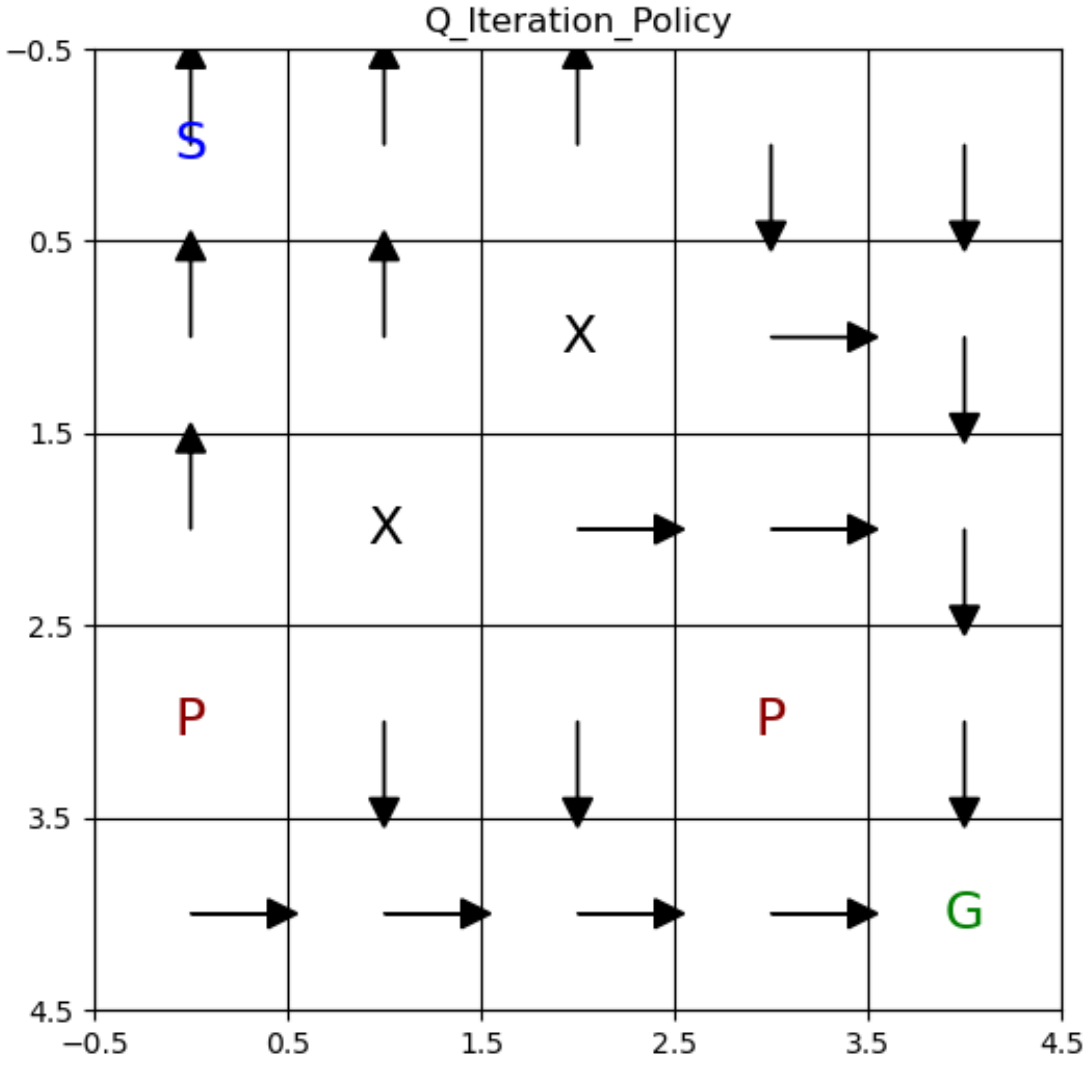
Problem 1:

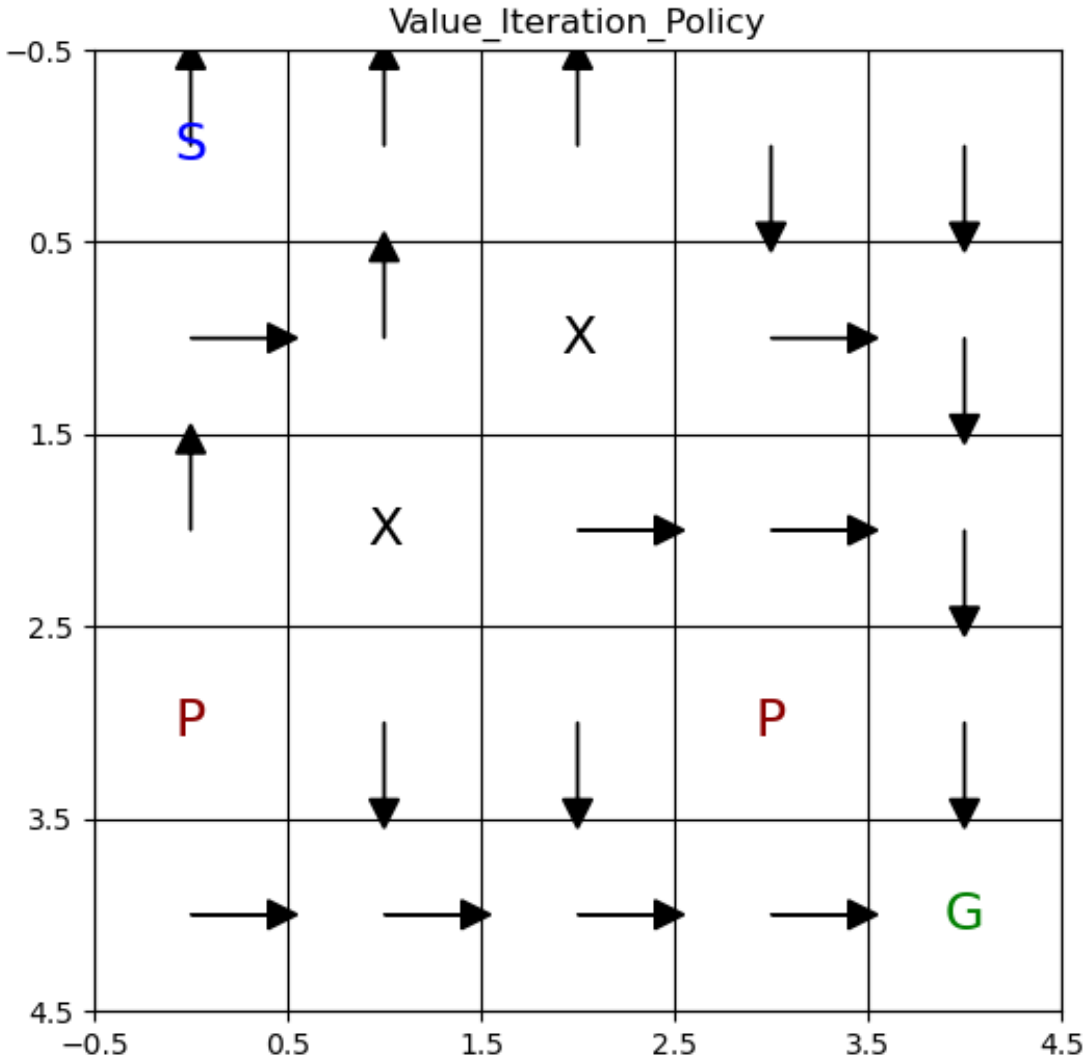
Both algorithms converge after 136 iteration

In general Value Iteration converges fasters as its not working in state-space but here as both algorithms took the same number of steps to converge to very similar solutions it is hard to comment on their individual convergence rates.

The stochastic transitions causes the optimizer to find the most conservative paths (moving away from hazards). This is very clearly visible as the policy near penalties is to avoid them and near boundaries is to go towards them as it is safer.




Q_Iteration_Policy
S
X
X
P
P
G
−0.5
0.5
1.5
2.5
3.5
4.5
−0.5
0.5
1.5
2.5
3.5
4.5

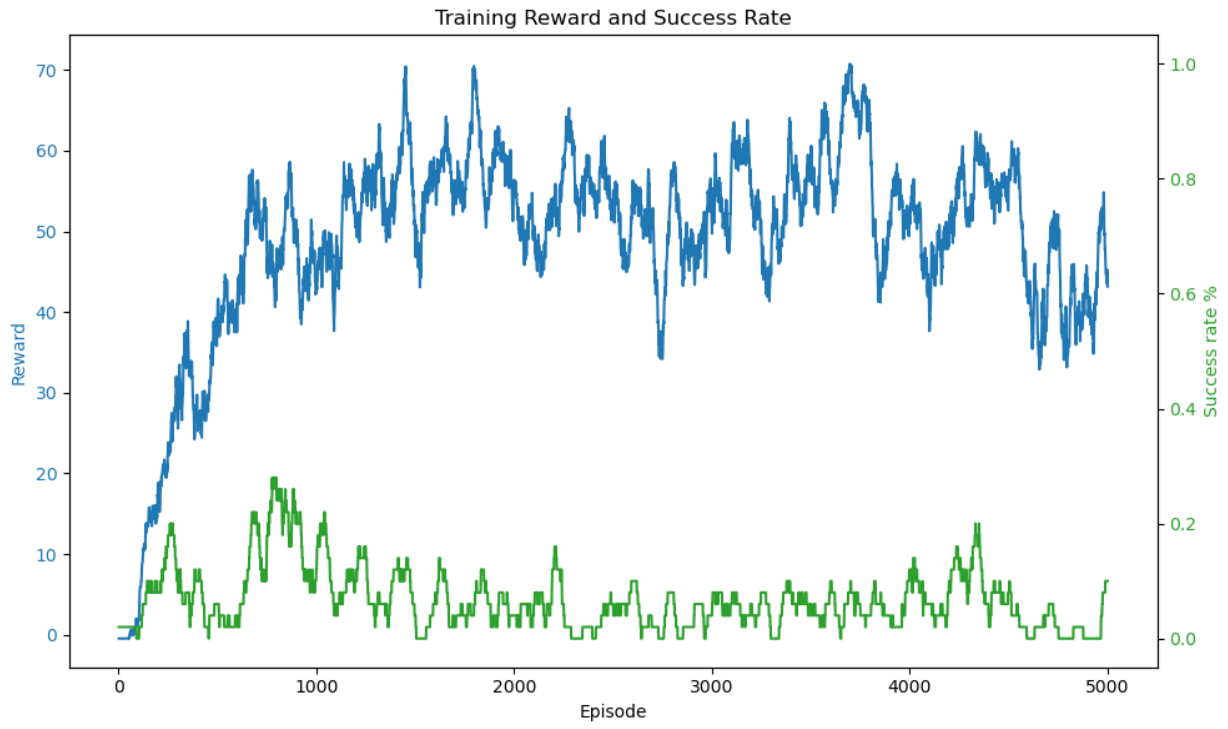

Value_Iteration_Policy
S
X
X
P
P
G
−0.5
0.5
1.5
2.5
3.5
4.5
−0.5
0.5
1.5
2.5
3.5
4.5

Problem 2:
Each model was trained with the default parameters.
The success rates for each model was very low <5%.
There was some emergent behavior of each configuration:

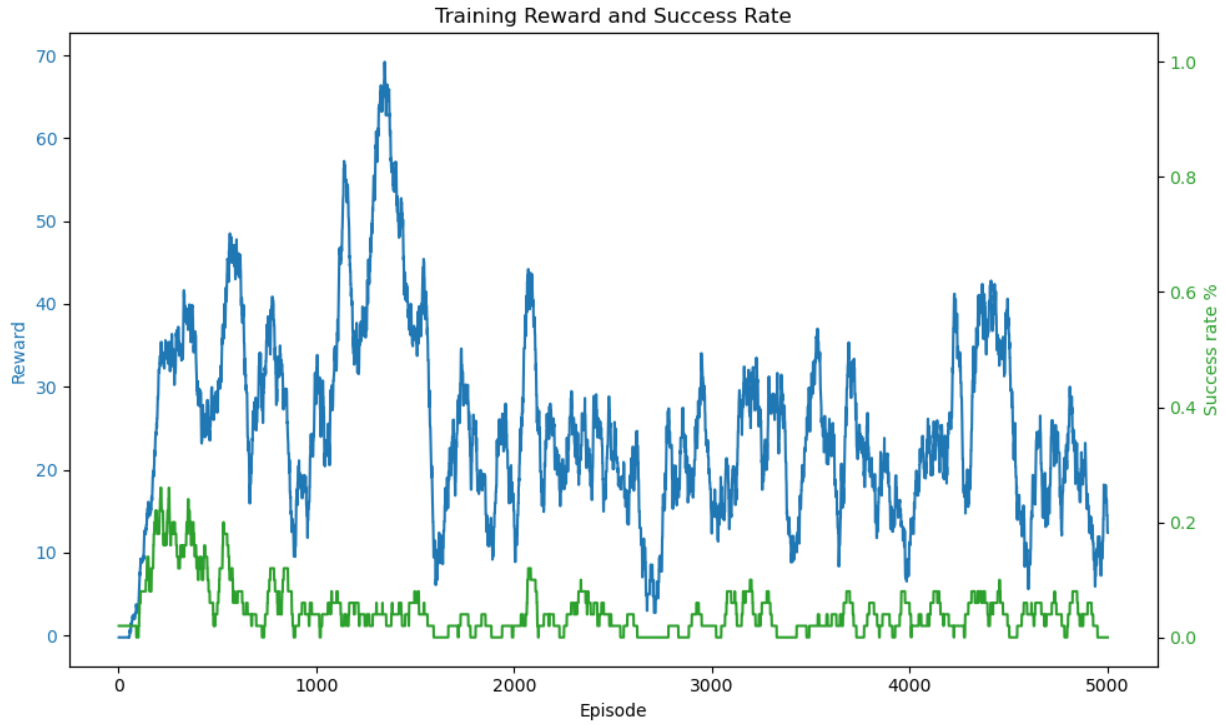
- In the ‘full’ case one agent became the ‘leader’ and the other agent became the ‘follower’. The agent that was closer to the goal usually tended to become the leader.
- In the ‘independent’ case the movements of the two models were totally uncoordinated form each other.
- In the ‘comm’ case the ‘leader’-‘follower’ dynamic reemerged.

The model with only communication performed the best with the highest training success rate and the best evaluation scores. This maybe due to the fact that the distance metric was interfering with the ‘leader’-‘follower’ dynamic between the two agents.

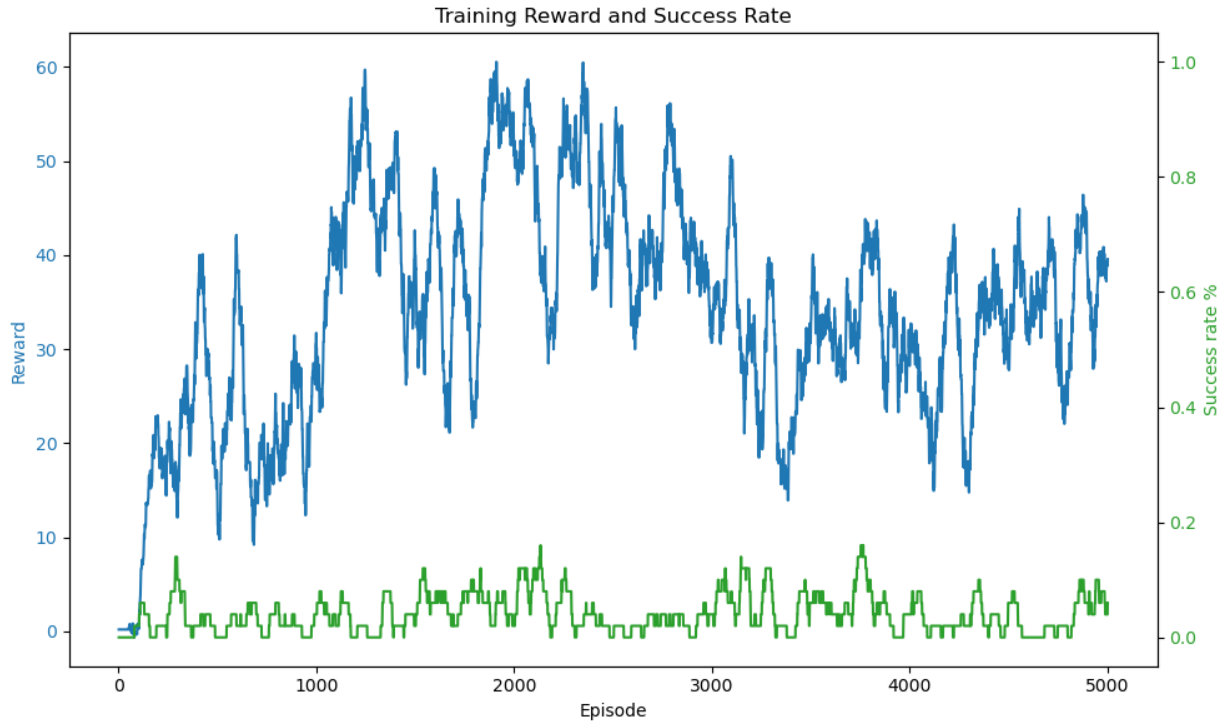
Training curve ‘FULL’:



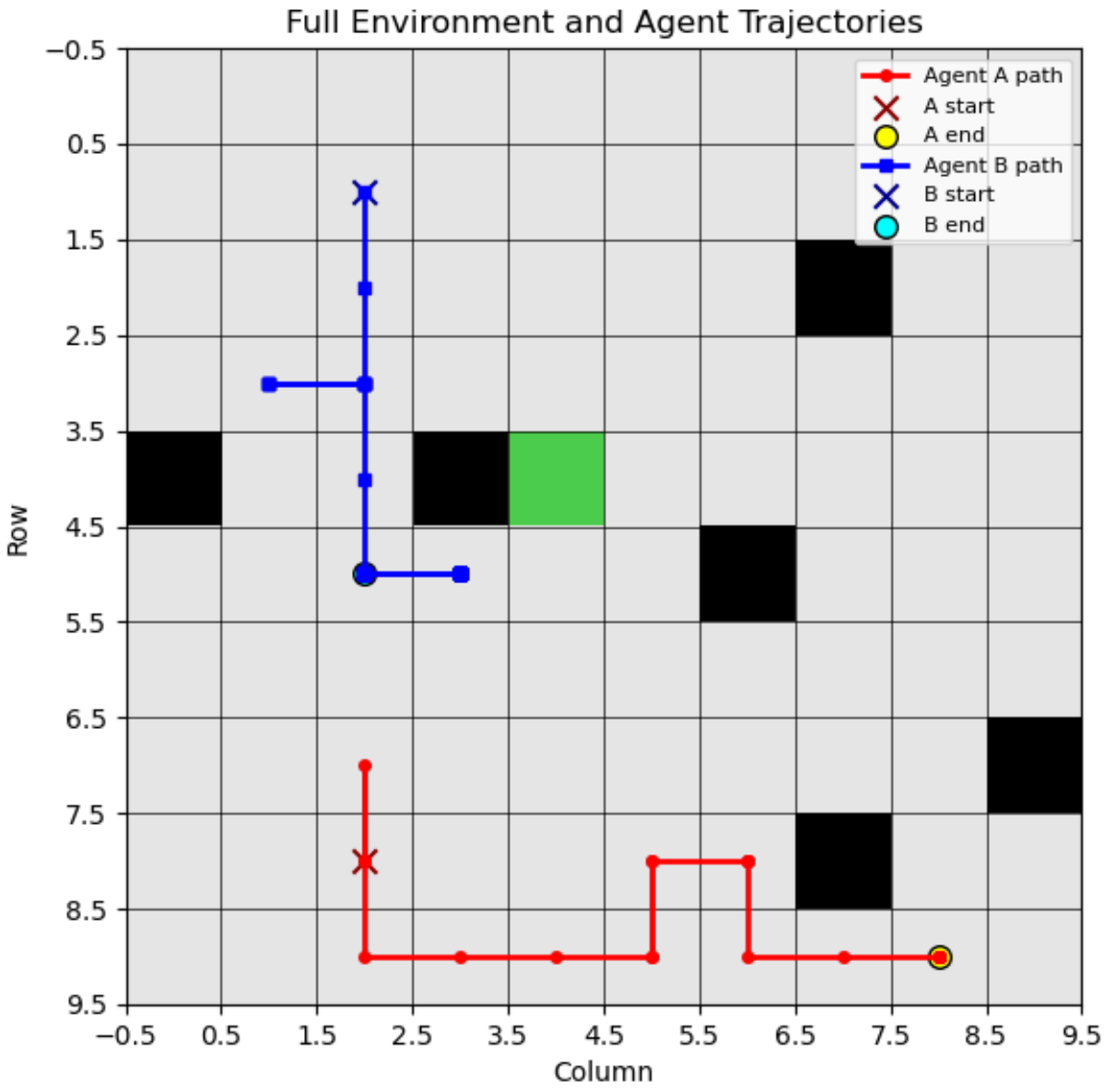
Training curve ‘INDEPENDENT’:



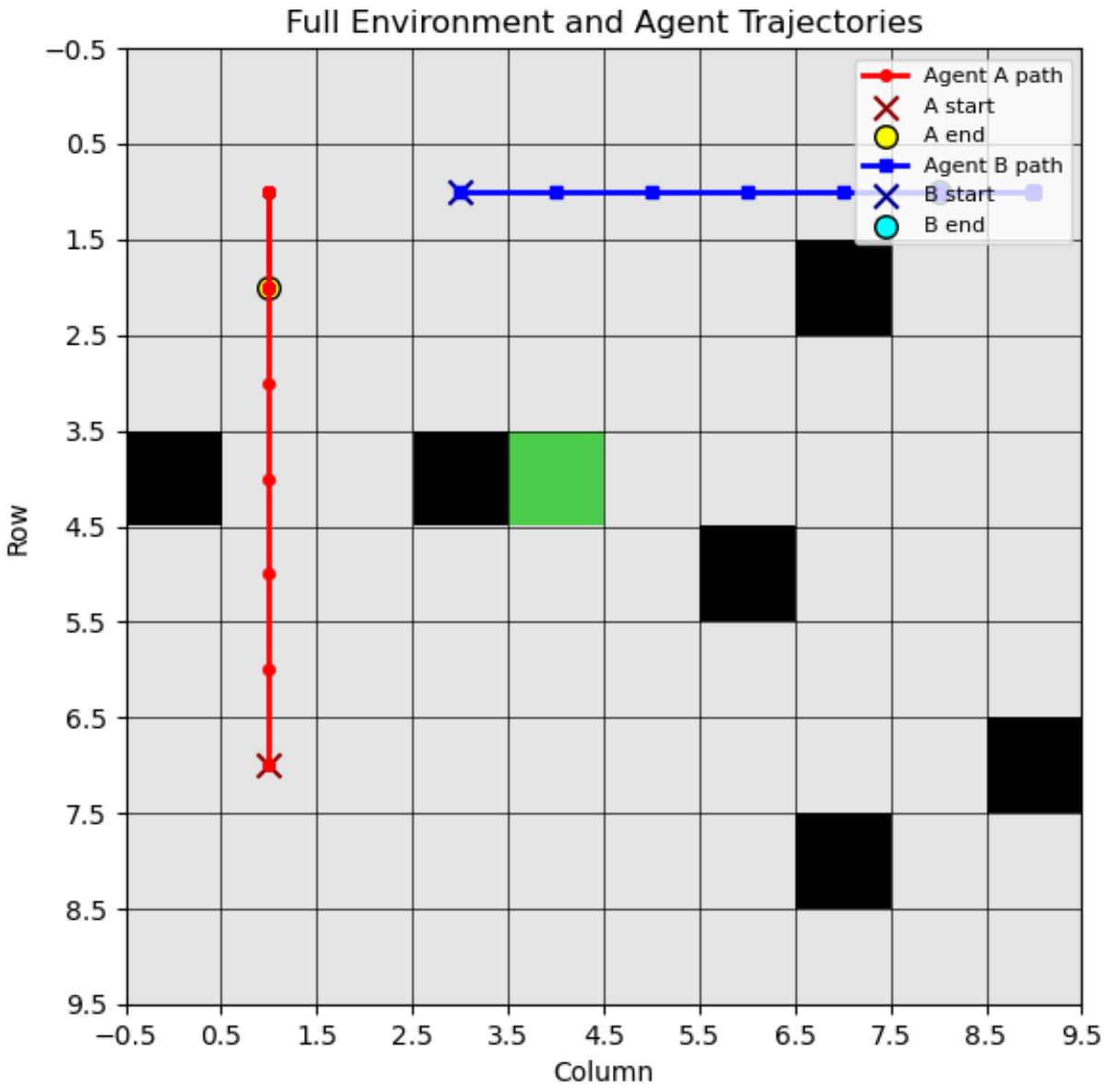

Training curve ‘COMM’:



Trajectory ‘FULL’:



Trajectory ‘INDEPENDENT’:



Trajectory ‘COMM’: