1 Problem 1

For both of my agents, I adopted a decaying exploration parameter(ϵ) which starts at 1.0, meaning the agent will always choose a random move from the environments movement space and decays to 0.1 using this formula:

$$\epsilon_{t+1} = \max\big(\epsilon_{\min},\ \epsilon_t - \Delta\epsilon\big), \quad \Delta\epsilon = \frac{\epsilon_0}{0.5 \cdot N}$$

where N is the amount of times the agent will go through the training loop. The number of episodes I chose is 100000000 episodes.

1.1 SARSA Agent

For my SARSA agent I found that a learning rate(α) of 0.01 and a discount rate(γ) of 0.95 to work the best for me. The agent with these parameters converges to an average reward of 2.4 after about 12 minutes of training. To run the script with these parameters you can run python taxi_sarsa.py 1.0 0.01 0.95.

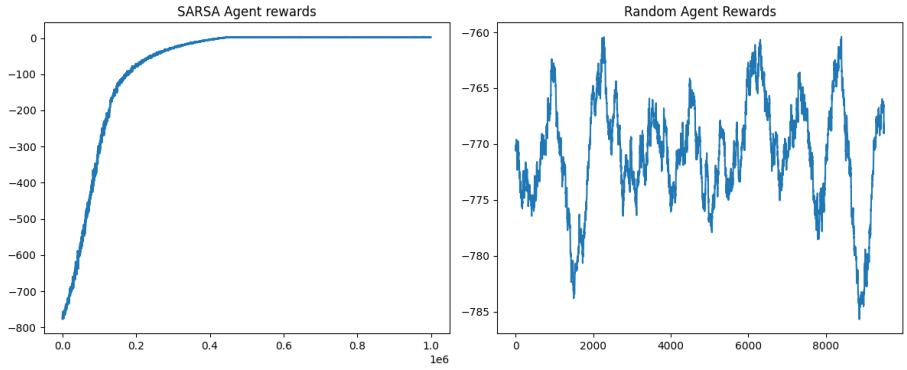

Figure 1: Graph of average reward per playout with a SARSA agent vs a Random agent. The SARSA agent converges to an average reward of 2.4

1.2 Q-Learning Agent

For my Q-Learning Agent I found a very small learning rate(α) of 0.001 and a discount rate(γ) of 0.95. The agent with these parameters converges to an average reward of 2.5 after about 11 minutes of training. To run the script with these parameters you can run python taxi_qlearning.py 1.0 0.001 0.95

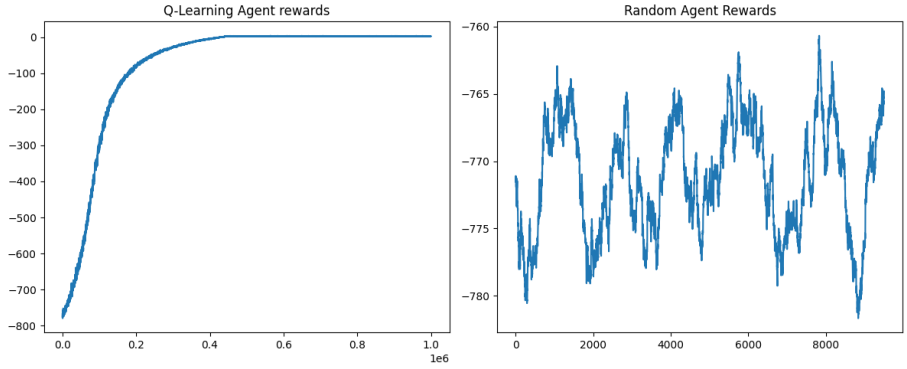

Figure 2: Graph of average reward per playout with a Q-Learning agent vs a Random agent. The Q-Learning agent converges to an average reward of 2.5

2 Problem 2

2.1 Agents

1. **Tit-fot-Tat**: For my Tit-for-Tat agent, I chose to have it randomly take an action if it moves first. The agent keeps track of the opponent's moves and then plays the move the opponent last played.

2. **Bully**: For my Bully agent I hard-coded the moves for each game. Where the Bully will always choose "Testify", "Straight", "Action", based on the game it is playing.

3. **Godfather**: My interpretation of the Godfather agent was to target certain pairs of actions for each game that the opponent cannot refuse. The pairs were (Refuse, Refuse), (Straight, Straight), and (Action, Action). For each game, I have a grace period, where if the opponent plays an action that goes against my target pairs, the godfather agent then chooses one that maximizes the opponent's score. After the grace period is over, the Godfather agent then chooses an action that minimizes the opponent's score, even at the cost of its own score, until they choose an action that follows along the target pairs. The grace period is then reset, and the Godfather agent continues playing actions for the target pair. I believe the godfather agent should have these grace periods as we are acting as a magnanimous agent, allowing our opponent to mess up a few times, but any more than that, and the godfather agent will have to punish the mistake.

4. **Fictitious Play** For my fictitious play agent, it keeps track of belief states of what the opponent will play and then plays the action that will give it the most reward. Each time the opponent plays a move the agent updates its belief to keep track of what is the most likely move the opponent will

play. If the fictitious play agent is playing the first move, it chooses a move out of the game's movement space at random.

2.2 Results

Prisoners Dilemma Results				
Agent Name	Tit-for-Tat	Fictitious Play	Bully	Godfather
Tit-for-Tat	(3.00, 3.00)	(0.99, 1.04)	(0.99, 1.04)	(2.99, 2.99)
Fictitious Play	-	(1.00, 1.00)	(1.00, 1.00)	(1.08, 0.98)
Bully	-	-	(1.00, 1.00)	(1.01, 1.00)
Godfather	-	-	-	(2.98, 2.98)

Chicken Results				
Agent Name	Tit-for-Tat	Fictitious Play	Bully	Godfather
Tit-for-Tat	(2.25, 2.22)	(2.19, 2.19)	(1.00, 1.00)	(2.25, 2.25)
Fictitious Play	-	(2.00, 2.00)	(1.50, 3.50)	(1.50, 3.50)
Bully	-	-	(1.00, 1.00)	(3.48, 1.50)
Godfather	-	-	-	(2.00, 2.00)

Movie Selection Results				
Agent Name	Tit-for-Tat	Fictitious Play	Bully	Godfather
Tit-for-Tat	(1.98, 2.97)	(2.85, 1.90)	(2.97, 1.98)	(3.00, 2.00)
Fictitious Play	-	(3.00, 2.00)	(3.00, 2.00)	(3.00, 2.00)
Bully	-	-	(3.00, 2.00)	(3.00, 2.00)
Godfather	-	-	-	(3.00, 2.00)