Intelligent Agents and Multiagent Systems

MSc in Artificial Intelligence

Equilibria computation
in Zero-Sum Games

Authors
Aris Tsilifonis,mtn2323

Supervisor
Georgios Vouros

26 February,2024

Contents

1. Abstract

In this assignment, two methods of computing equilibria are examined. Different combinations of the algorithms are tested to understand thoroughly how the agents interact in a competitive setting. Diagrams illustrate their behavior as well as their tendency to converge to equilibria. Firstly, Fictitious Play, a model-based approach, is introduced. Then, Reinforcement learning was applied to test more sophisticated agent behavior. The competition between the agents provides important results about which one behaves better and under which circumstances.

2. Zero Sum Repeated Games

Repeated games are those that are played over multiple rounds. The are distinguished between infinite and finite repeated games depending on if the game is played an infinite number of times or not. The latter considers more the theoretical side of the Game Theory field. Zero-sum games belong in a class of games that represent states of full competition where one player's loss is a profit for the other. In a zero-sum game for each strategy profile $a \epsilon A_1 \times A_2 : u_1(a) + u_2(a) = 0$. Zero-sum games can be presented in a way of a matrix, which is called Normal Form Games.

The concept of best response is very important when these games are examined.

If an agent knows what the other agents are playing the problem would become simple as they will always choose the action that maximizes their utility. In a two-player game we define s1 as the strategy profile for player 1. Then the strategy profile s1* is the best response for player 1 to player's 2 strategy if $u_1(s_1^*, s_2) \geq u_1(s_1, s_2)$. The best response characterizes a player behavior but it is not necessarily unique. Best response is essentially a strategy that generate the greatest payoff for him or her given what the other players are doing.

3. Nash equilibrium

Generally, the agent is agnostic about what the other's strategies are (mainly in non-cooperative, competitive setting). Best response allows us to introduce another

important concept in Game Theory, the Nash Equilibrium discovered by the Nobel prize winner John Nash. A strategy profile $s = (s_1, \dots, s_n)$ is a Nash Equilibrium if, for all agents i, s_i is a best response to s_{-i}. To put it simpler, if at one point all players announce their strategies simultaneously, no player would desire to deviate from their strategy.

There are two types of strategies, pure and mixed. If a player chooses playing strategies randomly, then this player is using a "mixed strategy". In a pure strategy a player chooses an action for sure, whereas in a mixed strategy, he chooses a probability distribution over the set of actions available to him.

A Nash equilibrium is a pair of strategies such that neither player can gain by a unilateral change of strategies

The expected utility of a mixed strategy u_i for a strategy profile s in a normal form game N is defined as $u_x(s) = \sum_{a \epsilon A} u_x(a) * \Pi_{y=1}^{n} s_y(a_y)$.

4. Teaching and Learning

Fictitious Play

Fictitious play refers to a dynamic process where at each stage, players play a (pure) best response to the empirical distribution of their opponent's play. This algorithm is based on this equation:

$$AverageSelfStrategy = (1 - a) * AverageSelfStrategy + a * best\ response$$

In this way agent balanced between exploration and exploitation stages. At the start, learning rate is relatively high and algorithm explore new strategies. Then, agent exploits what already knows to make an informed decision about the game.

The algorithm can be summarized by the steps below:

Fictitious Play algorithm

Initialize beliefs about the opponent's strategy

Repeat n episodes:

Play best response to the addressed strategy of the opponent

Observe the opponents actual play and update beliefs accordingly

This algorithm avoids getting trapped in a local equilibrium since it uses adaptive exploration. One major drawback of this strategy is that it assumes a stationary strategy of the opponent. We call a strategy stationary, when the agent adopts the same strategy over the course of the game. If the opponent changes strategies between game, then the average will not reflect opponent choices.

Reinforcement Learning

Reinforcement Learning (RL) is a type of machine learning paradigm that is concerned with how agents ought to take actions in an environment to maximize some notion of cumulative reward. RL can be classifying as:

Model-based RL: The agent builds a model of the environment's dynamics and uses it to make decisions.

Model-free RL: The agent learns policies or value functions based on the observed rewards without any model of the environment's dynamics.

Fictitious play mostly classifies as model-based RL while Q-Learning is a model free approach.

Q- Learning

In this project, Minimax Q-Learning was implemented. It essentially involves a maxmin (or maximin) strategy in game theory, which is a decision rule used by a player to maximize the minimum gain that can be achieved under any scenario. It reflects a strategy where a player prepares for the worst-case scenario by choosing the action that provides the best of the worst possible payoffs.

The π(policy) needs to be computed with linear programming

The agent chooses an action:

Based on an exploring probability, return an action uniformly at random

Otherwise, return action a_1 with probability π_1

The essence of the algorithm is the following three equations. After the agents receive the reward from the action a_1 and opponent action a_2:

$$Q(s, a_1, a_2) = (1 - a)Q(s, a_1, a_2) + a\big(r + \gamma V(s')\big)$$

$$\pi_1(s, \ .) = \underset{\pi_1 \epsilon PD(A_1)}{\operatorname{argmax}} \min_{a_2 \epsilon A_2} \sum_{a_{1\epsilon A_1}} \pi_1(s, a_1) Q_1(s, a_1, a_2)$$

$$V_1(s) = \min_{a_2 \epsilon A_2} \sum_{a_{1\epsilon A_1}} \pi_1(s, a_1) Q(s, a_1, a_2)$$

The $V(s)$ is essentially the maximum expected future reward. It is multiplied by a discount factor gamma in the first equation, since it is not certain that the agent will get it. The learning rate a is decreasing after each iteration raising the importance of the quality of the action (exploitation stage). The equation is effectively updating the value function V to the worst-case expected reward given the current strategy P and Q-values Q. A conservative approach was used since the minimum of reward (expected utility of the action) was considered. This is consistent with the approach of minimax q-learning where you prepare for the worst-case assuming the opponent will play the best possible counter strategy against you.

There is a theorem which states that Q-learning guarantees Q and V values will converge to those of the optimal policy π provided that each state-action pair is sampled an infinite number of times.

5. Experiments

In this project, the aim was to determine equilibria in zero-sum games, using Fictitious play and Q-Learning algorithms. The study involved implementing both learning algorithms to different games. It will be tested whether these algorithms converge to the same equilibria, analyze their convergence rate and explore the influence of initial configurations.

More specifically, two agents use each of these algorithms and compete against each other. The research elaborated on Fictitious play versus Fictitious Play, Q-Learning versus Q-Learning and Fictitious play versus Q-Learning.

5.1 Matching Pennies

Firstly, the matching pennies game was introduced to the agents. In the figure below, the payoff matrix of the game is shown.

	Heads	Tails
Heads	(1, -1)	(-1,1)
Tails	(-1,1)	(1, -1)

There is no pure strategy Nash equilibrium in this game. A pure strategy Nash Equilibrium happens when no player can benefit by changing their strategy unilaterally while the other players keep theirs unchanged. In a pure strategy a player chooses an action for sure, whereas in a mixed strategy, he chooses a probability distribution over the set of actions available to him. If there are no pure strategy Nash equilibria, then mixed strategy Nash equilibria definitely exist. Let row player choose heads with probability p and tails with probability 1-p. The row player should be indifferent between his two possible actions since if there were a difference in expected payoffs among the actions, the player would have an incentive to switch to a pure strategy, exclusively choosing the action that provides the higher expected payoff. So, for row player:

$$u_1(H) = u_1(T)$$

$$1 * p + (-1) * (1 - p) = (-1) * p + 1 * (1 - p)$$

$$p - 1 + p = -p + 1 - p$$

$$2p - 1 = 1 - 2p$$

$$4p = 2$$

$$p = 1/2$$

This illustrate that the row player plays the mixed strategy of (H, T) = [0.5, 0.5].

Both players have an expective payoff equal to zero. This can be computed by:

$$E(agent_x\ payoff) = \sum_{a \in A} s_1(a) * \bar{u}_1(a)$$

Where $\bar{u}_x(a_y) = s_{-i}(a_y) * u(a_y, a_1) + s_{-i}(a_2) * u(a_y, a_2)$

$$s_{-i}\ denotes\ the\ strategy\ of\ the\ opponent$$

The average reward refers to the expected payoff that a player receives over time. In the repeated games, which are examined here, they are computed over multiple periods. In those games, players engage in the same game multiple times, and the average reward captures the long-term payoff that a player can expect to achieve from their strategy. It can be regarded as the current average of the expected payoffs.

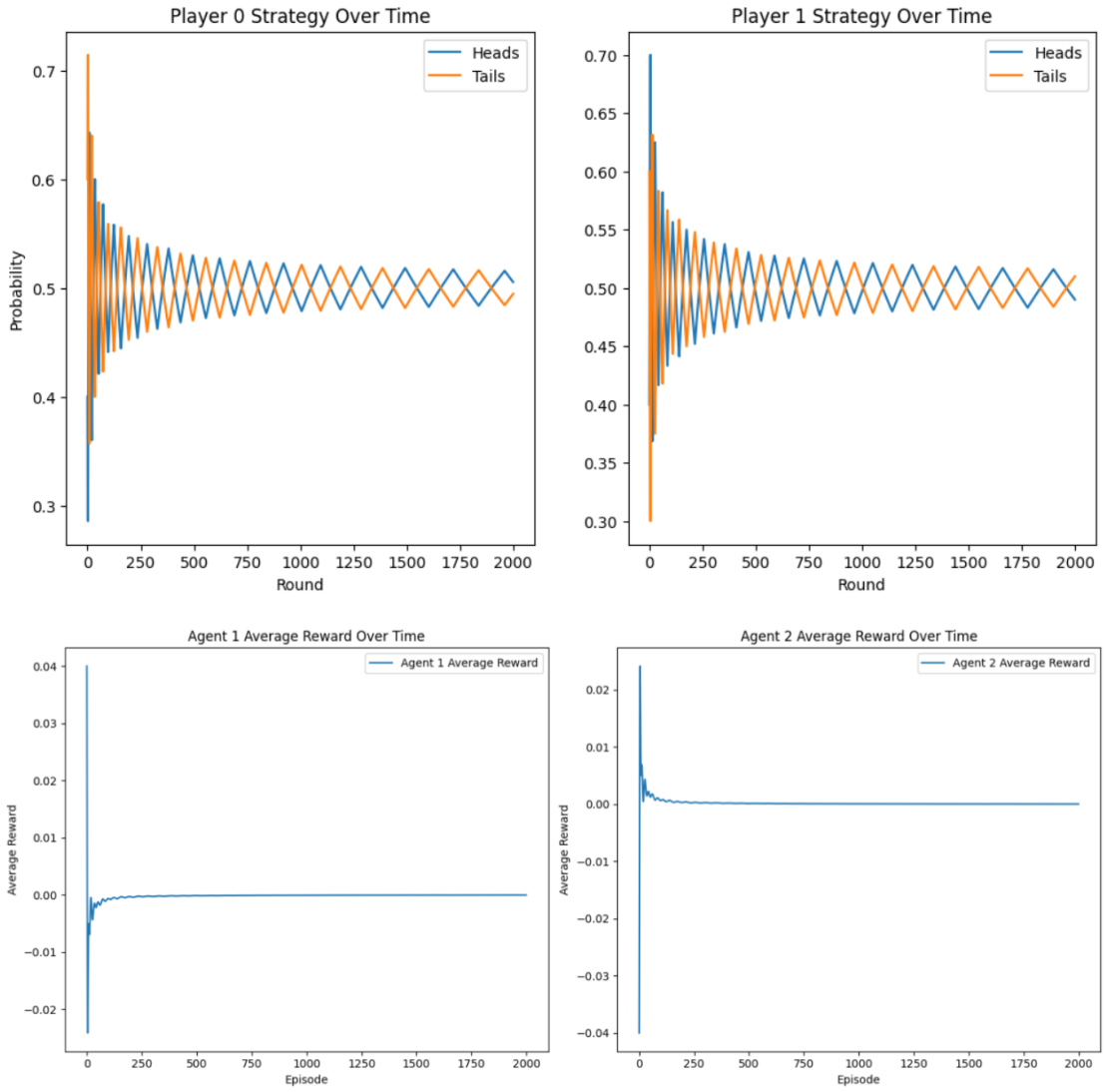
$$average\ reward = \sum_{i=1}^{k} \frac{r_i}{k}$$

, where k indicates the number of rounds

FP and RL algorithms were utilized to compute the mixed strategies Nash equilibria. Both algorithms were tested for 700-2000 iterations. Regarding learning rate, it is computed at each iteration with this equation: $\text{learning\_rate} = 1 / (\text{iteration} + 5)$ for fictitious play. Regarding Q Learning, it is initialized to 1.0. Regarding the exploration rate, epsilon, it is initialized as 0.3. The discounted factor Gamma=0.9. For clarification, action 0 and action1 in Q-Learning diagrams mean Heads and Tails respectively. The

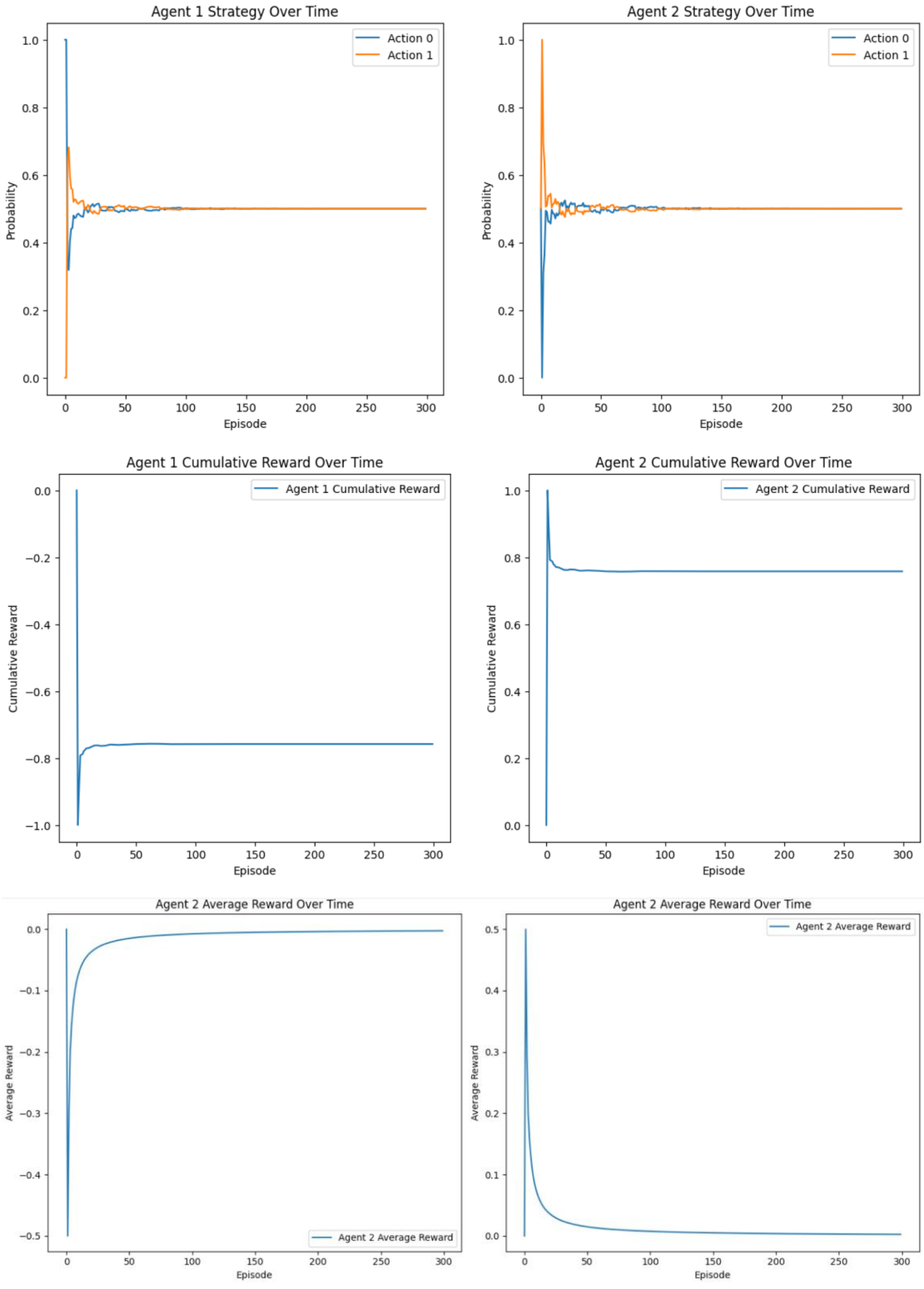
y-axis illustrates the probabilities of each of the two possible actions while the x-axis presents the number of iterations of the experiment.

Figure: Fictitious play vs Fictitious play Matching Pennies



Agent's beliefs were initialized [0.5,0.5] for each one respectively. Another experiment included beliefs of [0.1, 0.9] for the first and [0.6, 0.4] for the second. 10% of the previous times he has played Heads and 90% Tails. The results were very similar on both of the experiments about fictitious play.

Figure: Minimax Learning vs Minimax Learning Matching Pennies



From the diagrams, it can be understood that Q-learning converges a lot quicker than Fictitious play. The Q-Learning needs only 50 iterations while Fictitious converges a lot slower. The average reward diagrams are very similar, illustrating the effectiveness of the experiment since the values approach zero. As we know, the expected payoff of this game is zero, so that indicates that the agents manage to reach the mixed strategy

Nash Equilibrium. The fact that cumulative reward, as well as the average reward per iteration, stabilizes very quickly is indicative that the agents have no incentive to change their probabilistic policy.

The Q-Learning's figures prove that this algorithm is more robust since there are very few oscillations compared to Fictitious Play. Even when the agents are initialized with a preference over policies, they quickly shift their strategies to converge at the equilibrium. As you can see in the figures, agent 1 prefer only action 0 at the start and avoids action 1 completely (policy (1.0,0). The opposite happens for the second agent's policy.

5.2 Selling Damaged goods

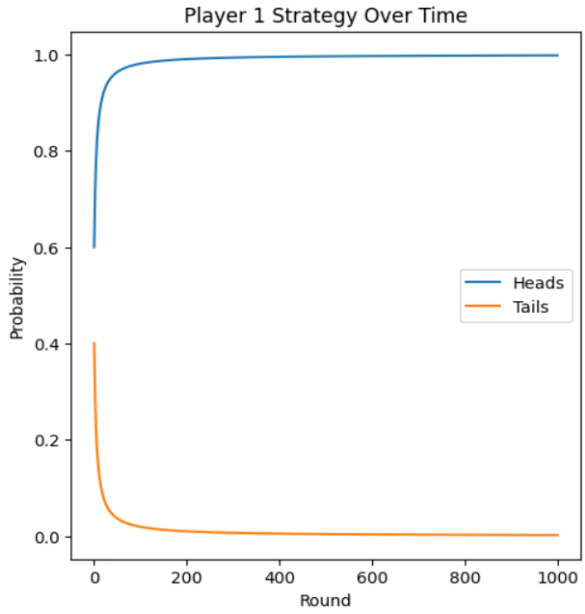
The other game to be studied, selling damaged goods. The payoff matrix of this game is demonstrated below:

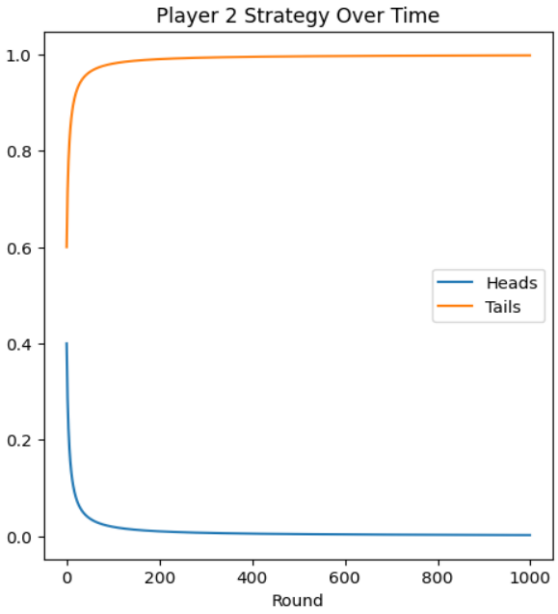
	Buy	Pass
Keep	(1, -1)	(-1,1)
Sell	(-1,1)	(-1, 1)

In this game, two pure strategies Nash Equilibria exist. These are, *(Keep, Pass)* and *(Sell, Pass).* Also, another equilibrium exists in the pure strategy of Pass and the mixed strategy (0.5,0.5) Buy, Pass of the column player. In each of the equilibrium, the row player expected payoff is -1 and 1 for the column player. That means that over the long run row player will win and column player will lose. The cumulative graph reflects this thought. Row's player cumulative reward is decreasing and column's one increasing consistently. This tendency sustains over the course of all episodes.

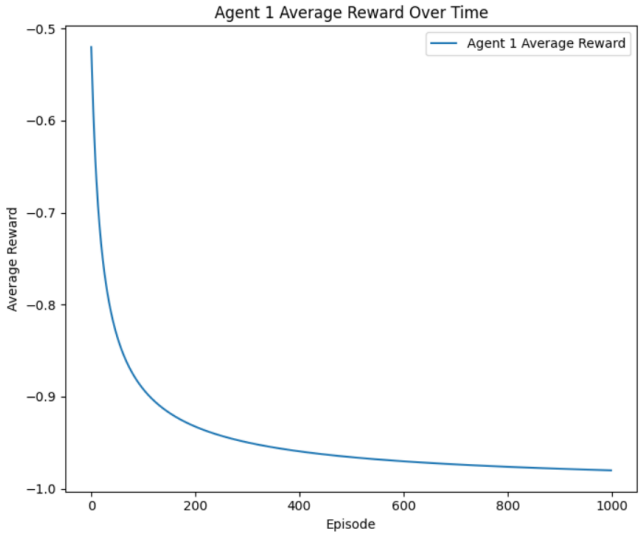
For fictitious play, each of the two agents was initialized with beliefs (0.5, 0.5) for each action. As it can be viewed at the diagrams below, agents quickly manage to reach equilibrium by adjusting their policies. Initial beliefs can be altered, in order to be biased about a specific policy. For instance, we can assign (0,1.0) and (0,1.0) for sell and pass respectively. In this way, there is a greater chance that they will reach that equilibria. However, this specific game has multiple equilibria and it is not guaranteed that the agents will converge at the same all of the time. When other initial beliefs were tested, some differentiations in the action probabilities diagram occurred. For clarification, they were not included in this report but someone can see them by testing the code. The accuracy equilibria computation was verified by the average reward plot. Since the expected payoff of this game -1 for the row player and 1 for the column, it is understood that the agents should get an average reward similar to that value. As it can be viewed afterwards, the agents' average reward consistently converges to that value. The lines of the cumulative reward plot follow a trend based on the value of the reward that the agents attain at each iteration. If they acquire negative reward the line will trend backwards, otherwise upwards.

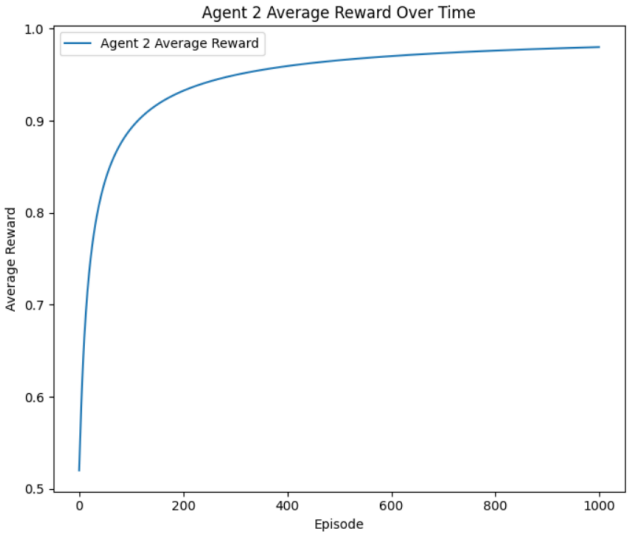
Figure: Fictitious play vs Fictitious play Selling Damaged Goods





Blue line is essential the first action for the row player 1 (Keep), orange line is the second (Sell). For the column player 2, blue line is (Buy) and Orange Line is (Pass). The action profile of the first agent becomes the pure strategy Keep, while the second agent's become pure strategy Sell.





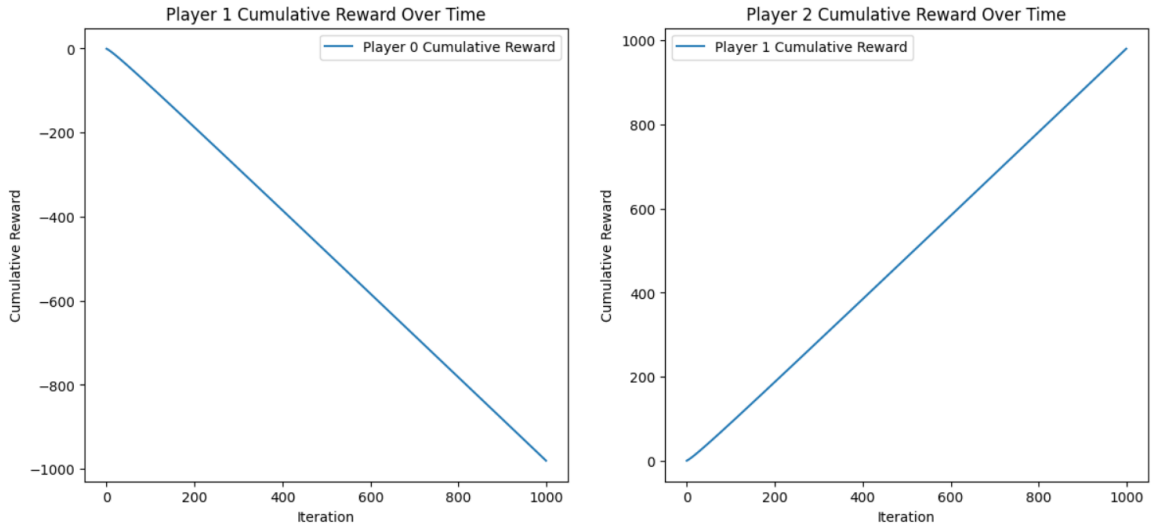
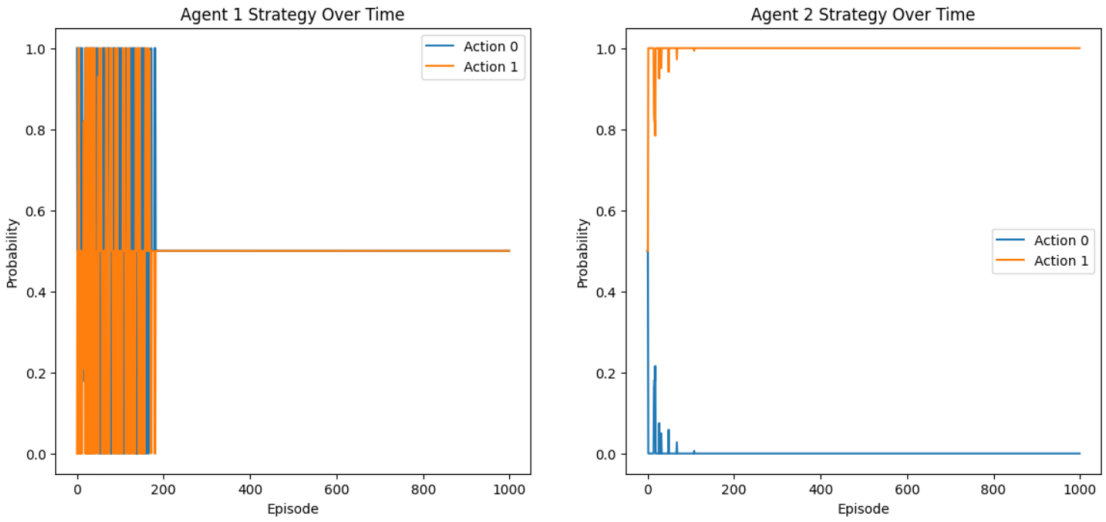
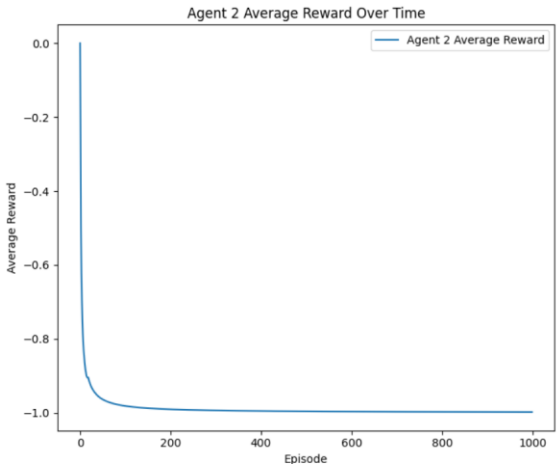
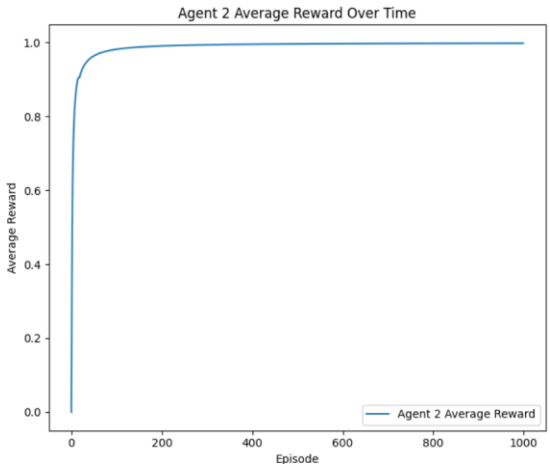


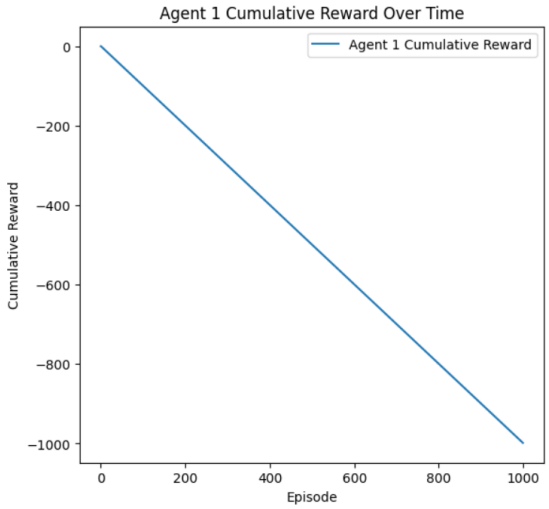
Figure: MinMax Q-Learning vs MinMax Q-Learning Selling Damaged Goods

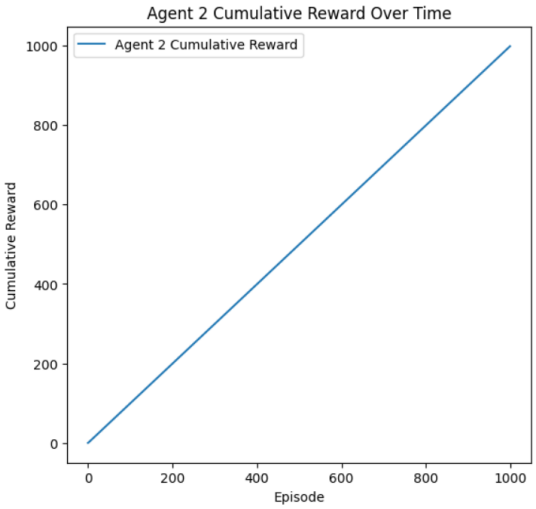
For Minimax Q-Learning, agents always converge to the same mixed Nash Equilibrium regardless of the parameters. The stabilization of the average reward illustrates that the agents have reached a stable equilibrium that maximizes the average reward for all the players over time. The rewards have no tendency to deviate after some steps, since the agents have no incentive to deviate their policies.











5.3 Rock Paper Scissors

The last game that was examined was rock, paper, scissors. It involves three possible actions from which the player has to choose. Its payoff matrix is depicted below

	Rock	Paper	Scissors
Rock	(0,0)	(-1,1)	(1, -1)
Paper	(1, -1)	(0,0)	(-1,1)
Scissors	(-1,1)	(1, -1)	(0,0)

No pure Nash equilibria exist in this game too. Player 1ís goal is to make Player 2 indifferent among his pure strategies.

The following equations hold for expected values of player 2:

$$E_2[Rock] = 0 * p1_{rock} + (-1) * p1_{paper} + 1 * p1_{scissors}$$

$$E_2[Paper] = 1 * p1_{rock} + 0 * p1_{paper} + (-1) * p1_{scissors}$$

$$E_2[Scissors] = (-1) * p1_{rock} + 1 * p1_{paper} + 0 * p1_{scissors}$$

We know have 3 unknowns, $p1_{rock}, p1_{paper}, p1_{scissors}$. It must be that player 2 has

$$E_2[Rock] = E_2[Paper]$$

$$E_2[Paper] = E_2[Scissors]$$

$$0 * p1_{rock} + (-1) * p1_{paper} + 1 * p1_{scissors} = 1 * p1_{rock} + 0 * p1_{paper} + (-1) * p1_{scissors}$$

$$1 * p1_{rock} + 0 * p1_{paper} + (-1) * p1_{scissors} = (-1) * p1_{rock} + 1 * p1_{paper} + 0 * p1_{scissors}$$

$$p1_{rock} + p1_{paper} + p1_{scissors} = 1$$

By solving this system of equations, we get $p1_{rock} = \frac{1}{3}, p1_{paper} = \frac{1}{3}, p1_{scissors} = \frac{1}{3}$

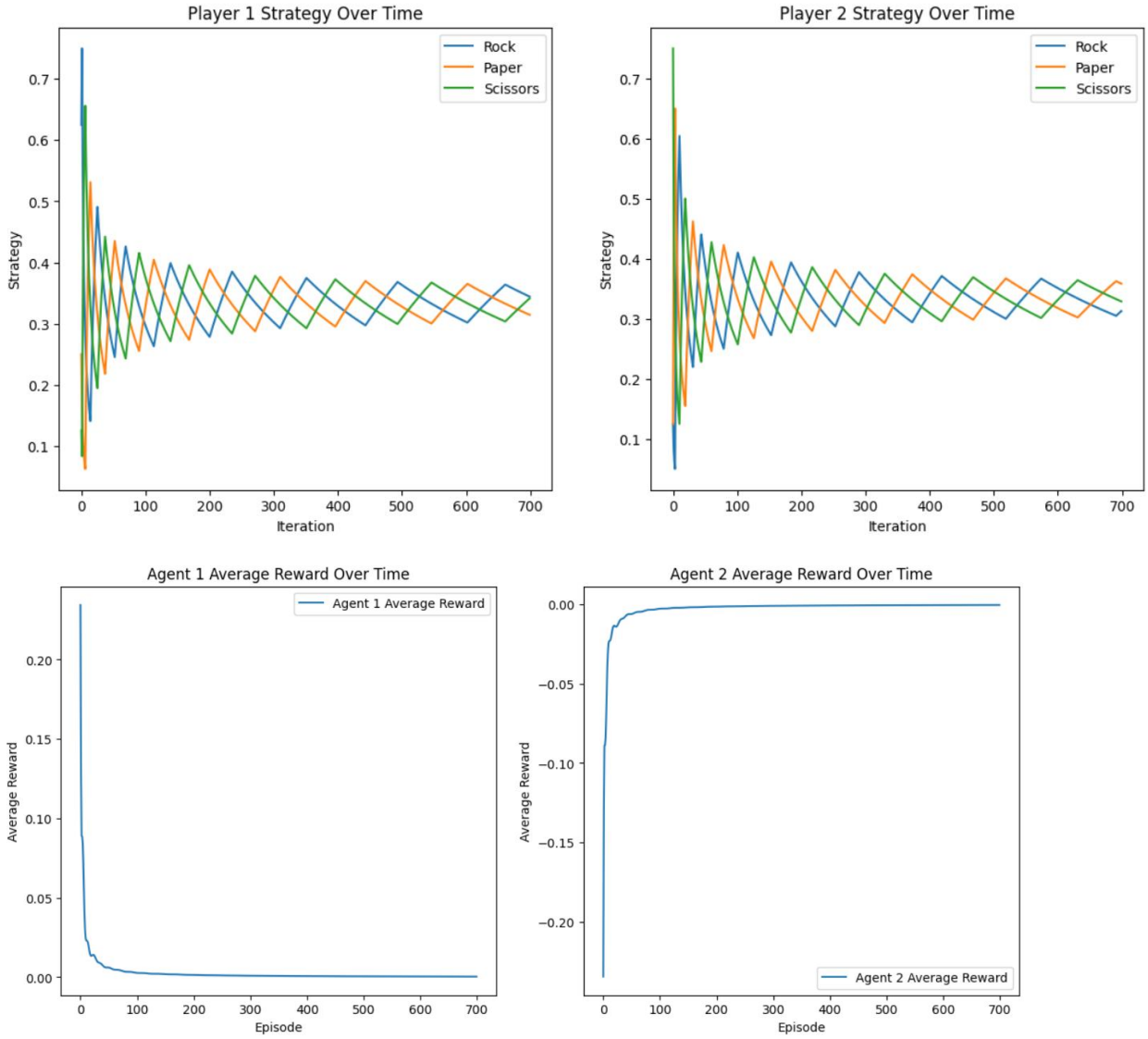
Since payoffs are symmetric, so that the probabilities for Player 2 that make Player 1 indifferent between his strategies are the same for player 2 too.

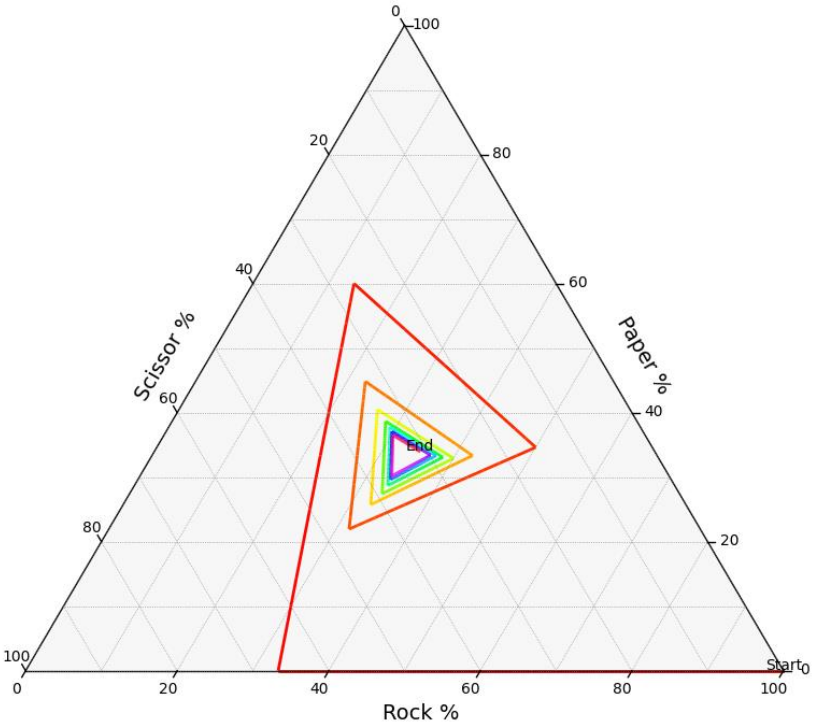
From the graphs below, someone can understand that both agents setting manages to find the mixed strategy Nash equilibria, which is [1/3,1/3,1/3]. The expected payoff of this game is 0 which means that neither player loses or winning in the long run. The cumulative reward remains stead after some initial iterations which verifies what we expected.

Fictitious play converges slower to the equilibria as opposed to Q-learning. On the other hand, Q learning converges a lot more abruptly since there is a steep decline in the line after around only 30 iterations. Average reward stabilizes quickly indicating that agent approach and reach equilibria. Reward graphs are very similar for both algorithms. They are both effective in determining the equilibria.

Simplex graphs were plotted to show the evolution of each agent's policy more clearly. Each corner of the triangle represents a pure strategy. It is a method of determining if the algorithm reach Nash Equilibria, which is essentially the center of the triangle. The rest of the triangle depicts the mixed strategies and the fluctuation of the line illustrates how the agent adjust its policy in this competitive setting. Fictitious play follows a clear pattern which can constitute an algorithm more predictable against more intelligent agents. On the other hand, Q-learning shows no clear pattern in the learning process and can be unpredictable against more sophisticated agents. As a result, Q-Learning can be exploited harder by another agent compared to the fictitious agent.

Figure: Fictitious Play vs Fictitious play Rock Paper Scissors




Scissor %
Paper %
Rock %
Start
End

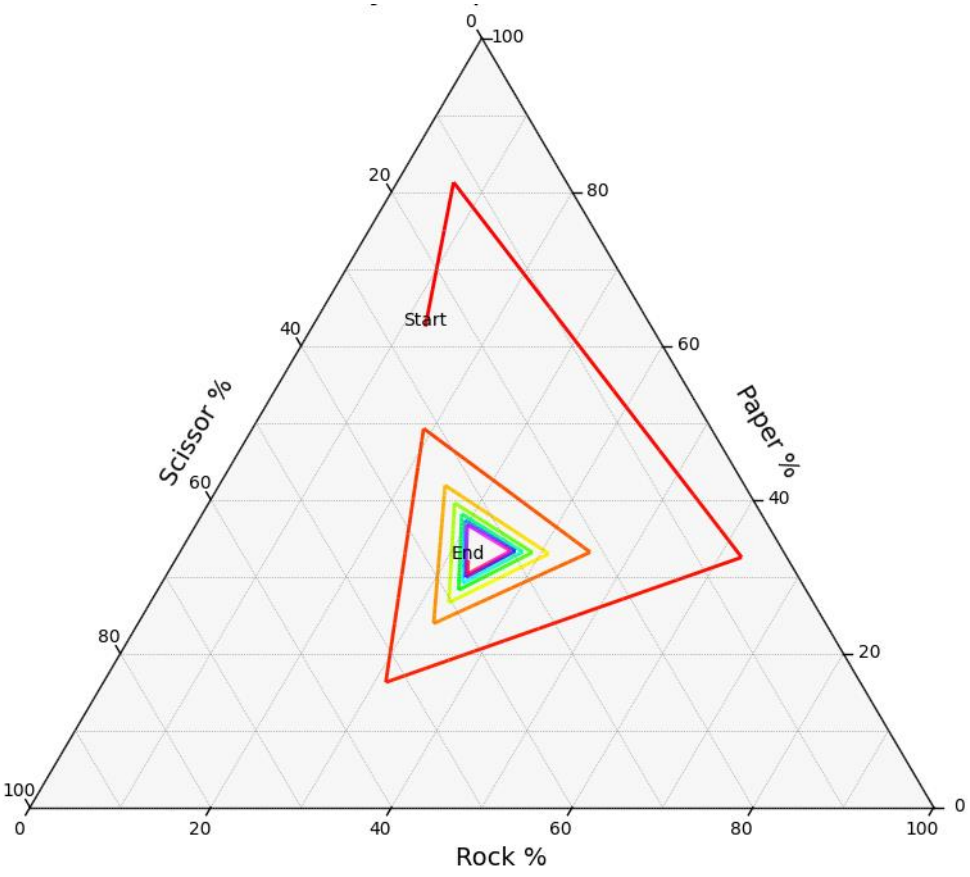
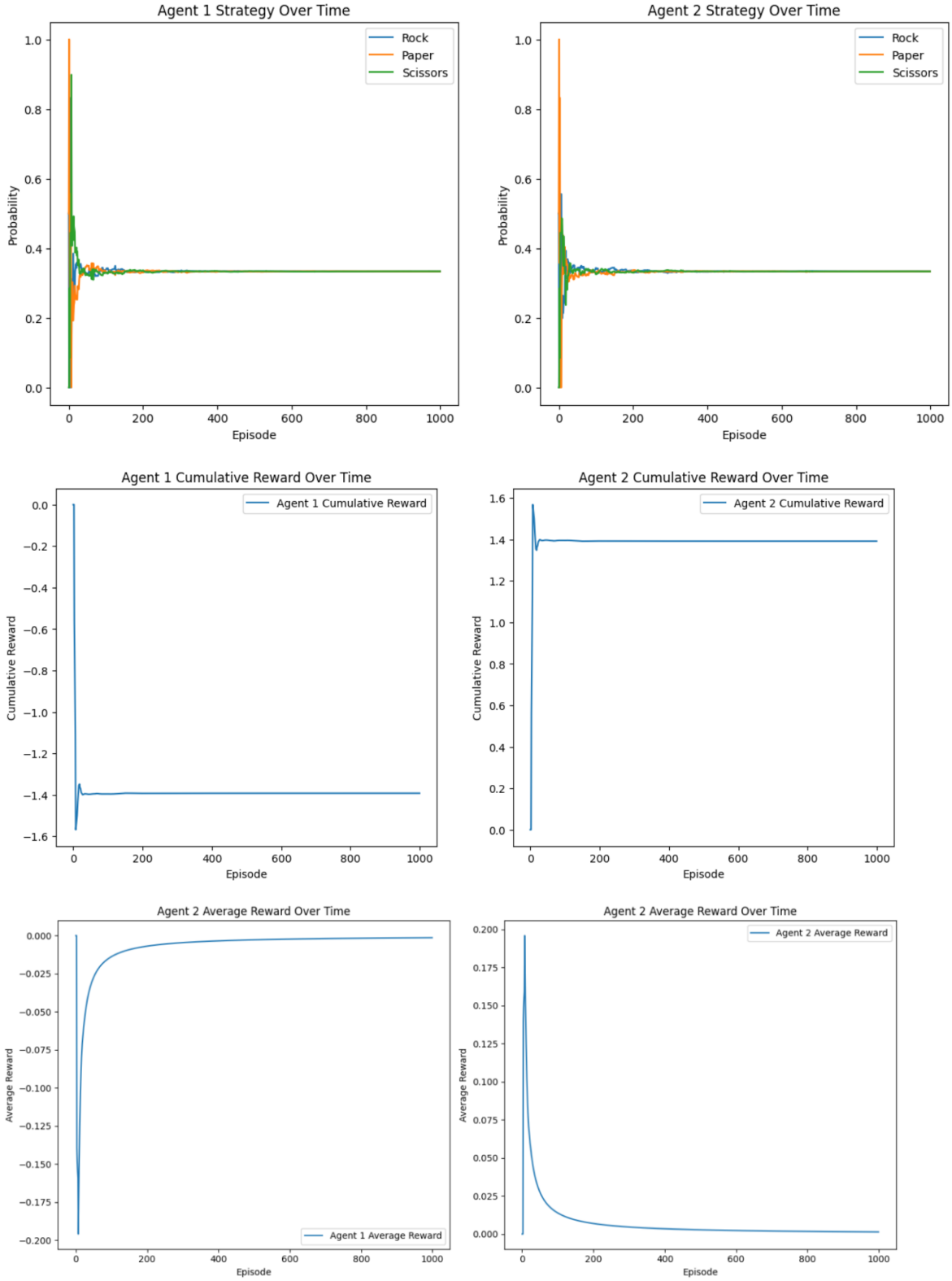
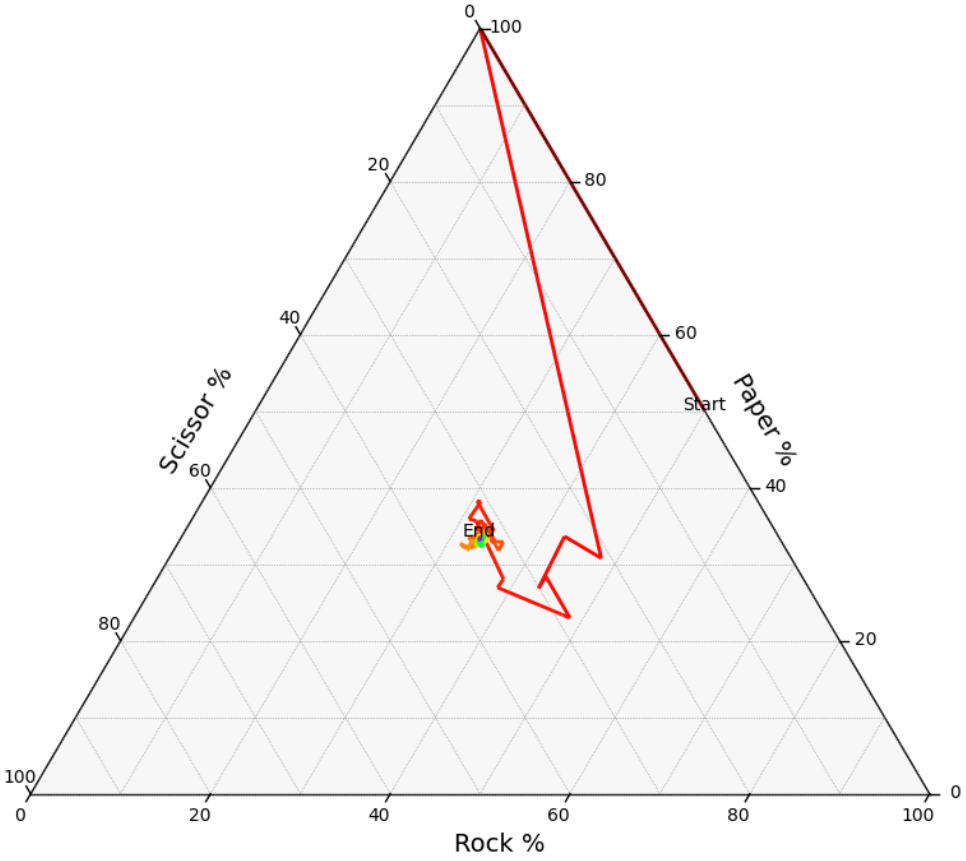
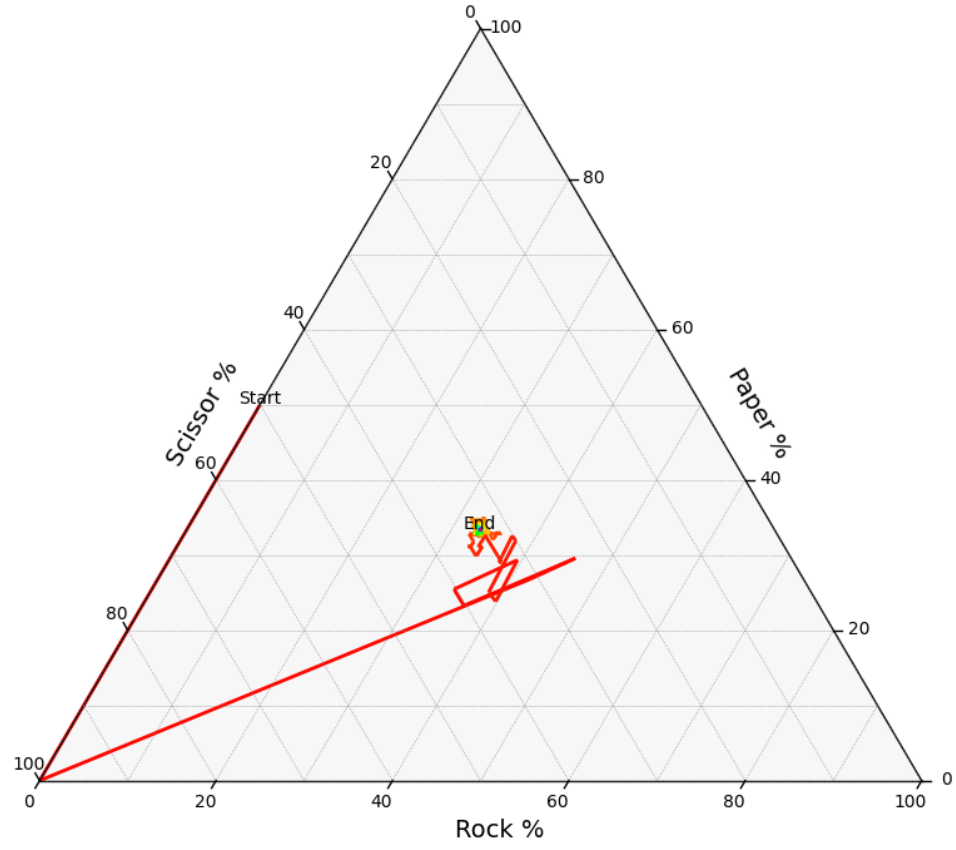

Scissor %
Paper %
Rock %
Start
End

Figure: Minimax Q-Learning vs Minimax Q-Learning Rock Paper Scissors

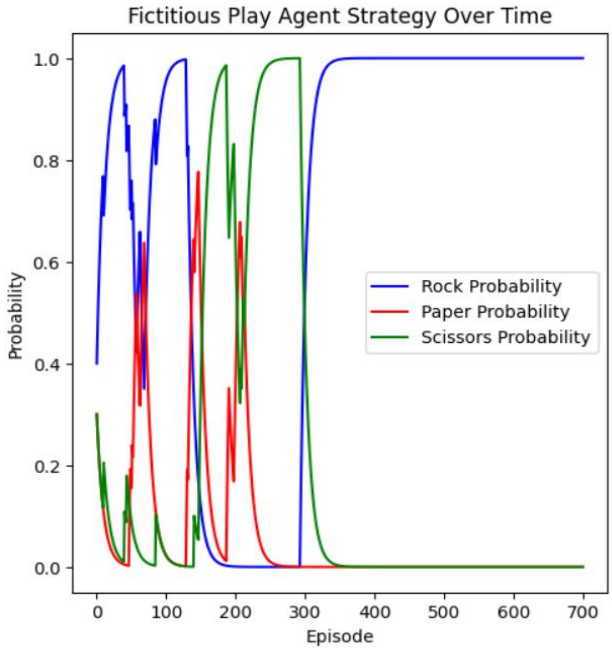


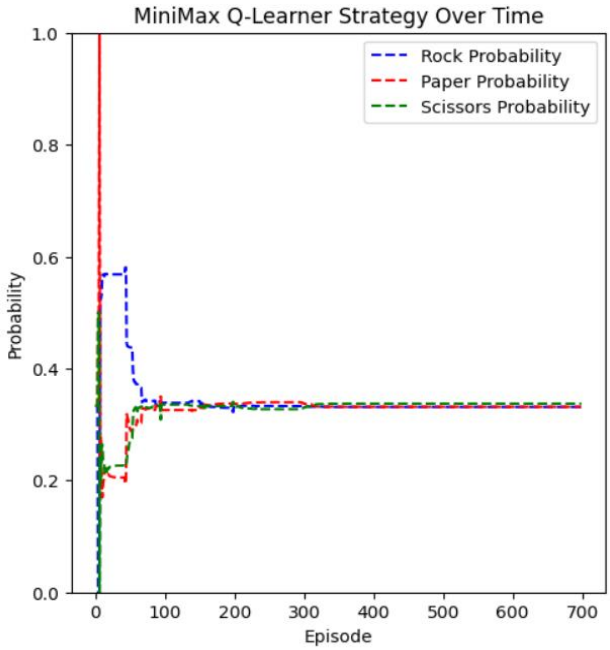

Scissor %
Paper %
Rock %
Start
End
0
20
40
60
80
100

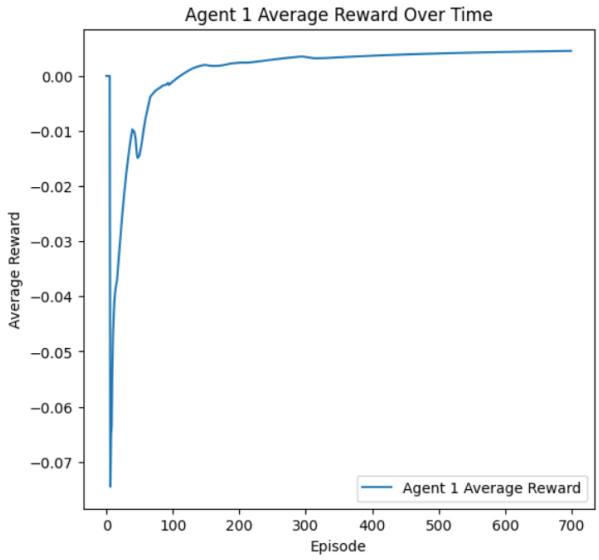

Scissor %
Paper %
Rock %
Start
End
0
20
40
60
80
100

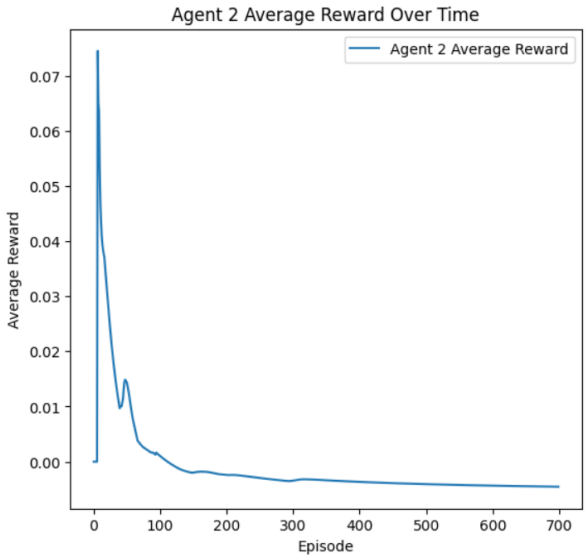
Fictitious Play agent versus RL agent Rock Paper Scissors

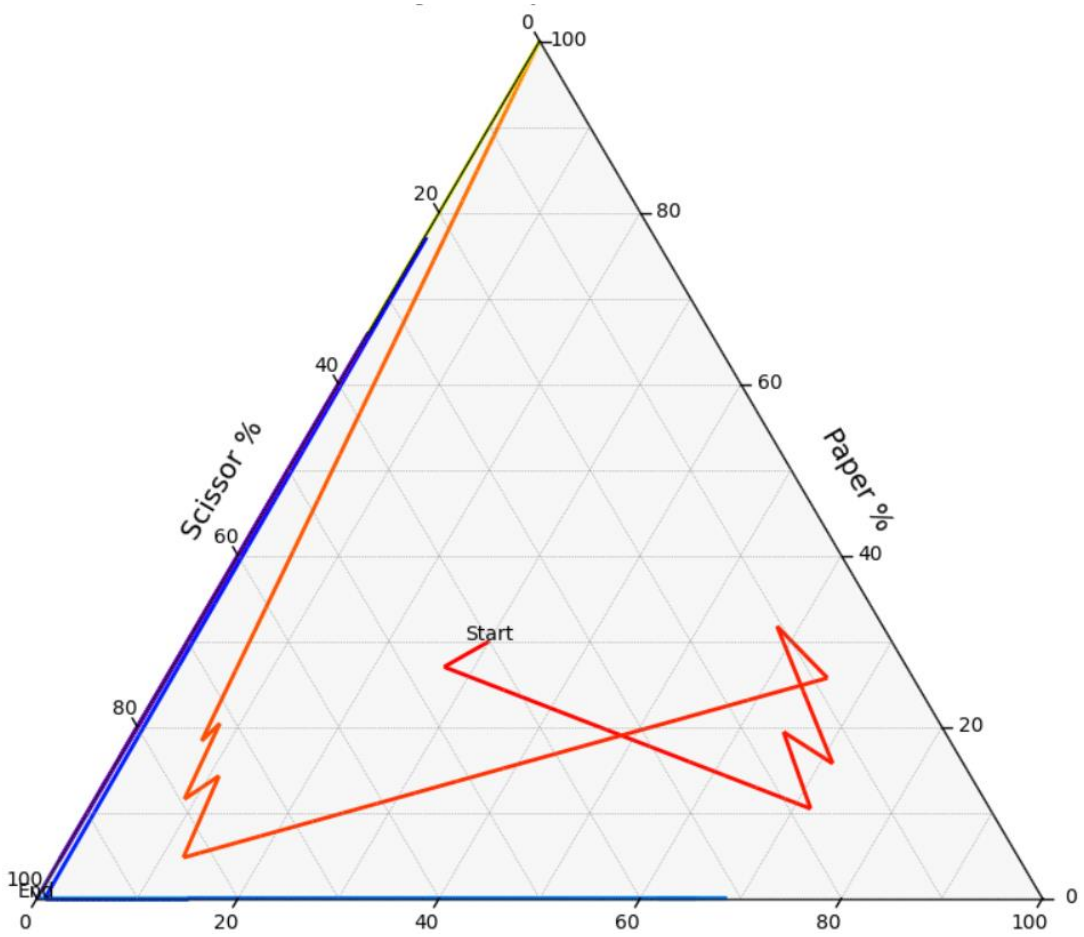
In this setting, Fictitious play and RL were put together to compete at the game rock paper scissors. It is obvious from the graph that MiniMax Q-Learner converges to the game's Nash equilibrium after just only 100 iterations. On the other hand, fictitious play agent tries to convergence to a policy but then it quickly shifts to another strategy. Apparently, Fictitious play agent struggles against Q-Learning Opponent, since the latter has quite unpredictable strategy as discussed previously. Fictitious play stabilizes its policy after 400 episodes but it will not remain the same after a while. The experiment was performed for around 5 times more iterations and fictitious agent did not manage to reach the game's standard Nash equilibrium most of the times. The simplex plot indicates that Minimax Q Learner finds the mixed Nash Equilibrium (1/3,1/3,1/3) of the game but Fictitious agent does not. Average reward graphs demonstrate that the agent stabilized(Q-Learner) or try to stabilize their policy (fictitious agent) because the line becomes flat after around 100 iterations. The simplex plot verifies that fictitious agent is unable to find the standard Nash Equilibria since it is not approaching the center of the triangle.

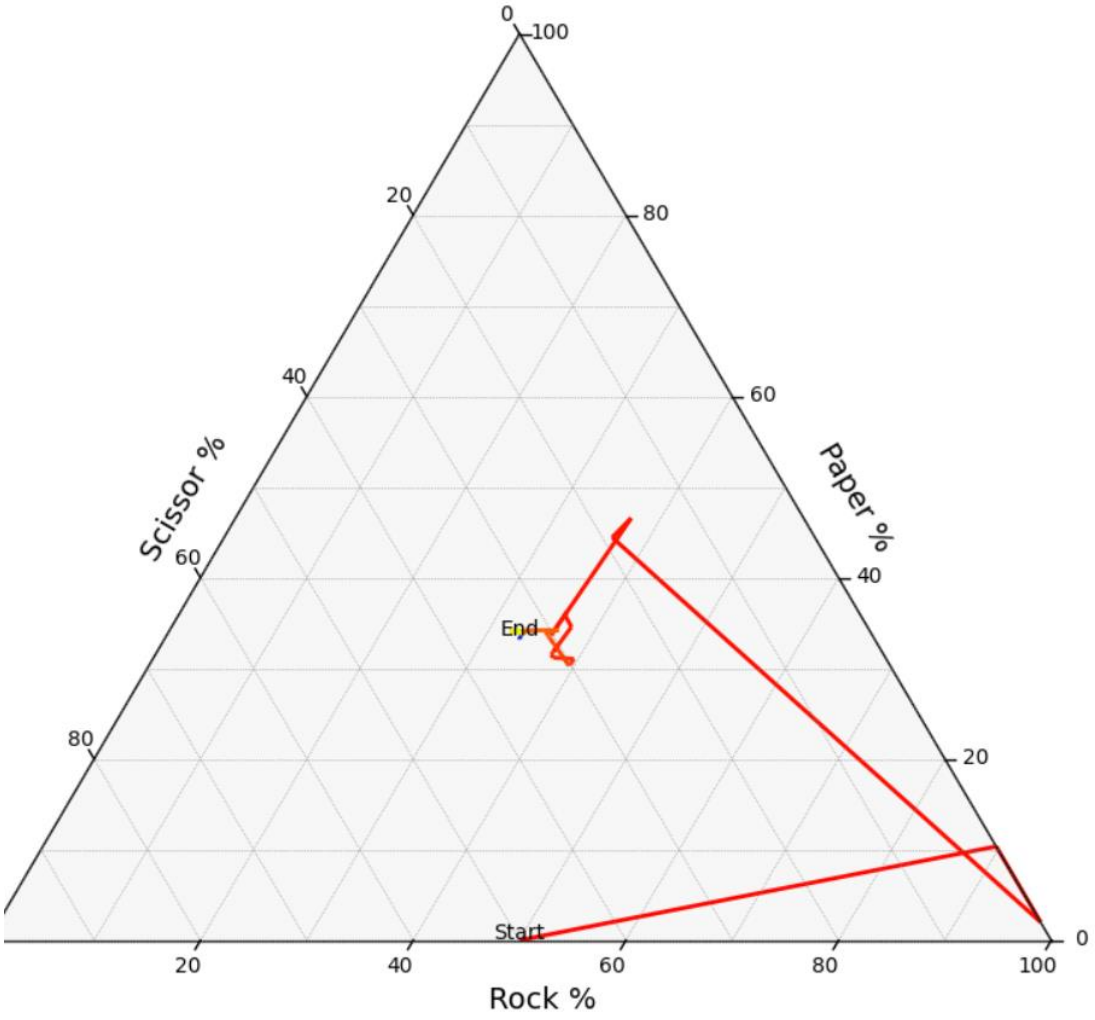





Agent 1 Average Reward Over Time
Average Reward
Episode
Agent 1 Average Reward


Agent 2 Average Reward Over Time
Average Reward
Episode
Agent 2 Average Reward

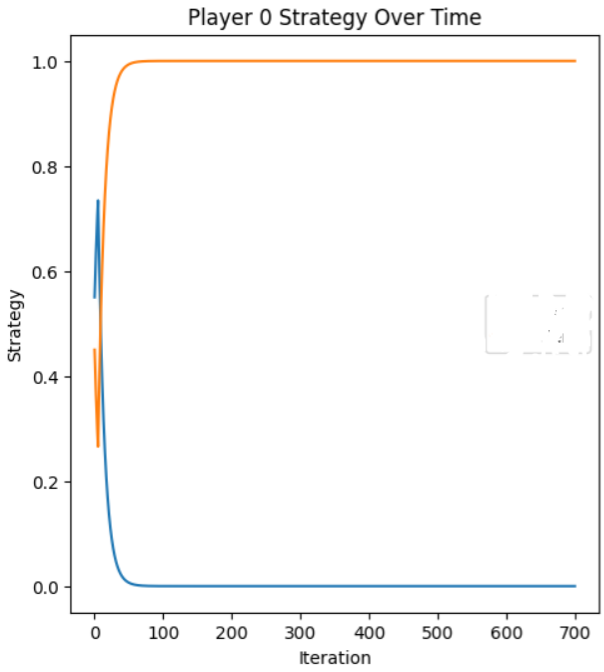

Scissor %
Paper %
Start
End

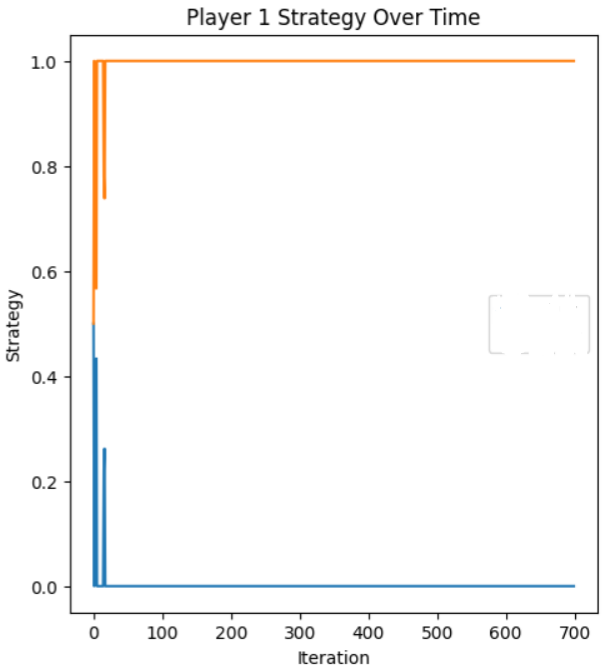


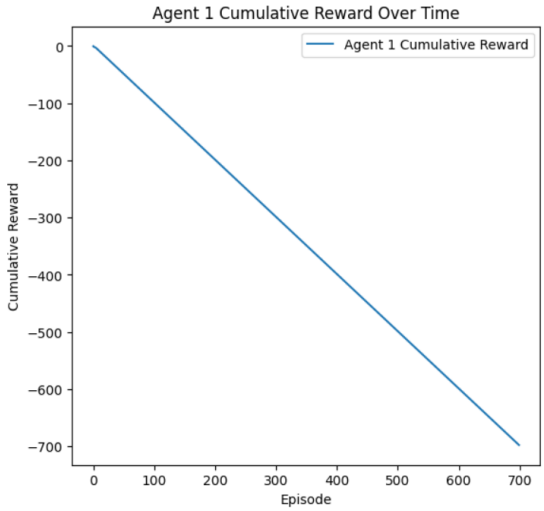
Fictitious Play agent versus Matching Pennies

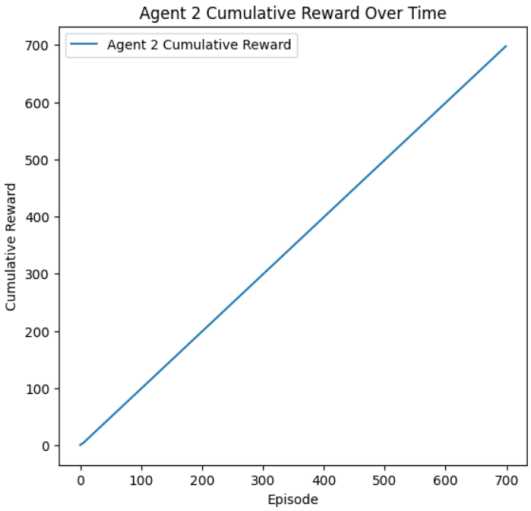
On this occasion, agents reached a Nash equilibrium but it was not the standard that was mentioned before. Regarding their rewards, one player consistently loses and the other consistently wins. Both agents show a preference in their strategy since they converge to a pure policy. The action probability lines become flat very quickly which can be interpretated as that they reach the equilibrium very fast. Sometimes, another pattern emerged on this occasion. The plots became very similar to the previous situation where fictitious play did not converge to a policy but q-learning converged to the known mixed Nash equilibrium of the game.

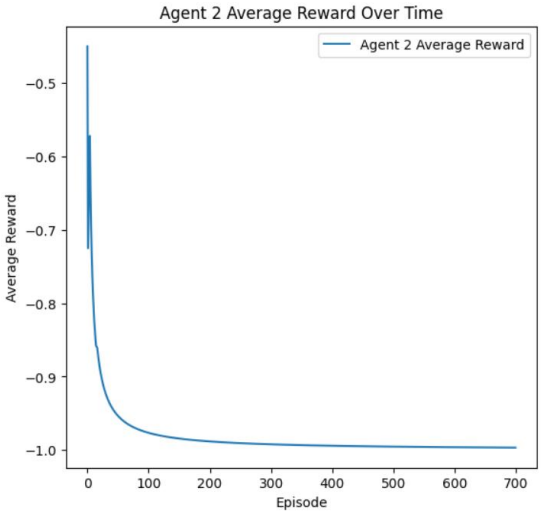
The Blue line depicts Heads and the orange tails.

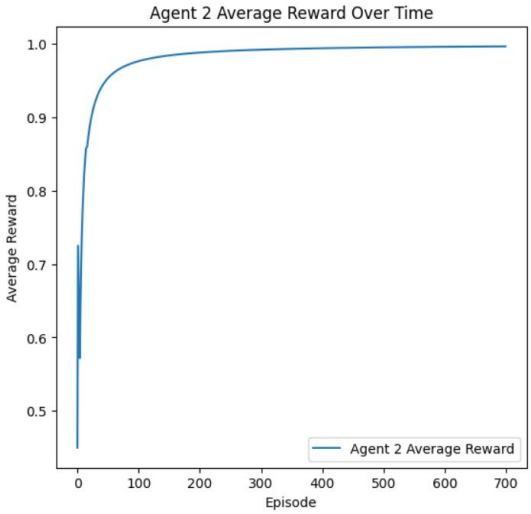

Player 0 Strategy Over Time
Strategy
Iteration

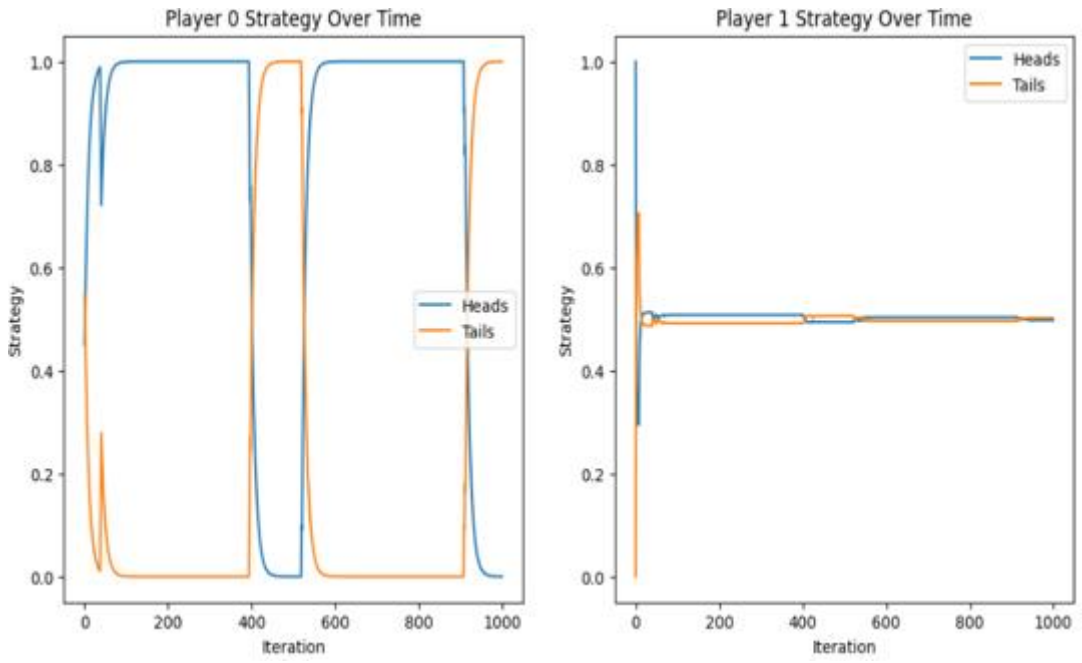

Player 1 Strategy Over Time
Strategy
Iteration


Agent 1 Cumulative Reward Over Time
Agent 1 Cumulative Reward
Cumulative Reward
Episode


Agent 2 Cumulative Reward Over Time
Agent 2 Cumulative Reward
Cumulative Reward
Episode

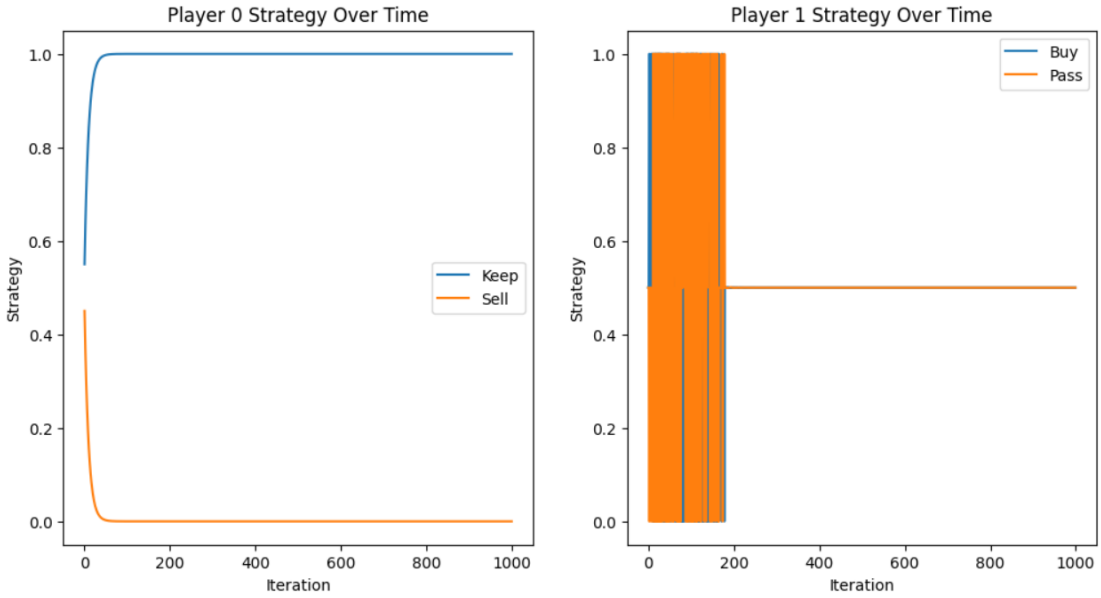

Agent 2 Average Reward Over Time
Agent 2 Average Reward
Average Reward
Episode

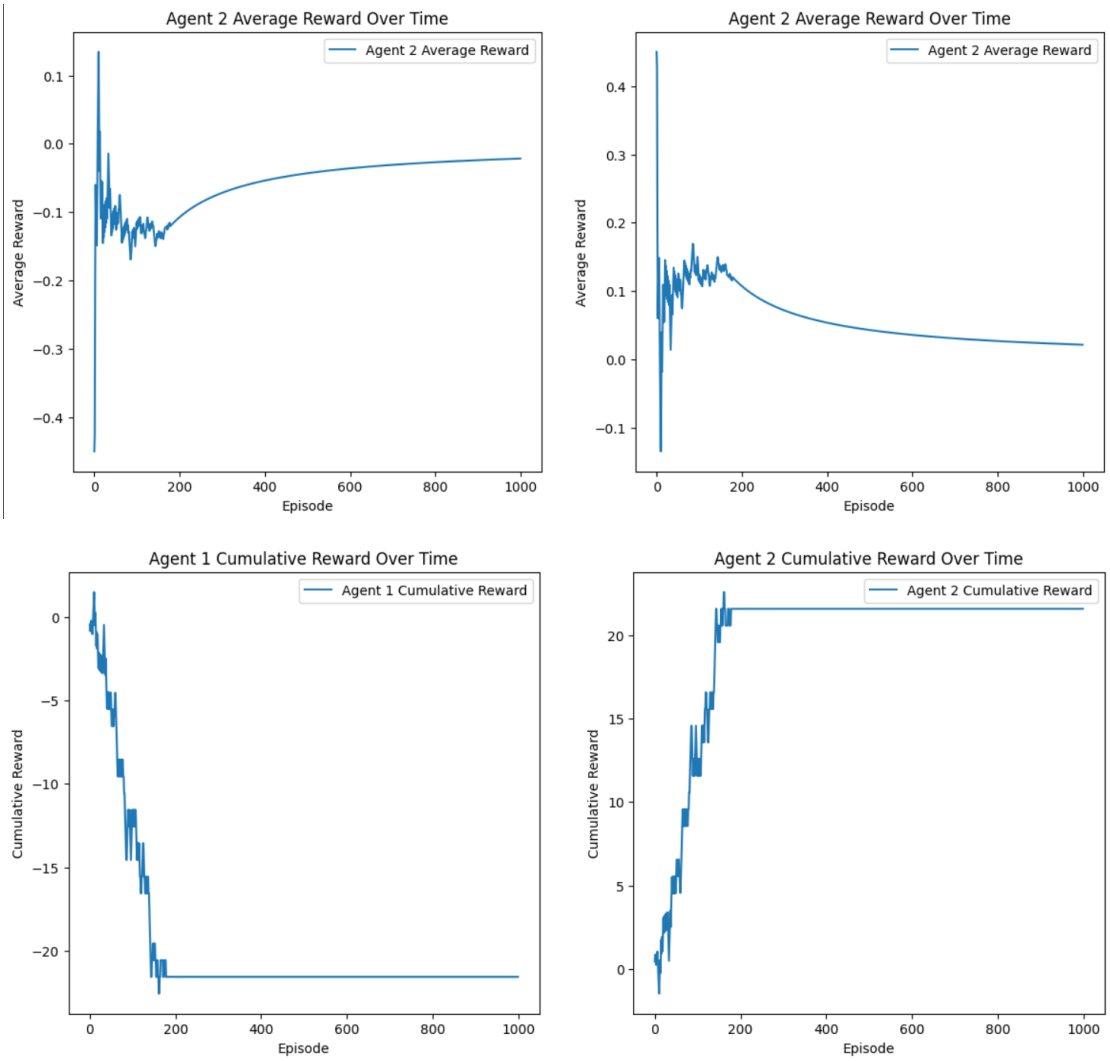

Agent 2 Average Reward Over Time
Agent 2 Average Reward
Average Reward
Episode



Fictitious Play agent versus RL agent Selling damaged goods

On this setting, fictitious row player rapidly converges to the pure strategy Keep, while column Q-Learning player becomes indifferent between its actions after some initial fluctuations in its policy. This mixed strategy Nash equilibria is expected based on the payoff matrix of the game. Average and cumulative reward become flat after 200 iterations, just when column's player fluctuations in its policy stop. The diagrams below ensure that the agents manage to find a Nash equilibrium because their policies are stable.





6.Conclusions

In this assignment, Reinforcement learning proved a more robust approach compared to fictitious play. When both agents competed against each other, RL did not struggle as much to reach equilibria comparing to Fictitious play. Moreover, RL algorithm converges to the equilibria faster than fictitious most of the times. Also, fictitious play did not converge to the same equilibria every time, while RL consistently found the same. The trend of the lines in cumulative reward, average reward and simplex ternary Diagrams verified the computation of equilibria. Most of the times, when the line of the reward flattens or reaches a plateau, then an equilibrium was reached by the agents. It is worth noting that FP vs RL diagrams differentiate often after each experiment. Initial beliefs affect FP. Also, FP has a tendency of reaching mainly pure Nash Equilibria and it converges a lot more smoothly than RL. Overall, both algorithms proved effective in computing equilibria. Against less sophisticated opponents, Fictitious play should be performing very well. When it competed against more intelligent ones, like Minimax Learning, then the task becomes harder.