Studying Game Theory Optimal Solution through Q-learning

Giovanni M. D'Antonio[1]

Harvard University
School of Engineering and Applied Sciences
[1]giovannidantonio@college.harvard.edu

Abstract

In this project, we explore how independent Q-learning agents learn in preset environments. More precisely, we define increasingly complex game theoretic systems to look at possible shortcomings and interactions these agents have in finding optimal solutions. We will start with some key definitions and past research, to then analyze behaviors in more and more complex games. The paper will explore the behavior of reinforcement learning agents in environments with clear optimal strategies, noticing how such agents do indeed find optimal solutions when those are clear, deterministic, and within a limited number of agents' interactions. Then, we will explore more complex games where a clear solution is often difficult to find, is non-deterministic, and where the amount of agents' interactions scales exponentially. It is in these setups that the biggest shortcomings are to be found; thus, suggesting, that the architecture of independent Q-Learning agents is incapable of dealing with games where optimal choices change constantly or when the number of agents is very high. On the other hand, the architecture proves to be a perfect fit, and faster to converge to optimal solutions than humans, in an environment where there is a limited number of agents' interactions as well as an always static optimal strategy.

1 Introduction and Motivation

Through the field of Game Theory, humans have successfully modeled some real-life games. After a complete and correct model is achieved, it is easier to find a "solution" to this game, thus "solving" the game. A Game is considered solved when the outcome (win, lose, draw) can be correctly predicted from any position, assuming that both players play perfectly. These "perfect moves" are called dominant strategies. A strategy S dominates another strategy S' if S leads to a better outcome, independent of what the opponent does.

Many game theory researchers believe that the field can successfully model the vast majority of decision-making scenarios. Through the concept of "games" we can look at different kinds of interactions and consequences in complex systems. Looking at how reinforcement learners behave can be fundamental in understanding new, and unexpected, ways to act on these games and act on real-world scenarios. Learning the shortcomings could be

fundamental in understanding when we can trust reinforcement learning agents to find the correct solution to real-world problems and when we should not. We can envision a future in which we have good enough and trained models, capable of deciding if it is optimal to go to dinner at this time at this restaurant or not. This future is not any more than distance, and studying reinforcement learning is a promising area of research in decision-making science. In this initial analysis, we will investigate the behavior of independent Q-learning agents. More precisely, we are interested in looking if they will eventually solve different kinds of games, always playing the dominant strategies against each other. Thus, converging to an equilibrium, called Nash Equilibrium. Furthermore, we are interested in understanding the shortcomings and the solution of these agents, so as to have new insights into problem solutions as well as correct usage of these models. Thus, converging to an equilibrium, called Nash Equilibrium.

2 Previous Research

Researchers have long struggled to find a reinforcement learning model that could fit the concept of Game theory with its complexities and generalization. Q-Learning or policy gradient result in being poorly suited to multi-agent environment as well as being unstable under constantly changing agent's policy. At first, the most promising results have been in learning in multi-agent settings using independently learning agents with Q-Learning (Tan 1997) but achieved poor results in practice (as the updates could not be immediately communicated). Another Interesting approach has been attempting to converge to specific behaviors such as cooperating via sharing of policy parameters (Gupta, Egorov, and Kochenderfer 2017), which required that every agent could have the same capabilities (an assumption that does not always hold in game theory Nash equilibrium scenarios). Promising research is present in the case of Q-Learning in a noncooperative multi-agent context. Here, researchers have the framework of sum stochastic games. The results have been very promising but limit themselves to stationary strategies, strategies that do not change over time over conditioning on historic plays (The framework used is very useful, but results are slightly different from my goal, I found this very useful regarding pseudocode) (Hu and Wellman 2003) 1 In other cases, researchers did not try to perfect reinforcement learning but instead preferred to use another approach altogether. Spike-based Decision Learning outperformed reinforcement learning in a number of games. This is more in the field of Neuroeconomics. (Friedrich and Senn 2012) Talking about new approaches, as I will explain more in detail later. I want to use as a key resource a paper working around the concept of different Q-Learning agent criticizing and acting on each other (more details below). This is an innovative approach that is generalizable on different kind of games (cooperative, mixed, competitive) (Lowe et al. 2020) Why is this important? The question of whether reinforcement learning agents can reach Nash equilibrium and how fast they can do so is important as it touches on the core of decision-making in complex, interactive environments. Once we understand and improve the efficiency of these algorithms we will be able to optimize interactions in fields such as economics, biology, and artifi-

cial intelligence, where strategic interactions and optimal decision-making are vital. One important concern is the speed of convergence as it is particularly important for applications requiring real-time decision-making or operating in dynamic environments.

3 Foundational Definitions

- **Game**: A game, in the mathematical sense, is a situation in which players make rational decisions according to defined rules in an attempt to receive some sort of payoff.
- **Strategy**: A strategy is a complete algorithm for playing the game, telling a player what to do for every possible situation throughout the game.
- **Dominant Strategy**: A dominant strategy S is a strategy that provides optimal outcomes for the player regardless of what the other player does. We say that Strategy S strictly dominates strategy S' if all the outcomes of S are higher than the outcomes of S'. Or strategy S' is strictly dominated by strategy S.
- **Nash Equilibrium**: A Nash equilibrium occurs when each player in a non-cooperative game has chosen a strategy such that no one can increase their own expected payoff by changing their strategy, assuming other players keep theirs unchanged. If the strategies remain constant this is also called a Pure Strategy Nash Equilibrium. This obviously happens when everyone plays their own dominant strategy.
- **Mixed Strategy Nash Equilibrium**: A mixed strategy Nash equilibrium involves at least one player playing a randomized strategy and no player is able to increase his or her expected payoff by playing an alternate strategy.

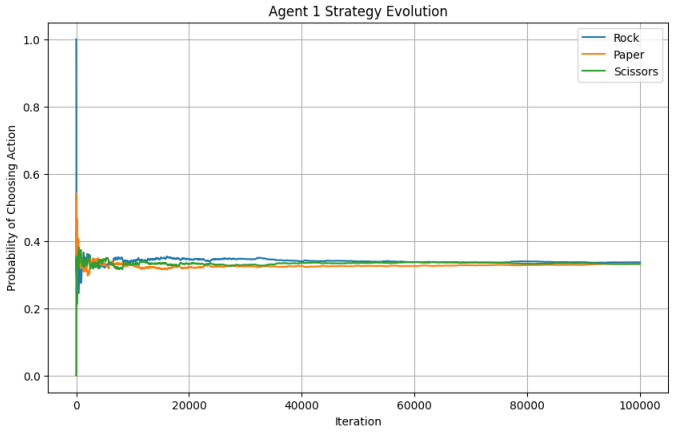
4 Pure Nash Equilibrium Games with Finite Agents

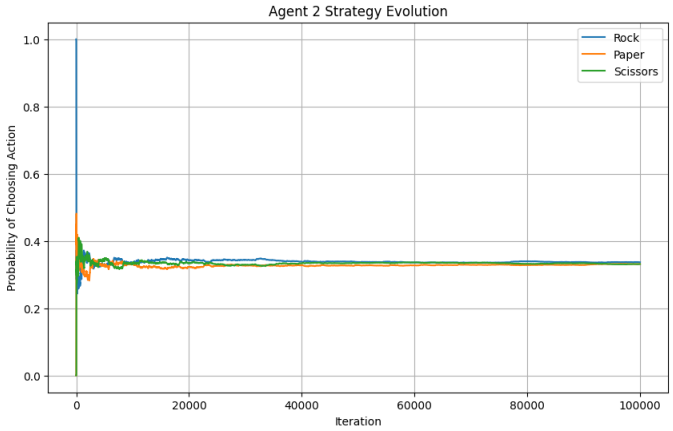
4.1 Rock, Paper, and Scissor

We all used to play Rock, Paper, and Scissors as kids. The game is simple. We have two players choosing, at the same time, one of three actions: Rock, Paper, or Scissor. Then the two actions are compared according to the following order where R that beats S beats P that beats R.

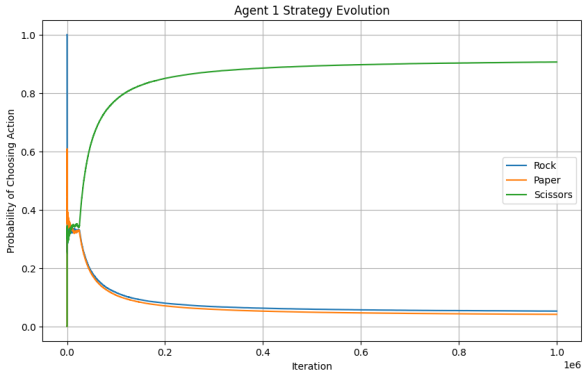
4.2 Practical and Mathematical Analysis of Rock, Paper, and Scissor

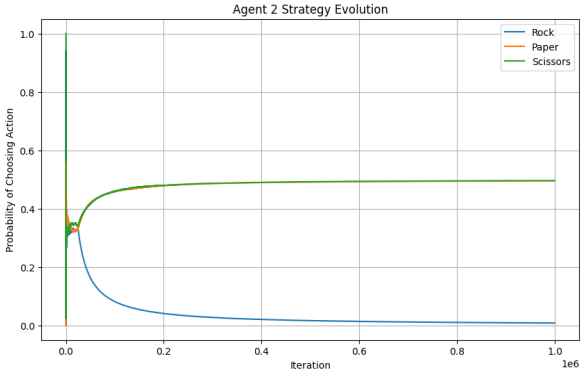
In the practical analysis I have undertaken, I have worked through training two independent Q-Learning agents in order to make them learn the most optimal strategy in the game. According to game theory optimal dominant strategies, in RPS, we should always randomly choose among the three possibilities with equal probabilities. Interestingly enough, in the following graphs, we notice that the two agents actually do converge to this solution. After an initial period of exploration they then converge to pick each choice with probability 0.3. This is very promising results suggesting that in an environment with finite agents and pure nash equilibrium the RL models do converge to the optimum.





We then perturbate the system, making Agent 2 unable to play Scissor. This is because I was interested in seeing how fast agents would react to suboptimal strategies by the opponent. This same result is applicable in every successive category and is therefore omitted from the next ones. As demonstrated by the subsequent graphs, agents are pretty fast in adaptation. This is due to the inherent structure of the Q-Learning algorithm (for the purposes of this paper, we have assumed the reader has knowledge of it). More precisely, the Q-Learning algorithm works as an expected payoff maximizer at each iteration (other than exploration) and updates these payoffs based on the results of its actions. By the nature of this algorithm, when we perturbate the system the new iterative updates will immediately change making the agents converging to the second most optimal strategy available. If we make Agent 2 unable to play rock, it then follows that to not be exploited playing always the same action among paper and rock it will randomly pick 0.5 of the times each, which is what we see it does. On the other hand, Agent 1 now knows that in expectation (by Q-Learning) Scissor will have always the highest expected payoff since it can not be beaten since Agent 2 is unable to play Rock. This makes Agent 1 not really playing the most optimal action but the least exploitable.





4.3 Prisoner's dilemma

Let's now look at a slightly more complex game with no such clear outcomes. The prisoner's dilemma is a game in which two "prisoners" are placed in two different rooms. The police think that one of the two committed a crime and want to understand who, so they ask each separately to talk and say who did what. One prisoner does not know what the

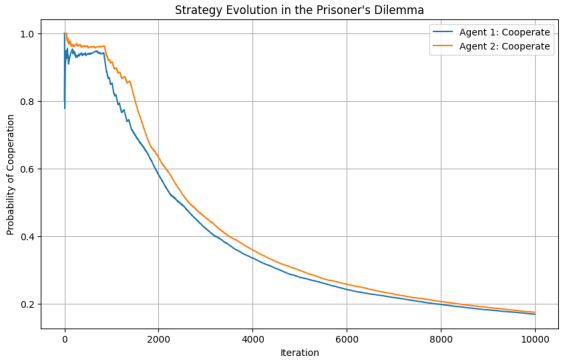
other says. They have 2 possible actions C (Cooperate) and D (Defect). The payoffs, in my version, are as follows:

- If Both Cooperate they both get +3
- If Both Defects they both get +1
- If One Cooperates and the other Defects the one who cooperates gets +0 and the one who defects gets +5

Let's look at what the two independent Q-Learning agents do in practice

4.4 Practical and Mathematical Analysis of Prisoner's Dilemma

The two agents start by Exploring different kinds of actions, starting from cooperation. We do notice that it seems that agents will constantly cooperate with a probability of about 0.9. Then something happens around iteration 1000, and one of the two starts defecting which creates a vicious cycle. Now, the other agent is punished for cooperating, and in expectation if Agent One has 0.9 probability of cooperating Agent 2 by defecting has an expected payoff of $0.9 * 5 = 4.5 > 3$, thus, it will stop cooperating. This creates a vicious cycle in which they both stop earning way less payoff than what they would do in any other possible strategy. It is interesting to see that they fall in the same traps as humans, but, again, in a Pure Strategy Nash Equilibrium game they still reach the Nash Equilibrium (never cooperating in this case), even when it is not the most optimal. Thus, empirically, independent Q-Learning Agents reach the Nash Equilibrium, in Pure Strategy Nash Equilibrium with finite agents games both when it is optimal to do so (RPS) and when it is not (Prisoner's Dilemma). This result has serious implications, as it shows that Q-Learning Agents may be not useful when searching for optimal solutions, but they may still be studied in order to find new and possible Nash Equilibriums.



5 Pure Nash Equilibrium Games with possibly Infinite Agents

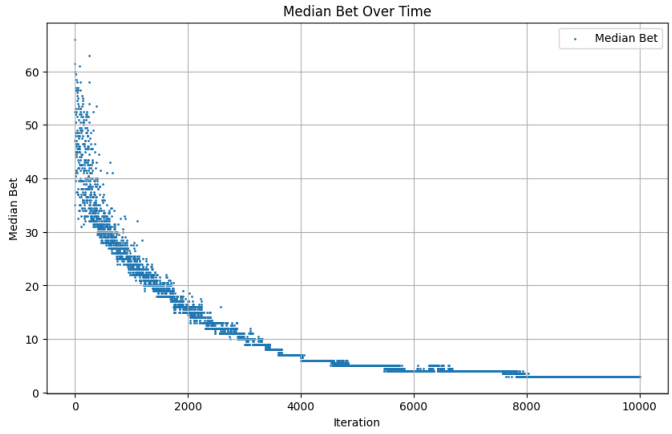
5.1 p-Beauty Game

A different, more complex, kind of game is the set of games that have a Pure Nash Equilibrium but that can possibly have infinitely many agents, causing the overall complexity of interactions to increase. One of these games is the p-Beauty game, an auction game popularized by Game Theorists to study how our minds work and how we think of ourselves compared to our peers. In my version of it, the rules are as follows:
"Choose a number between 0 and 100. The winner is the person whose number is closest to 2/3 times the average of all chosen numbers. The winner gets a fixed prize of 20 dollars. In case of a tie, the prize is split amongst those who tie." According to a game theoretic analysis, the dominant strategy of this game is based on the interactions with all the other players. If all the players know game theory and play rationally, we expect the equilibrium to be everyone betting 0 dollars. This is because, if you assume that everyone bets 66 (around 2/3 of 100) you should then play 44 (around 2/3 of 2/3 of 100) continuing like this up until reaching 0. Will the

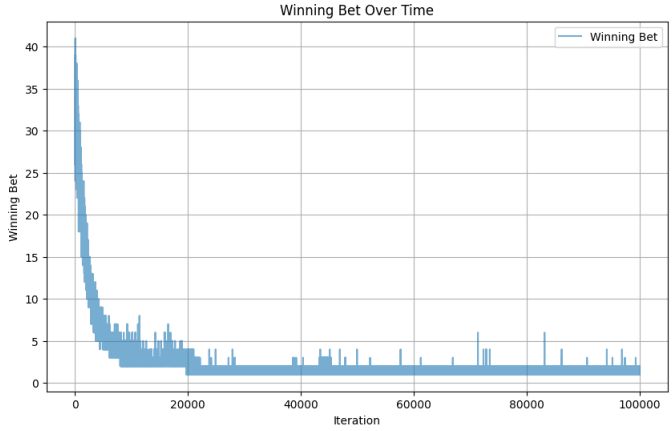
Q-Learning agents understand this? In my p-beauty game, I have chosen 50 agents and ran some experiments

5.2 Practical and Mathematical Analysis of the p-Beauty Game

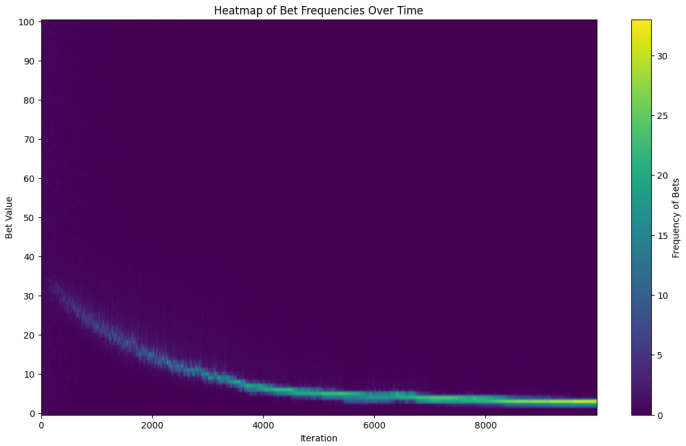
Plotted below there is the evolution of the median bet over time, the winning bet over time, and the heatmap of the bet frequencies over time.



All the agents start by giving higher expectations to around 66 (which makes sense at the first iteration) but we then notice that the winning bet is way lower than that. At this point, agents understand that they need to lower their own bets and gradually move toward 0. However, they then asymptotically reach around 1 instead of actually reaching 0, again, a sub-optimal solution that is not even the correct Nash Equilibrium strategy.



Interestingly enough, instead of humans, who arrive at an understanding through multiple iterations that the best solution is for everyone to bet 0, Q-Learning agents do not arrive at it opening different scenarios.



By the heatmap, one would think that the next reasonable step is to lower the bet, but Q-Learning does not do that, as it is stuck with the expected payoff it has. This is a very surprising result, one that is worth investigating further, and that I am currently working on for future papers.

6 Mixed Nash Equilibrium Games with possibly Infinite Agents

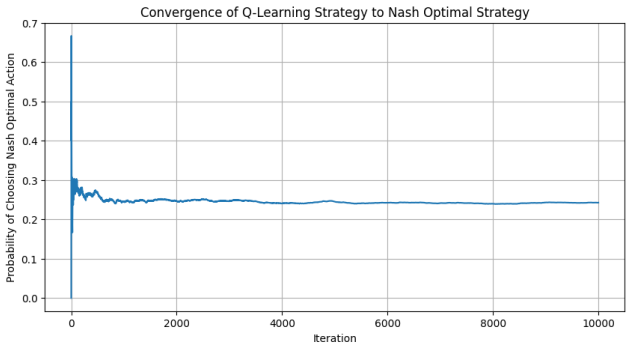
6.1 Kuhn's poker

Real life is even more complex than that. One of the most complex games to have been analyzed is poker. Poker has the peculiarity of not having optimal strategies unconditional on what happens in the game, but it only has game theory optimal strategies conditional on the cards drawn. This adds a layer of complexity, as the Q-Learning algorithm updates at the end of the iteration, thus, not truly capturing the meaning of action with the drawn cards (agents will choose their actions unconditionally on the cards). More precisely, the mixed strategy nash equilibrium of this game is for the first player

(Player 1), the equilibrium mixed strategy is to bet with probability 2/3 when holding a King, 1/3 when holding a Queen, and never bet with a Jack. For the second player (Player 2), the equilibrium mixed strategy is to call with probability 2/3 when the first player bets, and 1/3 when the first player checks. (van der Werf 2022)

6.2 Practical and Mathematical Analysis of Kuhn's poker

The implementation of Kuhn's poker resulted to be the hardest, with the agents learning in a subotpimal way. The end results is that agent converge to the optimal strategy just 1/3 of the times almost if they were guessing the next move. This is most likely caused by the shortcomings of the Q-learning algorithm that do require additional engineering to be overcome. More precisely, the Q-learning algorithm has a payoff matrix uncoditional on the card drawn or the action see, maybe creating a multidimensional Q-Learning would solve this problem. Overall, it is interesting to note how Agent 1 bets way more often than agent 2 which is what we know to be correct. Furthermore, both agents move their rate of bluffing to almost zero, most likely as they seem to learn that bluffing is almost never a good strategy! This is indeed true in real life too, where professional poker players tend to bluff way less than the past decades due to the concerns that the opponent is playing game theory optimal strategies regarding of what the other does or says. In this way, you are guaranteed to win in the long run if the other does not play always game theory optimal moves.



7 Future Work

- **Alternative path of learning**: One possible area of improvement is to change the algorithms through which reinforcement learning agents learn. More precisely, right now, we are dealing with multiple independent Q-learning agents instead of working through an actual multi-agent re-

inforcement learning model. To the same extent, it would interesting to investigate other techniques that could achieve similar results.

- **Real Poker**: Another really interesting extension of the work done so far, would be to train reinforcement learning agents on real poker. This would be different from "poker bots" as we would make them learn unsupervisingly from the outcome of each iteration they go through. It would be interesting to see if, in a more complex environment where nash equilibrium is not clear, reinforcement learning agent will, nonetheless, converge to a solution.
- **Real life situation**: The ultimate goal of this work is to have games capable of correctly testing the reinforcement learning agents we have. Once these games are achieved then we can model real life situation as games, and, hopefully, have an RL model capable of giving an answer. Imagine you need to know if you should go out with John or Katherine, input their characteristics and yours, and a model suggests the most optimal expected utility.

8 Conclusion

As it was my very first time with coding outside of introductory courses and the first working with RL agents and game theory, I feel very satisfied with the work done. These results could be useful to build more advanced model capable of modeling real life decision making. Through testing, the independent Q-Learning agents approach proved very useful in Pure Nash Equilibrium game with finite agents (RPS), correctly predicting the dominant strategies and assessing the optimal payoffs. To the same extent, it was interesting to see how, by increasing the number of agents, the randomness due to the exploration rate causes noise inside the system which blocks other agents to learn effectively (p-beauty game) ultimately leading to not reach the Nash Equilibrium. Furthermore, interestingly enough, RL agents are not capable to understand the overall best strategy but ultimately remain stuck in short-sighted suboptimal scenarios (prisoner's dilemma). Moving to a more complex environment, Kuhn's poker unveiled some of the main concerns and shortcomings regarding the independent Q-Learning approach. The iterative updates of the payoff matrix do not take into account the conditionality caused by drawing one card rather than another. This,

References

Friedrich, J.; and Senn, W. 2012. Spike-based Decision Learning of Nash Equilibria in Two-Player Games. *PLOS Computational Biology*, 8(9): 1–12.

Gupta, J. K.; Egorov, M.; and Kochenderfer, M. J. 2017. Cooperative Multi-agent Control Using Deep Reinforcement Learning.

Hu, J.; and Wellman, M. 2003. Nash Q-Learning for General-Sum Stochastic Games. *Journal of Machine Learning Research*, 4: 1039–1069.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2020. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments.

Tan, M. 1997. Multi-Agent Reinforcement Learning: Independent versus Cooperative Agents.

van der Werf, L. 2022. Analysis of Nash equilibria for Kuhn poker and its extensions.