# Report for the First Assignment of Reinforcement Learning Course, CISC 856, Fall 2019

Milad Khademi Nori (20187733), Azadeh Motamedi (20188887)

## I. First Policy

### A. Prisoners' Dilemma

(a) For the first game, namely the prisoners' dilemma, the policy of the player $j$ almost converges to $p^j = [p_1^j, p_2^j] = [0,1]$, where $p_1^j$ and $p_2^j$ denote the probability of cooperation and defection. These number indicate that player $j$ (prisoner $j$) has the tendency not to cooperate. According to simulations, Figure 1 and 2 show the trajectory of policies that converge to the origin where indicates that prisoners are discouraged from cooperating. It is worth mentioning that the policy distribution at the beginning of iteration is shown by violet; it changes color gradually until it reaches the final point shown in red.

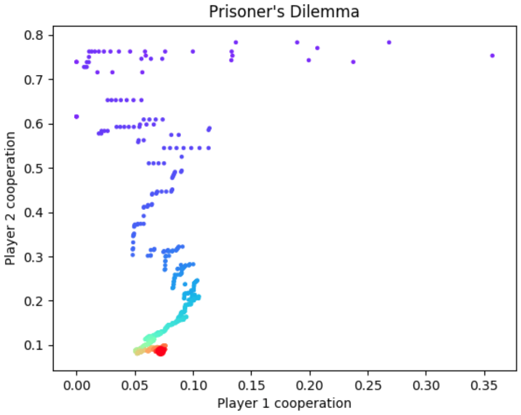

Fig. 1. Policy convergence trajectory of player 1 cooperation versus player 2 cooperation. Start point $(0.35, 0.75)$, end point $(0.07, 0.08)$

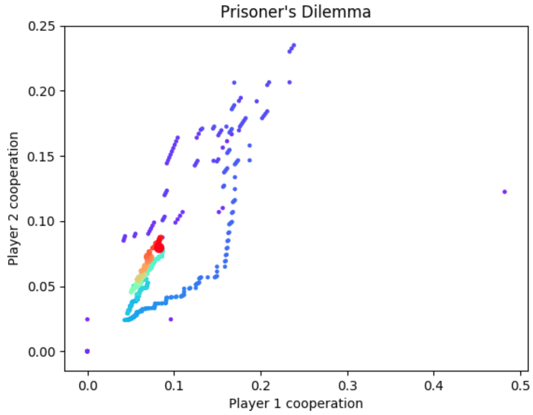

Fig. 2. Policy convergence trajectory of player 1 cooperation versus player 2 cooperation. Start point $(0.48, 0.12)$, end point $(0.07, 0.07)$

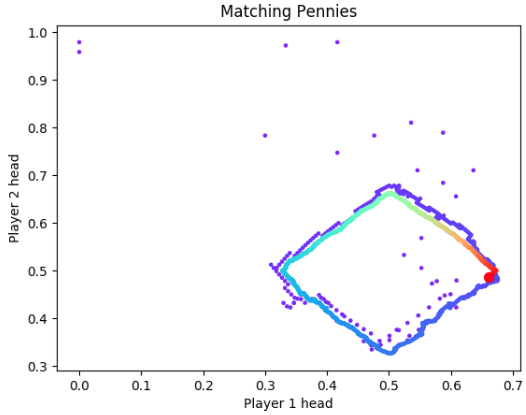

Fig. 3. Policy convergence trajectory of player 1 head versus player 2 head. Start point $(0.91, 0.92)$, end point $(0.53, 0.62)$

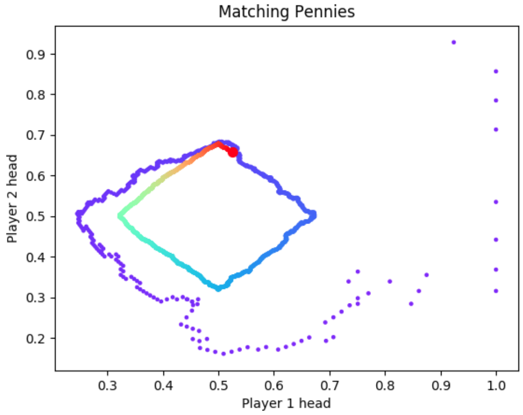

Fig. 4. Policy convergence trajectory of player 1 head versus player 2 head. Start point $(0.48, 0.12)$, end point $(0.07, 0.07)$

(b) No, the policy is not optimal (or Pareto optimal), it is a Nash equilibrium. Nash equilibrium is an outcome in which every player is doing the best he possibly can, given the other players' choices. Therefore, no player can benefit from unilaterally changing his choice. In this example if any player changes his choices he get 0 as opposed to 1 that he is currently receiving.

### B. Matching Pennies

(a) The policy of the player $j$ almost gets close (does not meet) to $p^j = [p_1^j, p_2^j] = [0.5, 0.5]$, where $p_1^j$ and $p_2^j$ denote the probability of selecting head and tail. According to simulations, Figure 3 and 4 show the trajectory of policies that do not converge to the center $(0.5, 0.5)$ and circulates it [1]. Players try to select each side of the coin with equal probability which has the most uncertainty for the opponent.

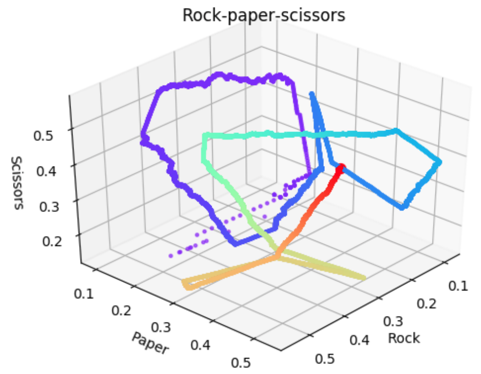

Fig. 5. Policy convergence trajectory of the first player. Start point $(0.55, 0.3, 0.15)$, end point $(0.22, 0.40, 0.38)$

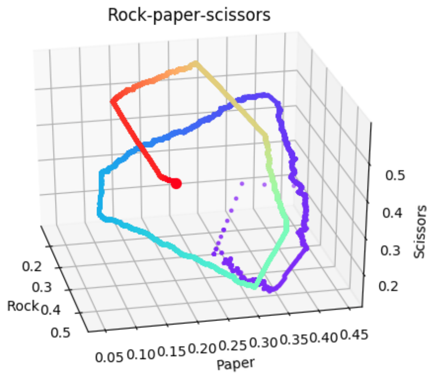

Fig. 6. Policy convergence trajectory of the first player. Start point $(0.25, 0.45, 0.30)$, end point $(0.30, 0.25, 0.45)$

(b) Yes, the policy is optimal (Pareto optimal) and it is a Nash equilibrium as well. In this example if any player changes his choices from 0.5 to more or less, he would decrease the uncertainty of the opponent. Besides, this game is a zero-sum game (fair game) which implies that expectation of reward is 0 for both.

### C. Rock-paper-scissors

(a) The policy of the player $j$ almost converges to $p^j = [p^j_1, p^j_2, p^j_3] = [1/3, 1/3, 1/3]$, where $p^j_1$, $p^j_2$, and $p^j_3$ denote the probability of selecting rock, paper, and scissors, respectively. According to simulations, Figure 5 and 6 show the trajectory of policies that gets close to the $(0.33, 0.33, 0.33)$ and circulates [1]. Players select each action with equal probability which has the most uncertainty for the opponent.

(b) Yes, the policy is optimal (Pareto optimal) and it is a Nash equilibrium as well. In this example if any player changes his choices from 0.33 to more or less, he would decrease the uncertainty of the opponent.

## II. Second Policy

### A. Prisoners' Dilemma

(c) This time with the new policy iteration algorithm, it has been observed that the algorithm converges to the origin considerably smoother because the appended term serves as a stabilizer; more precisely, when the first term spikes and pushes the policy value to increase, the second term (appended term) somehow damps the immediate changes and guides the policy toward the expected value. This term might have some downsides when the process is non-stationary that makes the policy less responsive to new environments. Figure 7 and 8 corroborates our expectations.

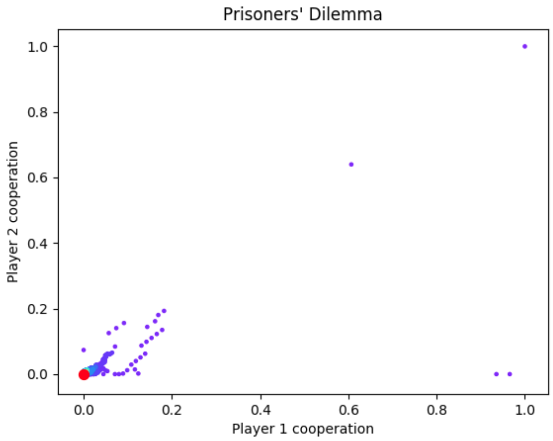

Fig. 7. Policy convergence trajectory of player 1 cooperation versus player 2 cooperation. Start point $(1, 1)$, end point $(0, 0)$

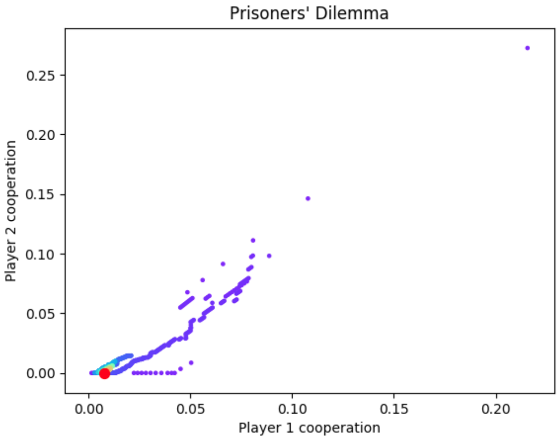

Fig. 8. Policy convergence trajectory of player 1 cooperation versus player 2 cooperation. Start point $(0.22, 0.26)$, end point $(0.01, 0)$

(d) As mentioned before, it is not optimal since there is another point on which both players can receive more reward, but we can say it is a Nash equilibrium.

(e) The value of the game is typically about 1.7. The decimal part stems from the occasion when players have not yet been trapped in Nash equilibrium.

(f) Because the extra term helps to stabilize policy iteration; more precisely, when the first term bounces and pushes the policy value to fluctuate, the second term (added term) somehow hampers the instant changes and leads the policy toward the expected value.

### B. Matching Pennies

(c) The new algorithm performs better for matching pennies game, gets closer to $(0.5, 0.5)$ point, and circulates it but still

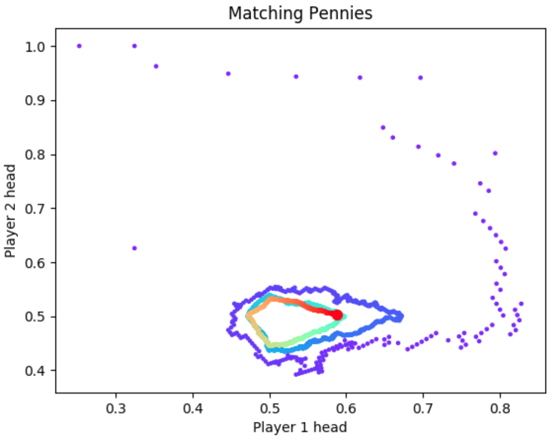

Fig. 9. Policy convergence trajectory of player 1 head versus player 2 head. Start point $(0.10, 1)$, end point $(0.59, 0.50)$

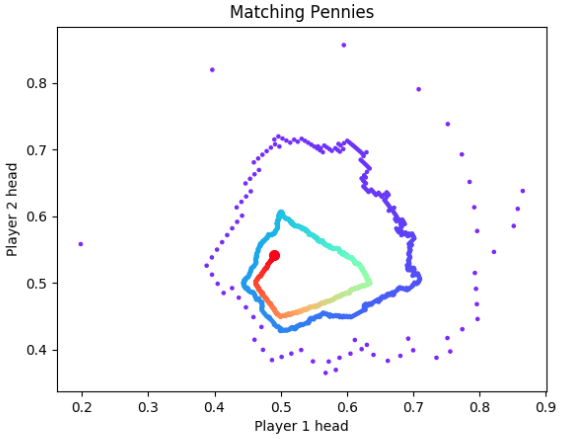

Fig. 10. Policy convergence trajectory of player 1 head versus player 2 head. Start point $(0.20, 0.56)$, end point $(0.50, 0.55)$

does not converge to it [1] (shown in Figure 9 and 10).

(d) This policy is both a Nash equilibrium and a Pareto optimal because it imposes the maximum uncertainty to the opponent.

(e) The value of the game is typically between $-0.1$ and $0.1$ since this game is a zero-sum game (fair game) which implies that expectation of reward is $0$ for both.

(f) As explained before, the added term hampers the bounce of policy value and directs it toward expected value.

### C. *Rock-paper-scissors*

(c) The modified algorithm works better for rock-paper-scissors, approaches $(1/3, 1/3, 1/3)$ point, and wanders around it nevertheless does not meet it [1] (shown in Figure 11 and 12).

(d) The policy is both a Nash equilibrium and a Pareto optimal because it has the maximum entropy possible.

(e) The value of the game is typically between $-0.1$ and $0.1$ since this game is a zero-sum game (fair game) which implies that expectation of reward is $0$ for both.

(f) The added term hampers the bounce of policy value and directs it toward expected value.

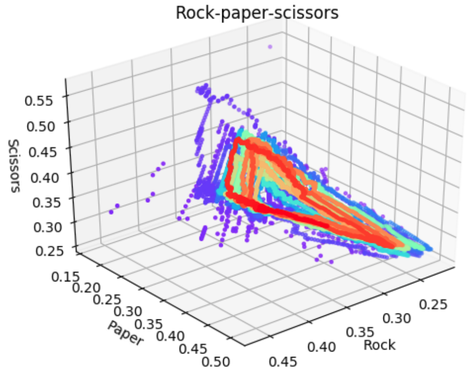

Fig. 11. Policy convergence trajectory of player 1 head versus player 2 head. Start point $(0.10, 1)$, end point $(0.59, 0.50)$

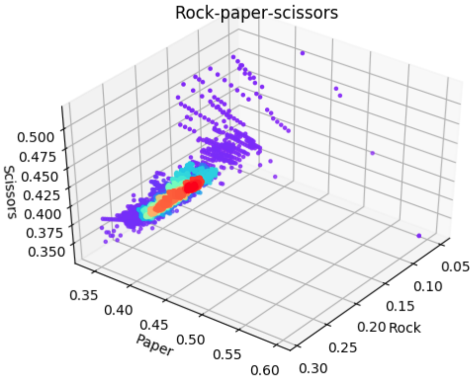

Fig. 12. Policy convergence trajectory of player 1 head versus player 2 head. Start point $(0.20, 0.56)$, end point $(0.50, 0.55)$

## III. Notes About the Code

There are 12 files of code, 6 of them are the main code, and the others serve as a function file for the main. The names of the files are chosen appropriately, such that they totally indicate for which part of the assignment they are. Codes are written in Python 3.6.8 with standard libraries, namely, Numpy and Matplotlib. The VS code is used as the editor of the code.

## References

[1] T. N. Cason, D. Friedman, and E. Hopkins, “Cycles and Instability in a Rock-Paper-Scissors Population Game: A Continuous Time Experiment,” *The Review of Economic Studies*, vol. 81, no. 1, pp. 112–136, 09 2013. [Online]. Available: https://doi.org/10.1093/restud/rdt023