Discrete-Time Heavy-Anchor Reinforcement Learning in Multiagent Finite Games

by

Rupin Khadwal

Supervisor: Prof. Lacra Pavel

April 2024

Abstract

This paper introduces a reinforcement-learning approach designed to converge towards the true Nash Equilibrium in multiagent finite games. Building upon Gao et al.'s model,[1] which utilized passivity-based controls and Q-learning for Nash Equilibrium determination, this paper addresses limitations observed in achieving only a logit Nash Equilibrium in monotone games. To rectify this, this paper proposes a modification by transitioning from Q-learning to P-RL and incorporating an additional layer of Heavy Anchor Dynamics. The inclusion of Heavy Anchor Dynamics feedback is crucial to prevent the P-RL model from cycling in monotone games, ensuring convergence towards the true Nash Equilibrium. Furthermore, this was completed in the discrete-time domain, and a new set of models that integrate both Heavy Anchor Dynamics and P-RL was introduced. All simulations have been completed in Python and plotted using the Matplotlib and Plotly libraries.

Table of Contents

Abstract ii
1. Introduction 1
2. Literature Review 2
2.1 Introduction 2
2.2 Definitions 2
2.3 The Motivation to use RL in NE-Seeking 3
2.4 Game Theory 3
2.5 NE-Seeking 4
2.6 Passivity-Based Controls for NE Seeking 5
2.7 P-RL & Q-Learning in Continuous Time 6
2.8 Heavy Anchor Dynamics 7
2.9 Q-Learning with Side Information 8
3. Methods 10
4. Results and Discussion 12
4.1 Game 1 RPS 12
4.2 Game 2 Anti-coordination 13
4.3 Game 3 Matching Pennies 15
4.4 Game 4 Shapley 16
4.5 Game 5 Modified RPS 18
5. Conclusion 20
6. References 20
List of Figures & Tables iv

1. Introduction

This investigation aims to explore the potential development of a reinforcement-learning (RL) scheme that converges to the true Nash Equilibrium (NE) in multiagent finite games. This study builds upon the 2021 paper by Bolin Gao and Lacra Pavel, which utilized a passivity-based approach and achieved convergence to a perturbed NE [1]. Their work on Q-learning research introduces a different RL scheme, P-RL [2]. Furthermore, research on Heavy Anchor Dynamics and Q-Learning with side-information proved that convergence to the true NE can be attained in monotone games [3] [4]. While Gao and Pavel's model demonstrated convergence in games characterized by the monotonicity property of negative payoff vectors, the proposed approach in this paper aims to identify a solution in a discrete-time model with side-information, as most models are implemented in such time.

2. Literature Review

2.1 Introduction

This investigation ambitiously embarks on the mission of developing an advanced reinforcement-learning (RL) scheme aimed at achieving the authentic Nash Equilibrium (NE) in the complex landscape of multiagent finite games. Building upon the pioneering work of Gao and Pavel, this novel approach incorporates Heavy Anchor Dynamics and Q-Learning with Side-information to assess the feasibility of attaining convergence to the true NE within a discrete-time model. The quest for NE-seeking in non-cooperative games propels our purpose [3][4]. RL, with its adaptable nature and minimal informational requirements compared to traditional methods like fictitious play and gradient play, emerges as a potential solution. The overarching goal expands beyond mere convergence, delving into the unexplored realms of RL's applicability in scenarios marked by incomplete information. The purpose is dual-fold: extending Gao's model and proposing innovative RL approaches, with a distinct emphasis on convergence within discrete-time models, aligning with the practical implementation of most models.

Exploring the basic concepts of Game Theory and Nash Equilibrium seeking, will deliver a base understanding of important notions to help move into advanced and specific concepts. Thereafter, we introduce four pivotal components: Passivity-Based Controls for NE Seeking, P-RL & Q-Learning in Continuous Time, Heavy Anchor Dynamics, and Q-Learning with Side Information. The comprehensive roadmap extends from foundational concepts to detailed analyses of each critical component, providing readers with a holistic view of our pursuit for true NE convergence in the dynamic landscape of multiagent finite games. A few important definitions are also mentioned.

2.2 Definitions

Nash Equilibrium: The point in a non-cooperative game wherein players have no incentive to unilaterally deviate from their strategy. [5]

Reinforcement Learning: An area of machine learning that focuses on rewarding/punishing desired/undesired behaviors to "teach" the agent an optimal policy. [6]

Continuous-Time System: A continuous-time system is a dynamical system where the input, output, and state variables are defined at every instant of time in a continuous manner. The system evolves smoothly and continuously over time. [7]

Discrete-Time System: A discrete-time system is a dynamical system in which the input, output, and state variables are defined at distinct, separate time instances. The system evolves in discrete steps or intervals, with time progressing in a quantized manner. [8]

2.3 The Motivation to use RL in NE-Seeking

Nash Equilibrium-seeking is a challenging problem without a general solution for non-cooperative games. Reinforcement learning (RL) in multiagent finite games, especially those with incomplete information, has gained interest due to its weak informational requirements.

While RL shows potential for solving games that other methods cannot, challenges remain. Prior research mainly focused on convergence results in potential games and two-player zero-sum games. Gao and Pavel's research addressed this gap, proposing that passivity-based control theory could make existing RL schemes converge in N-player monotone and hypomonotone games.[1]

2.4 Game Theory

Game Theory, a branch of applied mathematics, provides a robust framework for analyzing situations characterized by interdependent decisions among multiple parties, known as players. The essence of Game Theory lies in unraveling the strategic intricacies that unfold when players formulate decisions, considering the potential moves of others. Originating from the collaborative efforts of John von Neumann, a Hungarian-born American mathematician, and Oskar Morgenstern, a German-born American economist, Game Theory was initially conceptualized to address challenges in economics. Their seminal work, "The Theory of Games and Economic Behavior" [9], argued for the inadequacy of traditional physical sciences and mathematics in capturing the strategic dynamics inherent in economic interactions. Instead, they proposed Game Theory as a novel mathematical approach suited to the nuanced decision-making processes involved in economic activities. [10]

Game Theory encompasses scenarios where players may have similar, opposed, or mixed interests, leading to a diverse array of potential outcomes. The strategic considerations in decision-making, as opposed to pure chance, set Game Theory apart from classical probability theory. Its applications extend far beyond traditional parlour games, permeating various fields such as politics, business, pricing strategies, voting dynamics, jury selection, and even ecological studies of animal and plant behaviors. The theory aids in predicting and understanding the formation of political coalitions, determining optimal pricing strategies in competitive markets, assessing the power dynamics of voters or voter blocs, and optimizing decisions related to manufacturing plant locations. [10]

The versatility of Game Theory is evident in its application to challenges ranging from legal disputes about voting systems to the optimal placement of manufacturing plants. It has been instrumental in shedding light on the dynamics of strategic interactions in various contexts, offering valuable insights into decision-making processes influenced by complex interdependencies. [10]

2.5 NE-Seeking

At the core of Game Theory lies the concept of Nash Equilibrium (NE), a pivotal outcome in noncooperative games for multiple players. Coined after the eminent American mathematician John Nash, the NE represents a state where no player can enhance their expected outcome by unilaterally altering their strategy. This foundational idea serves as a cornerstone in Game Theory, especially in N-player noncooperative games, earning Nash the 1994 Nobel Prize in Economics for his ground-breaking contributions [11] [12] [13].

Crucial to understanding NE-seeking is the classification of games as noncooperative, meaning players lack mechanisms for binding agreements. A classic example is the prisoner's dilemma, where two accused individuals face the dilemma of confessing or remaining silent without any enforceable agreement. The absence of external enforcement renders the game noncooperative, emphasizing the strategic nature of decisions where betrayal incurs no penalty. [14]

Understanding when and where this state occurs, along with predicting player payoffs at that point, is crucial in competitive Game Theory. While algorithms like fictitious play and gradient play have been devised to find the NE, they are limited by informational requirements.

RL algorithms, explored in detail, have gained prominence due to their applicability in games with limited information, where traditional algorithms fall short.

2.6 Passivity-Based Controls for NE Seeking

The scrutiny of Reinforcement Learning (RL) within the realm of passivity-based control theory unfolds a series of substantial contributions and breakthroughs, elucidating intricate facets of convergence in multiagent finite games. The cardinal contributions articulated in the study can be expounded upon to unveil a more nuanced understanding of the advancements in RL.

The utilization of passivity-based control theory is pivotal in establishing the convergence of an existing RL scheme. The study illuminates that this convergence is not limited to potential games but instead encompasses a broader spectrum, specifically extending to N-player monotone and hypomonotone games. The significance of this extension, lies in the broadened applicability of RL in diverse game scenarios, transcending previous constraints and providing a more encompassing solution for convergence challenges. [1]

Expanding the horizon of passivity-based control, this research introduces the concept of higher-order learning dynamics, ushering in a paradigm shift in the design of RL extensions. By delving into the integration of higher-order dynamics through auxiliary states, this study underscores the potential advantages in fostering convergence for expansive classes of games, effectively surmounting the limitations posed by conventional first-order schemes. [1]

As depicted in Figure 1, there are limitations in convergence with passivity-based approaches that still need to be addressed, specifically with monotone games. The Anti-coordination is a benchmark strategic game which cannot be resolved via traditional passivity-based approaches and requires further research to identify a solution.

This study further extends its purview to discrete-time reinforcement learning, particularly focusing on a scheme with noisy updates grounded in the stochastic-approximation method. The revelations encapsulated in Theorem 3 underscore the adaptability and convergence capabilities of passivity-based control in discrete-time settings. This extension enhances the

practicality and versatility of RL algorithms, catering to real-world scenarios where continuous-time implementations may be impractical, thereby broadening the scope of RL applications. [1]

In essence, this deep dive into passivity-based control theory within the ambit of RL not only broadens the scope of convergence but also pioneers a framework for the design of higher-order learning dynamics. These developments offer a sophisticated and nuanced comprehension of RL's potential in navigating complex decision spaces involving multiple agents. The robust theoretical foundations, fortified by meticulous proofs and numerical results, contribute substantively to the scientific dialogue, paving the way for more resilient and adaptable RL applications across diverse scenarios and challenges.[1]

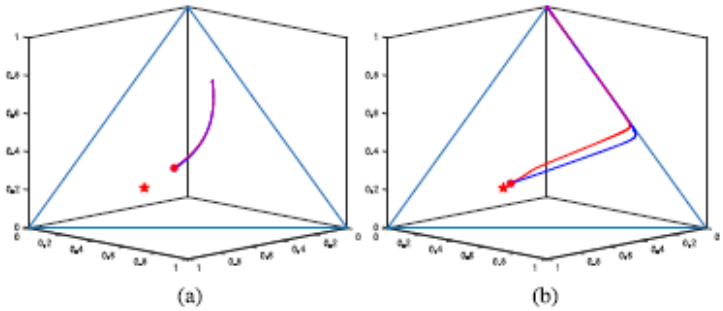


Figure 1. The convergence of the 123 Anticoordination game using the EXP-D-RL approach. (a) *epsilon* = 1. (b) *epsilon* = 0.1. Taken from [1]

2.7 P-RL & Q-Learning in Continuous Time

Understanding the importance of P-RL and Q-Learning methods in continuous time is vital. There are two interconnected models represented by equations (1) and (2) that are vital to understanding these concepts. These models are visualized in Figures 2(a) and 2(b), with the key difference lying in the configuration of the forward path. Figure 2(a) involves a bank of integrators, while Figure 2(b) incorporates a bank of low-pass filters.

Further analysis is carried out to leverage OSEIP payoff the asymptotic stability of the Q-learning closed-loop model (equation 2) shown in Figure 2(b). Gao & Pavel explore this idea in their 2022 paper. The discussion emphasizes the generality of the results, asserting stability for any epsilon greater than zero in any N-player monotone game, where the monotonicity of the negative payoff game mapping (-U) plays a crucial role. Additionally, the passage explores the limitations of the P-RL model in Figure 2(a), where only mere stability is achievable when -U is

monotone, highlighting the importance of passivity techniques to extend convergence results to a broader class of hypomonotone games and designing higher-order Q-learning dynamics. The goal is to strike a balance between the passivity characteristics on the feedback and feedforward paths, ensuring convergence to an approximated Nash Equilibrium for any epsilon greater than zero.

$$P: \begin{cases} \dot{z} = U_i(x) - z, \;\; z(0) \in \mathbb{R}^n \\ x = \sigma_\epsilon(z) \end{cases}$$

Equation 1. Q-Learning Feeback Model

$$P: \begin{cases} \dot{z} = U_i(x), \;\; z(0) \in \mathbb{R}^n \\ x = \sigma_\epsilon(z) \end{cases}$$

Equation 2. P-RL Feeback Model

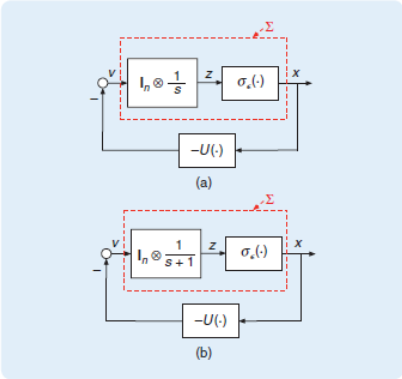


Figure 2. a) Payoff-based reinforcement learning (P-RL) and (b) Q-learning, represented as a feedback interconnected system (R,*U*), where *U* on the feedback path is the payoff game mapping. On the forward path, R is the composition between (a) a bank of integrators (P-RL) or (b) a bank of low-pass filters (Q-learning) and the soft-max mapping. Taken from [2]

2.8 Heavy Anchor Dynamics

Gao and Pavel's 2022 paper introduces an innovative solution, "Heavy Anchor," which stands out as a passivity-based modification of the conventional gradient-play dynamics. The primary objective of Heavy Anchor is to overcome the strict monotonicity constraints of the pseudo-gradient, which is imperative for gradient-play dynamics. This paper rigorously proves that Heavy Anchor achieves not only a relaxation of strict monotonicity but also ensures exact asymptotic convergence in merely monotone regimes, a pivotal contribution that extends the reach of convergence results beyond conventional boundaries. [3]

The study takes a bold step forward by extending the applicability of Heavy Anchor to scenarios where players possess only partial information about their opponents' decisions. In this setting, each player maintains a local decision variable and an auxiliary state estimate, fostering a decentralized learning approach. The modification of Heavy Anchor through distributed

Laplacian feedback becomes instrumental in leveraging equilibrium-independent passivity properties to attain convergence to a Nash Equilibrium, particularly in hypomonotone regimes. These findings mark a significant leap in the literature. Figure 3 and Figure 4 show the integration of Heavy Anchor in a monotone gradient field. [3]

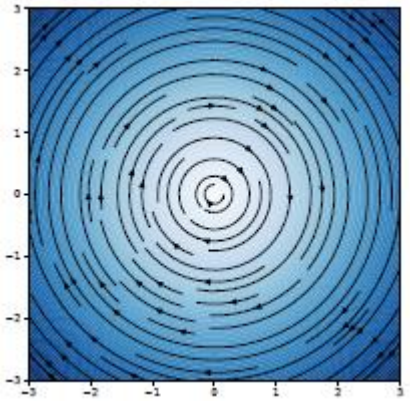
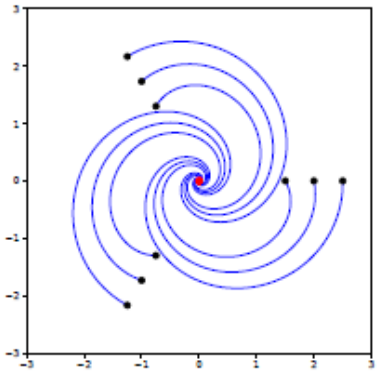


Figure 3. Gradient Vector Field. Taken from [3] Figure 4.



Decision trajectories under Heavy Anchor. Taken from [3]

The contributions of this paper are multifaceted. First and foremost, it introduces and rigorously analyzes the Heavy Anchor algorithm, showcasing its prowess in ensuring exact convergence to a Nash Equilibrium for positive parameter values in the full-decision information setting. Moreover, the extension of Heavy Anchor to hypomonotone games, under specified conditions, underscores its adaptability to more complex scenarios. In the partial-decision information setting, this paper achieves a notable milestone by proving convergence for monotone extended pseudo-gradient or hypomonotone and inverse Lipschitz pseudo-gradient, filling a significant gap in the existing literature. The exploration of Heavy Anchor's relationship to optimization dynamics, its similarity to approaches used in chaotic systems stabilization, and its connections to second-order dynamics in the optimization realm further enrich the scientific discourse. Overall, Heavy Anchor emerges as a versatile and powerful methodology, offering novel insights and solutions to long-standing challenges in distributed Nash Equilibrium seeking.[3]

2.9 Q-Learning with Side Information

This paper introduces a discrete-time Nash Equilibrium-seeking reinforcement learning scheme designed to exploit side information, ultimately achieving convergence in a specific class of finite games characterized by negative monotonicity properties in their utility. Notably, the

literature review emphasizes the limited research on reinforcement learning that effectively utilizes side information despite its potential practical applications. In response to this gap, the study explores various Q-Learning techniques, including FLQL, IQL, and QLSI 1-3, all derived from central equations (Figure 5) but featuring distinct simplifications based on initial conditions and assumptions that converge at different rates as depicted in Figure 6. [4]

$$
\begin{aligned}
z_{i_L}^{k+1} &= z_{i_L}^{k} + \gamma_i^k \left(U_i^L(e_{-i_L}^k) - z_{i_L}^k\right) \\
z_{i_G}^{k+1} &= z_{i_G}^{k} + \gamma_i^k \mathrm{diag}(\chi_i^k)^{-1} \mathrm{diag}\left(\Pi_{i_G}^k - z_{i_G}^k\right) S_i e_i^k \\
z_i^k &= z_{i_L}^k + z_{i_G}^k \\
x_i^k &= \sigma_i(z_i^k) \\
\chi_i^k &= S_i x_i^k
\end{aligned}
$$

Figure 5. Q-learning with Side Information (QLSI) updating functions. Taken from [4]

Through numerical simulations of representative games, the paper showcases that exploiting more side information leads to a faster convergence rate to a Nash Equilibrium, highlighting the practical significance of incorporating side information in reinforcement learning for finite games. [4]

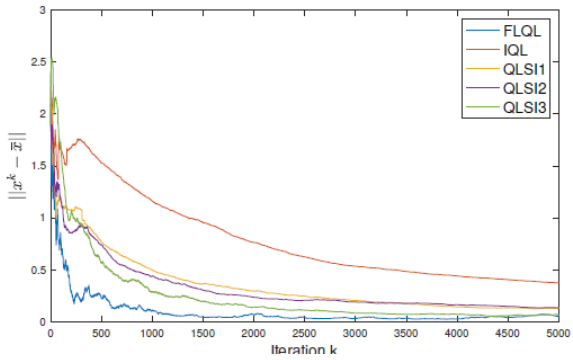


Figure 6. Distance from the Nash Equilibrium as a function of the iteration obtained with different algorithms for the standard RPS game. Taken from [4]

3. Methods

The methodology for integrating Q-Learning and Payoff-based (P-RL) into a unified model involves three main phases: Assessment, Game, and Choice. Each phase transitions between three key vectors: x, z, and π, representing different aspects of the learning process.

Figure 7. Algorithm map of Q-learning and P-RL

The Choice phase utilizes the SoftMax function applied to the z vector to generate the mixed strategy vector. In the Game phase, the dot product between the payoff matrix A and the vector x is calculated. The Assessment phase, in continuous time, involves passing through a set of high-pass filters. However, in discrete-time, the absence of high-pass filters necessitates adaptation.

$$(a)\ \sigma_i(z_i) \coloneqq \left(\frac{1}{\Sigma_{j\in A_j}\exp\left(\frac{1}{\epsilon}\right)z_{ij}}\right)\times\left[\exp\left(\left(\frac{1}{\epsilon}\right)z_{i1}\right)\ldots\exp\left(\left(\frac{1}{\epsilon}\right)z_{in}\right)\right]^T = x_i$$

$$(b)\ 2.x_i^T\cdot A = \pi_i$$

Equation 3. (a) SoftMax function on vector z. (b) Game Phase on vector x.

To address this, equations from Gao and Pavel's paper [2] are employed to update the vectors in the Assessment phase for both Q-Learning and P-RL models. In transitioning to discrete-time, adjustments are made to incorporate a time step, k, and the introduction of the alpha function, defined as $\frac{1}{k+1}$.

$$(a)\ z_i(k+1) = z_{i(k)} + \alpha_i(k)\pi_i(k)diag\left(\frac{1}{x_i(k)}\right)$$

$$(b)\ z_i(k+1) = z_{i(k)} + \alpha_i(k)diag\left(\frac{\pi_i(k) - z_i(k)}{x_i(k)}\right)$$

Equation 4. Assessment Phase update function in discrete-time taken from [2] for (a) P-RL model (b) Q-Learning model

Additionally, the Heavy Anchor Dynamics, introduced in continuous time with equations from Gao and Pavel's paper [2], are modified for discrete-time. This involves introducing a new auxiliary variable, such as beta, alpha, and a vector r, updated iteratively.

$$P: \begin{cases} \dot{x} = \text{proj}_{T\Omega(x)}(-F(x) - \beta(x - r)) \\ \dot{r} = \alpha(x - r), \end{cases},$$

Equation 5. Heavy Anchor Dynamics model

The equations are further modified to integrate the P-RL model into the Assessment phase alongside the Heavy Anchor Dynamics. Notably, variable names are changed to z to allow direct interaction between the heavy anchor layer and the Assessment phase. The projection phase is eliminated, and the r component is integrated into the revised equations.

$$P: \begin{cases} z_i(k+1) = \left(-(x_i(k) \cdot A) - \beta\big(z_i(k) - r_i(k)\big) + \pi_i(k) diag\left(\frac{1}{x_i(k)}\right) \right) \alpha(k) + z(k) \\ r_i(k+1) = (\alpha\big(x_i(k) - r_i(k)\big))\alpha(k) + r(k) \\ \sigma\big(z_i(k)\big) = x_i(k) \end{cases}$$

Equation 6. Integration of Heavy Anchor Dynamics with P-RL model in discrete-time

This final proposed method aims to effectively combine P-RL with Heavy Anchor Dynamics, leveraging Q-side learning for enhanced performance in reinforcement learning tasks. By adapting the methodology to discrete-time while maintaining the key components of the original model, it ensures applicability to real-world scenarios where discrete-time processing is often more practical and efficient, thus providing a comprehensive solution for complex reinforcement learning problems.

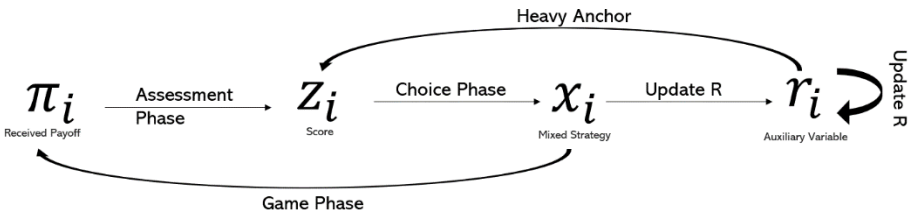


Figure 8. Algorithm map of P-RL with Heavy Anchor Dynamics

4. Results and Discussion

4.1 Game 1 RPS

We first examine the standard game of rock-paper-scissors. The payoff matrix for this game is denoted as:

$$A = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix}$$

To ensure a zero-sum game, $A^T = B$ is used for this setup. We denote B, to be the payoff matrix used by the second player in the game for all games, it can be seen in equation [4] that A is used however this matrix changes depending on which player is being updated. The Nash Equilibrium of this game occurs at the mixed strategy $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, which is considered the optimal strategy.

Here we can observe the convergence of the Q-Learning models in Gao and Pavel's paper compared to the discrete-time Q-learning model. The Nash Equilibrium in the Q-Learning model converges to the correct one.

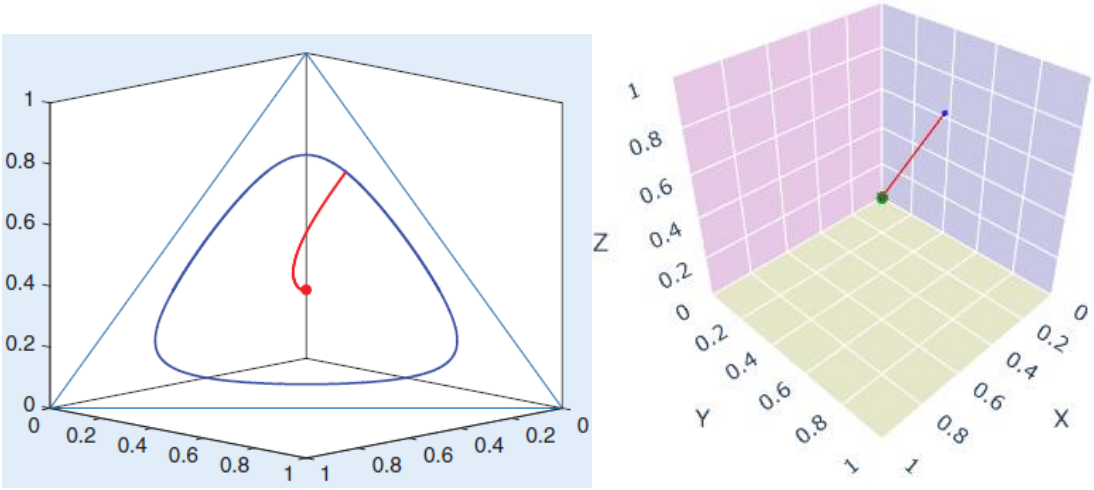


Figure 9. Mixed Strategy Trajectory for a simple RPS Game. (left) taken from [2] with blue line being P-RL and Q-Learning in red line (right) Q-Learning replicated in discrete-time

However, upon evaluating the models with P-RL and P-RL with Heavy Anchor Dynamics models, we notice that the P-RL game does not converge, while the P-RL with Heavy Anchor Dynamics does. The trajectory for the P-RL model resembles the one presented in Gao and Pavel's paper [2], indicating correct implementation.

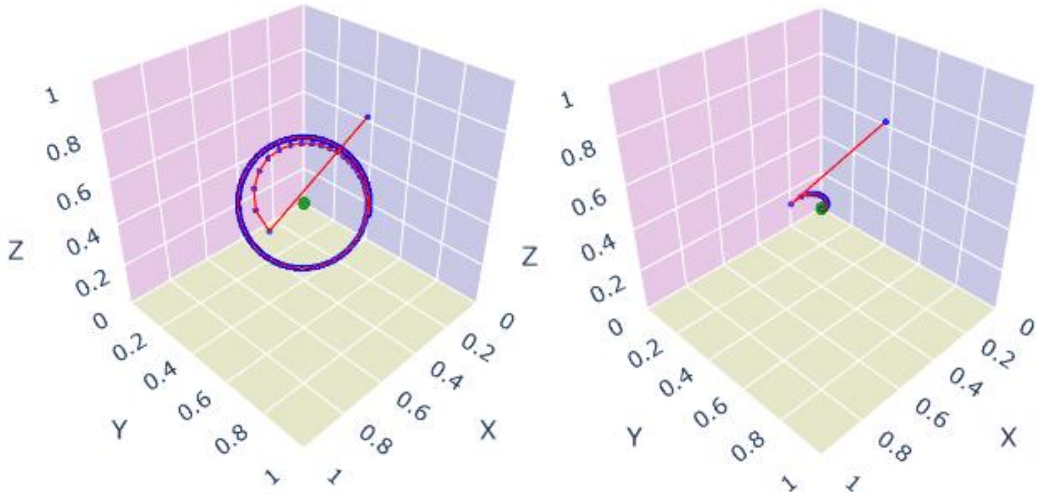


Figure 10. Mixed Strategy Trajectory for The RPS Game. (left) P-RL model (right) P-RL with Heavy Anchor Dynamics model both in discrete-time

The parameters used for P-RL with Heavy Anchor Dynamics include β=1, α=1, and ϵ=1. These parameters enable convergence in discrete-time but not in continuous time, a distinction that must be considered. Each dot on the graph represents a plot in a timestep; there are a lesser amount of time steps to convergence on the Q-Learning model than the P-RL with Heavy Anchor Dynamics model. However, this may not always be true for all games.

4.2 Game 2 Anti-coordination

We now examine the Anti-coordination game. The payoff matrix for this game is denoted as:

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix}$$

This game is not zero-sum, and the Q-Learning algorithm converges consistently to a logit Nash Equilibrium at $(0.40, 0.32, 0.27)$. The payoff matrix for player B is, $A^T = B$ and the true

Nash Equilibrium of this game occurs at the mixed strategy $(\frac{6}{11}, \frac{3}{11}, \frac{2}{11})$, which is considered the optimal strategy.

Here, we can observe the convergence of the Q-Learning models in Gao et al.'s paper [1] compared to the discrete-time Q-learning model. The Nash Equilibrium in the Q-Learning model converges to the incorrect Nash Equilibrium as expected.

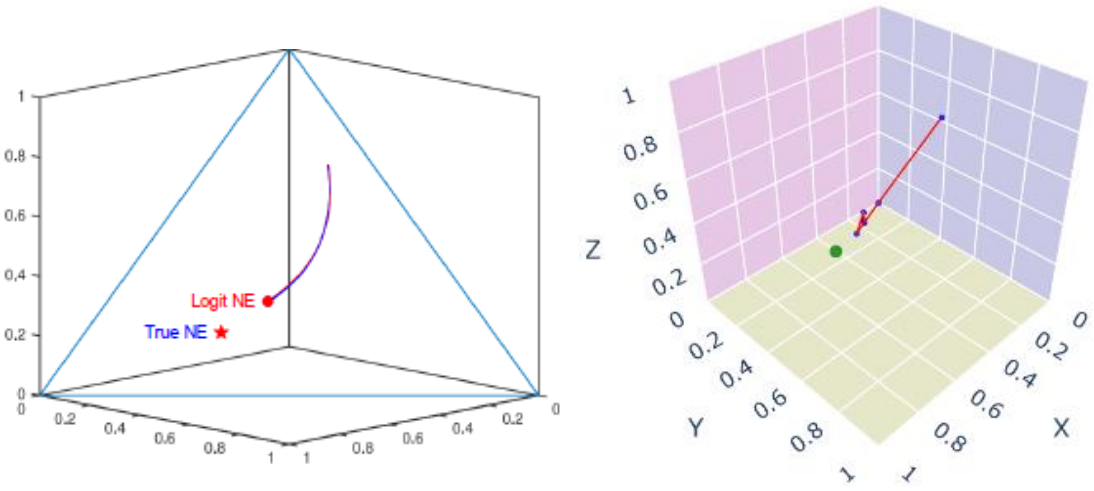


Figure 11. Mixed Strategy Trajectory for the Anti-Coordination Game. (left) Taken from Gao et. al.'s paper, with the blue line being EXP-D-RL and the red line being H-EXP-D-RL. (right) Plot recreated in discrete-time

However, upon evaluating the models with P-RL and P-RL with Heavy Anchor Dynamics models, we notice that the P-RL game does not converge, while the P-RL with Heavy Anchor Dynamics does.

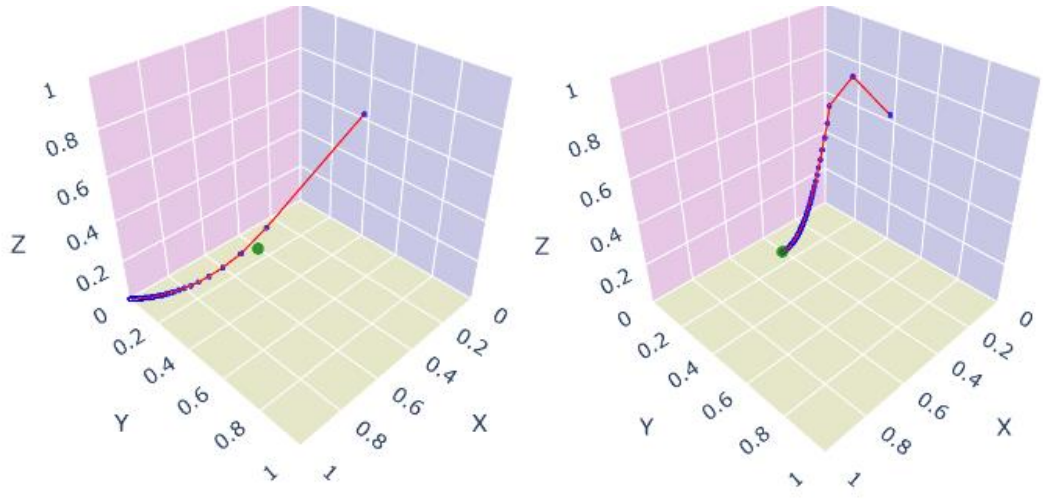


Figure 12. Mixed Strategy Trajectory for The Anti-Coordination Game. (left) P-RL model (right) P-RL with Heavy Anchor Dynamics model both in discrete-time

The parameters used for P-RL with Heavy Anchor Dynamics include $\beta = 2.5, \alpha = 5,$ and $\epsilon = 1.$

4.3 Game 3 Matching Pennies

Next, we examine the Matching Pennies game. The payoff matrix for this game is denoted as:

$$A = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

This game is zero-sum, the payoff matrix for player B is, $A = -B$ and the Nash Equilibrium of this game occurs at the mixed strategy $(\frac{1}{2}, \frac{1}{2})$, which is considered the optimal strategy. Here we can observe the convergence of the Q-Learning models in Gao and Pavel's paper [1] compared to the convergence of the P-RL model and the P-RL with Heavy Anchor Dynamics model. Since the game is zero-sum, the P-RL model does not converge while the P-RL with Heavy Anchor Dynamics does, that too at a very fast rate.

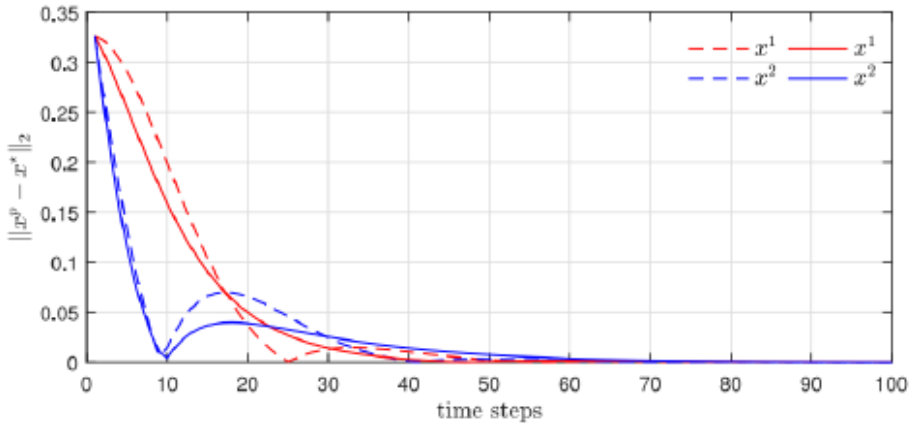


Figure 13. Mixed Strategy Trajectory for 2-player MP game taken from Gao et. al.'s paper, with the dashed lines being EXP-D-RL and the solid lines being H-EXP-D-RL

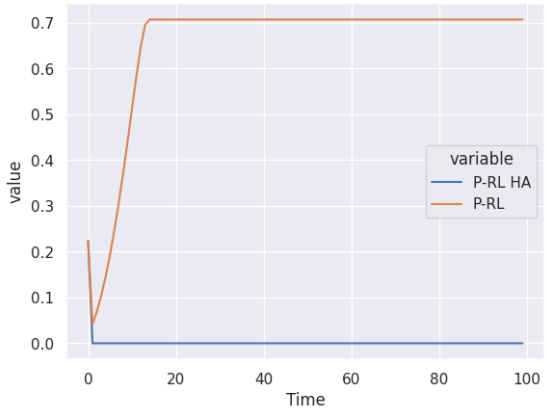


Figure 14. Mixed Strategy Trajectory for 2-player MP Game with P-RL and P-RL with Heavy Anchor Dynamics convergence

4.4 Game 4 Shapley

The Shapley game is very similar to the rock paper scissors game however is not considered a zero-sum game due to the payoff matrix. The payoff matrix for this game is denoted as:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

$A^T = B$ is used for this setup. The Nash Equilibrium of this game occurs at the mixed strategy $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, which is considered the optimal strategy.

Here we can observe the convergence of the Q-Learning models in Gao and Pavel's paper [1] compared to the discrete-time Q-learning model. The Nash Equilibrium in the Q-Learning model converges to the correct one.

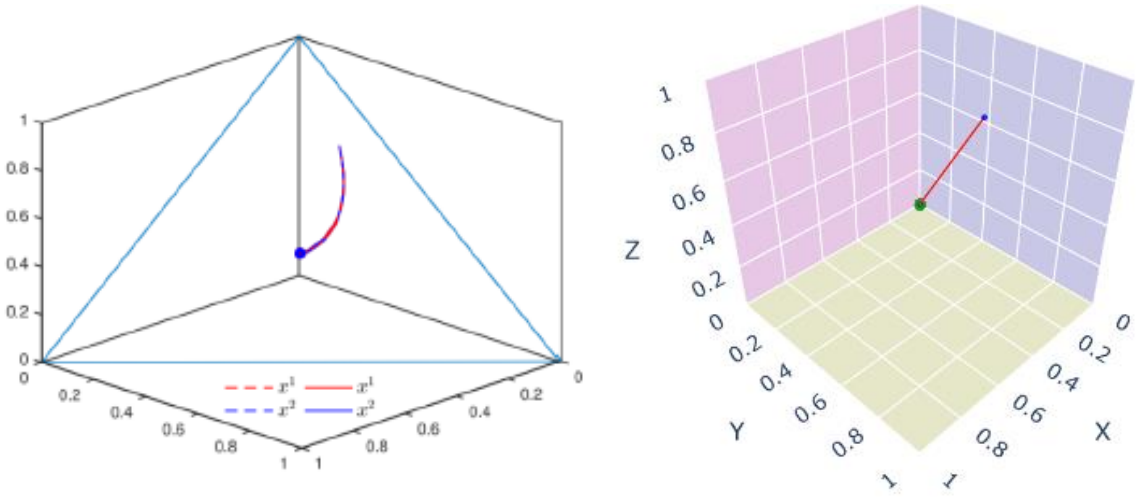


Figure 15. Mixed Strategy Trajectory for The Shapley Game. (left) taken from Gao et. al.'s paper, with the blue line being EXP-D-RL and the red line being H-EXP-D-RL (right) Plot recreated in discrete-time

However, upon evaluating the models with P-RL and P-RL with Heavy Anchor Dynamics models, we notice that the P-RL game does converge at a slower rate than the P-RL with Heavy Anchor Dynamics.

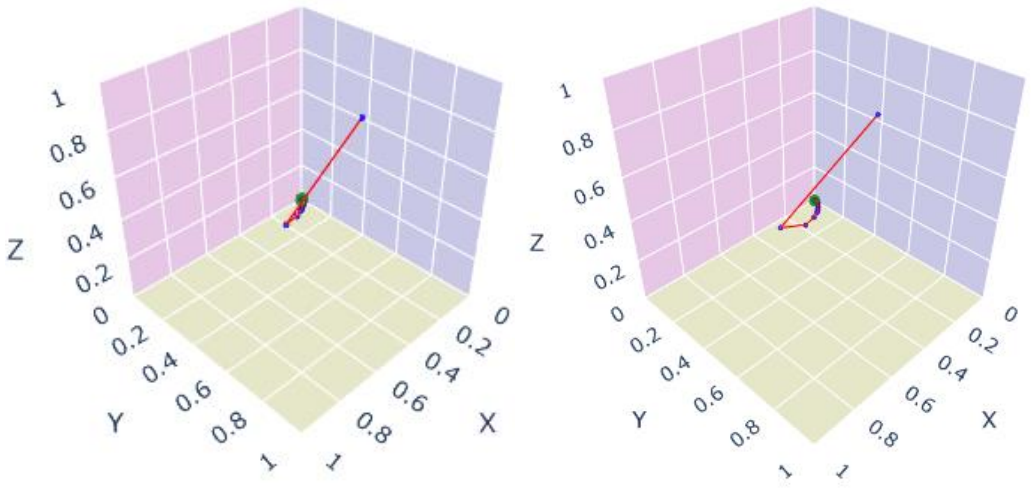


Figure 16. Mixed Strategy Trajectory for The Shapley Game. (left) P-RL model (right) P-RL with Heavy Anchor Dynamics model both in discrete-time

The parameters used for P-RL with Heavy Anchor Dynamics include $\beta = 0.1, \alpha = 2$, and $\epsilon = 1$. These parameters enable convergence in discrete-time but not in continuous time.

4.5 Game 5 Modified RPS

Finally, we examine a modified version of the standard game of rock-paper-scissors. The payoff matrix for this game is denoted as:

$$A = \begin{bmatrix} 0 & -1 & 3 \\ 2 & 0 & -1 \\ -1 & 3 & 0 \end{bmatrix}$$

This game is considered unstable due to the positive values in the payoff matrix which are greater than 1 and inconsistent. This makes this strategically different from the regular rock, paper, scissors game. The Nash Equilibrium of this game occurs at the mixed strategy (0.40625,0.3125,0.28125). which is considered the optimal strategy.

Here, we can observe the convergence of the Q-Learning models in Gao and Pavel's paper compared to the discrete-time Q-learning model. The Nash Equilibrium in the Q-Learning model converges to the incorrect Nash Equilibrium as expected.

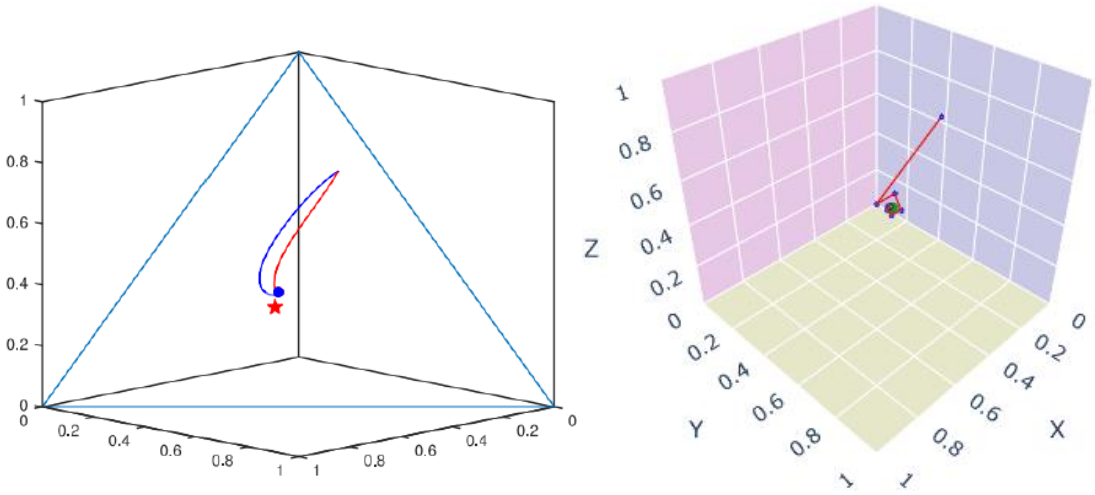


Figure 17. Mixed Strategy Trajectory for a Modified RPS Game. (left) taken from Gao et. al.'s paper, with the blue line being EXP-D-RL and the red line being H-EXP-D-RL (right) Plot recreated in discrete-time

However, upon evaluating the models with P-RL and P-RL with Heavy Anchor Dynamics models, we notice that the P-RL game does not converge, while the P-RL with Heavy Anchor Dynamics does.

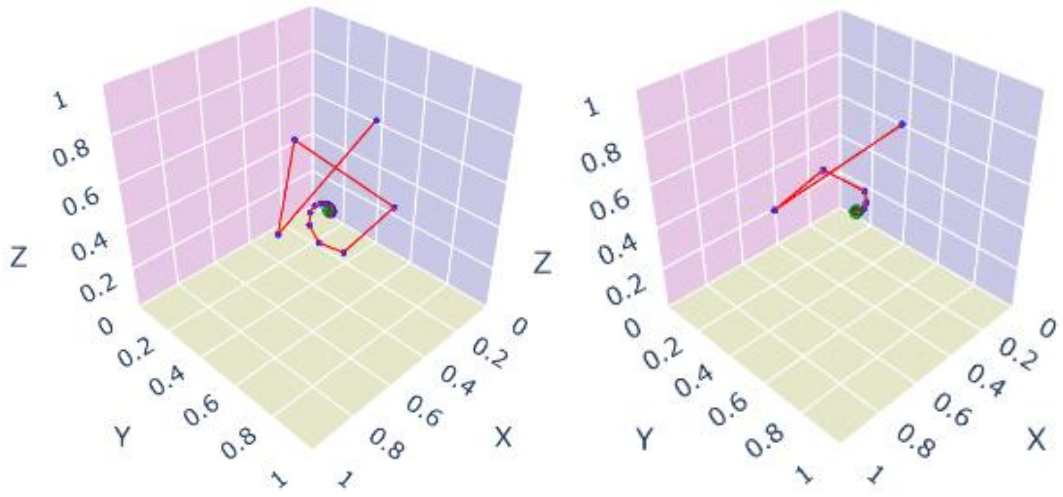


Figure 18. Mixed Strategy Trajectory for a Modified RPS Game. (left) P-RL model (right) P-RL with Heavy Anchor Dynamics model both in discrete-time

The parameters used for P-RL with Heavy Anchor Dynamics include β=0.5, α=0.5, and ϵ=1. These parameters enable convergence in discrete-time but not in continuous time, a distinction that must be considered.

5. Conclusion

This study introduced a modified version of the passivity-based method for identifying the Nash Equilibrium (NE) of a game, building upon the approach presented in a prior work by Gao and Pavel. While Gao and Pavel utilized Q-learning to reach a logit NE, this research adapted the model to employ P-RL instead. Additionally, this study introduced Heavy Anchor Dynamics on the feedback path to facilitate convergence to the true NE, particularly in monotone games. The findings were demonstrated using games outlined in Gao and Pavel's paper. Further research can explore the optimization of hyperparameters tailored to specific games, as the P-RL with Heavy Anchor Dynamics model involves distinct parameters for each game. Integrating this optimization process into the Nash Equilibrium seeking stages could be a valuable extension.

6. References

1. B. Gao and L. Pavel, "On Passivity, Reinforcement Learning, and Higher Order Learning in Multiagent Finite Games," IEEE Transactions on Automatic Control, vol. 66, pp. 121-136, 2021.

2. L. Pavel, "Dissipativity Theory in Game Theory," IEEE Control Systems, pp. 150-164, 2022.

3. D. Gadjov and L. Pavel, "On the exact convergence to Nash Equilibrium in hypomonotone regimes under full and partial-information," Proc. 59th IEEE Conf. Decision Control, pp. 2297-2302, 2020.

4. M. Sylvestre and L. Pavel, "Q-Learning with Side Information in Multi-Agent Finite Games," 2019 IEEE 58th Conference on Decision and Control (CDC), Nice, France, 2019, pp. 5032-5037, doi: 10.1109/CDC40024.2019.9029788.

5. O. Chatain, "Cooperative and Non-Cooperative Game Theory," in The Palgrave Encyclopedia of Strategic Management, 2014.

6. L. P. Kaelbling, M. L. Littman and A. W. Moore, "Reinforcement Learning: A Survey," Journal of Artifcial Intelligence Research, vol. 4, pp. 237-285, 1996.

7. K. Ogata, "System Dynamics," Pearson, 2010.

8. A. V. Oppenheim and R. W. Schafer, "Discrete-Time Signal Processing," Prentice Hall, 1999.

9. J. von Neumann and O. Morgenstern, "The Theory of Games and Economic Behavior," Princeton University Press, 1944.

10. M. Davis and S. Brams, “Game theory,” Encyclopædia Britannica, https://www.britannica.com/science/game-theory

11. J. Nash, "Non-cooperative Games," Annals of Mathematics, vol. 54, no. 2, pp. 286-295, 1951.

12. J. Nash, "The Imbedding Problem for Riemannian Manifolds," Annals of Mathematics, vol. 63, no. 1, pp. 20-63, 1956.

13. J. Nash, "The imbedding problem for differential manifolds," Annals of Mathematics, vol. 63, no. 1, pp. 20-63, 1956.

14. S. Eldridge, “Nash Equilibrium,” Encyclopædia Britannica, https://www.britannica.com/science/Nash-equilibrium.

List of Figures & Tables

Figure 1. The convergence of the 123 Anticoordination game using the EXP-D-RL approach. (a) epsilon = 1. (b) epsilon = 0.1. Taken from [1]. (a) *epsilon* = 1. (b) *epsilon* = 0.1. Taken from [0]…….....5

Figure 2. a) Payoff-based reinforcement learning (P-RL) and (b) Q-learning, represented as a feedback interconnected system (R,*U*), where *U* on the feedback path is the payoff game mapping. On the forward path, R is the composition between (a) a bank of integrators (P-RL) or (b) a bank of low-pass filters (Q-learning) and the soft-max mapping. Taken from [1]…….....6

Figure 3. Decision trajectories under Heavy Anchor. Taken from [2]…….....….7

Figure 4. Gradient Vector Field. Taken from [2]…….…….……...7

Figure 5. Q-learning with Side Information (QLSI) updating functions, Taken from [3]…….…..8

Figure 6. Distance from the Nash Equilibrium as a function of the iteration obtained with different algorithms for the standard RPS game. Taken from [3]……..…...……..9

Figure 7. Algorithm map of Q-learning and P-RL……10

Figure 8. Algorithm map of P-RL with Heavy Anchor Dynamics……11

Figure 9. Mixed Strategy Trajectory for a simple RPS Game. (a) taken from [2] with blue line being P-RL and Q-Learning in red line (b) Q-Learning replicated in discrete-time ……12

Figure 10. Mixed Strategy Trajectory for The RPS Game. (left) P-RL model (right) P-RL with Heavy Anchor Dynamics model both in discrete-time……..13

Figure 11. Mixed Strategy Trajectory for the Anti-Coordination Game. (left) Taken from Gao et. al.'s paper, with the blue line being EXP-D-RL and the red line being H-EXP-D-RL. (right) Plot recreated in discrete-time……...14

Figure 12. Mixed Strategy Trajectory for The Anti-Coordination Game. (left) P-RL model (right) P-RL with Heavy Anchor Dynamics model both in discrete-time……15

Figure 13. Mixed Strategy Trajectory for 2-player MP game taken from Gao et. al.'s paper, with the dashed lines being EXP-D-RL and the solid lines being H-EXP-D-RL……..16

Figure 14. Mixed Strategy Trajectory for 2-player MP Game with P-RL and P-RL with Heavy Anchor Dynamics convergence…….16

Figure 15. Mixed Strategy Trajectory for The Shapley Game. (left) taken from Gao et. al.'s paper, with the blue line being EXP-D-RL and the red line being H-EXP-D-RL (right) Plot recreated in discrete-time……...17

Figure 16. Mixed Strategy Trajectory for The Shapley Game. (left) P-RL model (right) P-RL with Heavy Anchor Dynamics model both in discrete-time……..17

Figure 17. Mixed Strategy Trajectory for a Modified RPS Game. (left) taken from Gao et. al.'s paper, with the blue line being EXP-D-RL and the red line being H-EXP-D-RL (right) Plot recreated in discrete-time……...18

Figure 18. Mixed Strategy Trajectory for a Modified RPS Game. (left) P-RL model (right) P-RL with Heavy Anchor Dynamics model both in discrete-time……..19