The Stochastic Evolutionary Dynamics of Softmax Policy Gradient in Games

Chin-wing Leung
University of Warwick
Coventry, United Kingdom
chin-wing.leung@warwick.ac.uk

Shuyue Hu*
Shanghai Artificial Intelligence Laboratory
Shanghai, China
hushuyue@pjlab.org.cn

Ho-fung Leung†
Independent Researcher
Hong Kong, China
ho-fung.leung@outlook.com

ABSTRACT

The theoretical underpinnings of multi-agent learning have recently attracted much attention. In this paper, we study the learning dynamics of the softmax policy gradient (PG) algorithm in multi-agent environments in the context of evolutionary game theory. We revisit the previous analyses based on mean dynamics and observe that previous models fail to characterize the effect of stochasticity. To this end, we propose a stochastic dynamics model to analyse the learning dynamics of PG under symmetric games. We model the parameter dynamics of the learning agent as a multidimensional Wiener process. Applying the Itô's lemma, we obtain the corresponding policy dynamics for the agent. From that, we study the convergence behaviour of the policy dynamics under the self-play training scheme for learning in games. We work out the sufficient conditions for the stochastic stability of the pure Nash equilibrium strategy, and we evaluate the sufficient conditions for the existence of stationary distribution for strictly stable games. Moreover, we express the dynamics of the parameter distribution with the Fokker-Planck equation. In the experiments, we demonstrate that our stochastic dynamics model always provides a significantly more accurate description of the actual learning dynamics than the mean dynamics model across different games and settings.

KEYWORDS

stochastic dynamics, evolutionary game theory, policy gradient

ACM Reference Format:
Chin-wing Leung, Shuyue Hu, and Ho-fung Leung. 2024. The Stochastic Evolutionary Dynamics of Softmax Policy Gradient in Games. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 14 pages.

1 INTRODUCTION

Multi-agent learning (MAL) has recently received much attention [3, 24, 25], with the theoretical foundation being far from well understood. Notably, the research in evolutionary game theory (EGT) [37, 38] and dynamical systems [36] have provided useful tools, by which one can study various properties, such as evolutionary stable strategy (ESS) and asymptotic stability, in various game environments. For example, in their seminal work, Tuyls *et al.* [42, 43] formally model the dynamics of Q-learning [44] in the 2-player game setting; moreover, the mean dynamics of Q-learning and some other learning algorithms [1, 20, 23] have been shown to be variants of the replicator dynamics in the MAL setting.

Policy Gradient (PG) [45] is a popular technique in reinforcement learning. The PG algorithm works on the policy space,[1] in which the policy is parameterized with a parameter vector. Researchers have developed different variants of PG, such as the A3C [33], DDPG [29] and SAC [11] algorithms.

Recently, Bernasconi *et al.* [2] have succeeded in establishing the equivalence between the mean dynamics of the softmax policy gradient in multi-agent settings and that of the replicator dynamics. While this represents a significant step in the research into the evolutionary dynamics of policy gradient, we note that only the *mean* evolutionary dynamics are considered and studied, as in many of the other related research [12, 39]. The mean dynamics approach effectively assumes that the agents are all learning from a deterministic environment, in which agents learn only based on the expected payoff they receive. However, in real situations, the learning environment is generally stochastic, and agents receive stochastic rewards. This is because the exploration mechanism of the PG algorithm presupposes that the choice of actions for agents is stochastic. In the existence of stochasticity, however, an agent's actual payoff, being stochastic, generally deviates from the expected payoff at every time step. Such discrepancies will accumulate over time, and it can, as we shall show in Section 5, lead to inconsistency between the actual evolutionary dynamics and the theoretical predictions.

To tackle this inadequacy, in this paper, we develop a model that captures stochasticity in the learning dynamics of PG under multi-agent settings. Specifically, we explicitly characterize the random factors during learning under the PG algorithm and model the parameter dynamics of an agent as a multidimensional Wiener process. Hence, we obtain the policy dynamics using the Itô's lemma. From that, we study the convergence behaviour of the policy dynamics under the popular "*self-play*" training scheme for learning in games. Analysis of the policy dynamics shows that the fully mixed Nash equilibrium strategy is never stochastically stable, and this rectifies the conclusions from the previous studies [2], which shows that a fully mixed ESS is asymptotically stable. In addition, we also work out the sufficient conditions for stochastic asymptotic stability of

*Corresponding Author
†Part of the research reported in this paper was done when the third author was with The Chinese University of Hong Kong.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

[1]The terms "policy," "strategy," and "action selection probabilities" are used interchangeably in this paper.

the pure Nash equilibrium strategy. In particular, it requires the strategy to be evolutionarily stable, with an appropriate learning rate and baseline. When the game is strictly stable, we work out sufficient conditions for the existence of the stationary distribution in the policy dynamics. Finally, we express the dynamics of the parameter distribution with the Fokker-Planck equation.

We conduct experiments using six canonical 2-player symmetric games, namely, Prisoner's Dilemma, stag hunt, hawk-dove, standard rock-paper-scissors, good rock-paper-scissors, and bad rock-paper-scissors. Results obtained from agent-based simulations show that the mean dynamics models are unable to depict the actual learning dynamics in stochastic environments, and our stochastic dynamics model always provides a significantly more accurate description of the actual learning dynamics across different games and settings.

2 RELATED RESEARCH

Learning Dynamics. In their seminal work, Tuyls *et al.* [42, 43] formally model the dynamics of Q-learning in 2-player games and develop the selection-mutation model. Based on this model, Kianercy and Galstyan [21] provide a comprehensive characterization of the rest point structure for different 2-player games. Gomes and Kowalczyk [10] develop another formal model for Q-learning with ϵ-greedy exploration in 2-player games. Wunder *et al.* [46] show that when ϵ-greedy exploration is applied, the Q-learning dynamics may exhibit chaotic behaviours. Researchers have shown that the policy dynamics of agents that apply other learning methods can also be characterized by replicator equations, examples including lenient learning [34], infinitesimal gradient ascent [19], regret minimization [23], SARSA [1], *etc.* We refer interested readers to [3] for a more thorough review of this line of research.

Recently, the study in MAL dynamics has also looked into population games. Hu *et al.* [14, 15] model the dynamics of Q-learning agents learning under the concurrent learning protocol. Using mean field theory, the authors capture the population dynamics by a system of three equations. Leung *et al.* [28] further consider the stochastic factors in population games and modelled the dynamics of Q-learning agents learning under the generalized social learning protocol. Leonardos and Piliouras [27] focus on the exploration-exploitation trade-off problem in Q-learning in weighted potential games. Using catastrophe theory, they show that the change in exploration hyperparameters can lead to phase transitions of a system, where the number and stability of equilibria change radically.

Learning Dynamics of Policy Gradient. Policy Gradient (PG) [45] is one of the most popular techniques in reinforcement learning. There have been multiple attempts to study the evolutionary dynamics of PG in multi-agent games. In particular, Srinivasan *et al.* [39] present the mean dynamics (all actions updates) of PG under the 2-players symmetric games. Focusing on the mean dynamics, the authors show that for two players in zero-sum games, the regret of an agent is bounded linearly in the order of the number of iterations. They also compare the mean dynamics with the replicator dynamics in certain games. Bernasconi *et al.* [2] establish the formal connections between softmax PG mean dynamics and the replicator dynamics and study various system behaviours in the context of EGT. They show that under the PG mean dynamics, the asymptotic convergence to the best response is proved when playing with a fixed agent. For general symmetric games, the asymptotic stability for the ESS is proved. Omidshafiei *et al.* [12] obtain insights from the PG dynamics and propose the NeuRD algorithm. They also show that the replicator dynamics belongs to the Hedge algorithm [8, 30].

Efforts have also been made to study the convergence behaviours for the gradient-based algorithms in multi-agent environments based on asymptotic analysis. Daskalakis *et al.* [5] show that the time average expected reward converges to the minimax value of the game in 2-player zero-sum games under the REINFORCE [45] algorithm. Leonardos *et al.* [26] focus on the multi-agent coordination scenario, in which convergence is proved for stochastic PG under the Markov Potential Games (MPG). Zhang *et al.* [48] prove the local convergence property of strict Nash equilibrium in stochastic games learning under the exact gradient play algorithm.

Our work studies the stochastic extension of previous works in mean dynamics for softmax policy gradient (PG) [2, 39], where we model the stochastic learning dynamics for PG under softmax exploration as a multidimensional Wiener process.

Stochastics Replicator Dynamics. The concept of Evolutionary Game Theory (EGT) originated from Smith and Price [37, 38] who developed the concept of evolutionary stable strategy (ESS) to study the evolution of a population of animals under natural selection. Later, Taylor and Jonker [41] proposed the replicator dynamics to model the evolution of the population strategy under 2-player symmetric games with random matching. Foster and Young [7] consider the stochastic nature of natural selection and model the replicator dynamics based on a stochastic differential equation. They have shown that the asymptotic behaviour of the population can be affected by the existence of randomness. Since then, research has been conducted to study the characteristics and behaviours of the population strategy under various stochastic replicator dynamics. Fudenberg and Harris [9] and Imhof [16] study the population dynamics where the payoffs of agents are subjected to aggregate shocks. Under such cases, the dynamics of population size for different strategies are modelled as a Wiener process with independent random factors, where the converging behaviour such as stochastic stability and extinction of dominated strategies are studied. Hofbauer and Imhof [13] further analyse the limiting behaviour of the time averages of the above process. Mertikopoulos and Moustakas [32] study the policy dynamics where players learn under the exponential learning scheme, where the payoffs are subjected to independent random perturbations.

Although research in EGT has analysed the stochastic evolutionary dynamics in various situations, there has not been research which studies the stochastic effects arising from the exploration behaviour of the RL agents.

3 PRELIMINARIES

Throughout this paper, we denote a column vector $\boldsymbol{x} \in \mathfrak{R}^d$ as $(x_1, ..., x_d)$.

3.1 Symmetric Games and Replicator Dynamics

Conventionally, a 2-player-d-action normal-form game involves a row player and a column player, each of which has a set of d available actions to choose from. During game plays, the two players simultaneously choose an action and receive an immediate payoff

(or reward) based on their joint actions. Formally, a 2-player-d-action game can be represented by two payoff matrices ($\boldsymbol{A}_{\text{row}}$ and $\boldsymbol{A}_{\text{column}}$) as

$$\boldsymbol{A}_{\text{row}} = \begin{bmatrix} A_{11} & \cdots & A_{1d} \\ \vdots & \ddots & \vdots \\ A_{d1} & \cdots & A_{dd} \end{bmatrix}, \boldsymbol{A}_{\text{column}} = \begin{bmatrix} B_{11} & \cdots & B_{1d} \\ \vdots & \ddots & \vdots \\ B_{d1} & \cdots & B_{dd} \end{bmatrix}$$

where each element represents the immediate payoff of the row or column player given joint action choices. The game is symmetric if (i) both the players have the same set of available actions, and (ii) the resulting payoffs depend *not* on the roles of the players, but only on their joint action choices, i.e. $\boldsymbol{A}_{\text{row}} = \boldsymbol{A}_{\text{column}}^{\top}$. Define $\boldsymbol{A} \coloneqq \boldsymbol{A}_{\text{row}} = \boldsymbol{A}_{\text{column}}^{\top}$. For any player, suppose it chooses action a_i and its opponent chooses action a_j, then its immediate payoff is given by

$$r = \boldsymbol{e}_i^{\top} \boldsymbol{A} \boldsymbol{e}_j \tag{1}$$

where $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ represent basis vectors such that $\boldsymbol{e}_i = (0...1...0)$ with the i^{th} element equaling 1.

Consider a population of pure strategy agents evolving through random matching, where the population growth is proportional to the payoff of the agents relative to the population policy $\boldsymbol{\pi}$ governed by the symmetric game ($r = \boldsymbol{e}_i^{\top} \boldsymbol{A} \boldsymbol{\pi}$), the (mean) replicator dynamics [41] is derived as

$$\partial_t \boldsymbol{\pi}_k = \boldsymbol{\pi}(\boldsymbol{e}_k^{\top} \boldsymbol{A} \boldsymbol{\pi} - \boldsymbol{\pi}^{\top} \boldsymbol{A} \boldsymbol{\pi}) dt$$

3.2 Policy Gradient

Softmax Policy gradient (PG) [45] is a reinforcement learning algorithm. It is defined in a Markov decision process (MDP) $\langle S, A, T, R \rangle$, where S is a set of states, A is a set of the available actions, $T : S \times A \times S \to [0, 1]$ is a state transition probability function, and $R : S \times A \to \Re$ is a function of immediate payoff. A learning agent aims to find a policy $\pi(a|s)$ that maximizes the expected sum of discounted rewards J. The PG method searches for the optimal policy by applying stochastic gradient ascent repeatedly over a parameter space $\boldsymbol{y} \in \Re^m$. To be specific, let $\pi(\cdot|s, \boldsymbol{y})$ be the parameterized policy evaluated as

$$\pi(a_k|s, \boldsymbol{y}) = \frac{\exp(f_k(s; \boldsymbol{y}))}{\sum_{l=1}^{d} \exp(f_l(s; \boldsymbol{y}))}$$

where $\boldsymbol{f} = (f_1(s; \boldsymbol{y}), \ldots, f_d(s; \boldsymbol{y})) : S \to \Re^d$ is the parameterized function approximator of the values of the actions in state s. Consider an episode of learning at iteration t, the gradient ascent is applied to the parameter vector $\boldsymbol{y}(t)$ in the direction of the approximated gradient $\nabla_{\boldsymbol{y}} \widehat{J(\boldsymbol{y}(t))}$,

$$\boldsymbol{y}(t+1) = \boldsymbol{y}(t) + \alpha \nabla_{\boldsymbol{y}} \widehat{J(\boldsymbol{y}(t))}$$

where α is the learning rate, and $\widehat{J(\boldsymbol{y}(t))}$ is the episode reward at time t.

3.3 Gradient of PG in Symmetric Games

We consider the case of a symmetric game, where there is only a single state (hence each episode lasts for only 1 step). Let the parameter vector be defined in the full action space (the tabular case) $\boldsymbol{y} = (y_1, ..., y_d) \in \Re^d$. We drop the state notation s, hence we have $f_k(\cdot; \boldsymbol{y}) = y_k$. Let $y_k \coloneqq y_k(t)$ be the parameter value of action a_k at t^{th} iteration of learning. The corresponding agent policy writes

$$\pi_k \coloneqq \pi(a_k|\boldsymbol{y}(t)) = \frac{\exp(y_k)}{\sum_{l=1}^{d} \exp(y_l)} \tag{2}$$

Given that at iteration t, let a_i and a_j be the actions performed by the learning agent and its opponent, let the episode reward be $r \coloneqq r(t) = \widehat{J(\boldsymbol{y}(t))}$, it can be shown [40] that the approximated gradient is $\nabla_{\boldsymbol{y}} \widehat{J(\boldsymbol{y})} = (r - b)(\boldsymbol{e}_i - \boldsymbol{\pi})$, where $b \coloneqq b(t) \in \Re$ is a baseline.[2] The change in parameter vector writes

$$\partial_t \boldsymbol{y} \coloneqq \boldsymbol{y}(t+1) - \boldsymbol{y}(t) = \alpha(\boldsymbol{e}_i^{\top} \boldsymbol{A} \boldsymbol{e}_j - b)(\boldsymbol{e}_i - \boldsymbol{\pi}) \tag{3}$$

4 THE STOCHASTIC DYNAMICS OF POLICY GRADIENT

In this section, we first model the dynamics of parameter $\boldsymbol{y}$ of gradient descent as a stochastic process described by a system of stochastic differential equations (SDE). The policy dynamics are then obtained by applying the Itô's lemma. Based on the derived models, we work out the sufficient conditions for the stochastic stability for the pure strategy Nash equilibria, and the sufficient conditions for the existence of stationary distribution for strictly stable games. Finally, we express the dynamics of parameter distribution by a Fokker-Planck equation (FPE).

4.1 Parameter Dynamics of PG

Let $\boldsymbol{\pi}$ be the policy of the agent with parameter $\boldsymbol{y}$, and $\boldsymbol{\phi} \coloneqq \boldsymbol{\phi}(t)$ be the policy of the opponent. Previous works of Bernasconi *et al.* [2] and others [12, 39] have performed the analysis based on the mean dynamics model which assumes expected gradient update. The parameter dynamics $\partial_t \boldsymbol{y}$ under the mean dynamics model is as follows:

$$\partial_t \boldsymbol{y} = \boldsymbol{\mu} dt \tag{4}$$

where $\boldsymbol{\mu} \coloneqq \boldsymbol{\mu}(\boldsymbol{y}, t) = \mathbb{E}[\partial_t \boldsymbol{y}] = (\mu_1, \ldots, \mu_d)$ is the mean vector, in which

$$\mu_k = \alpha \pi_k (\boldsymbol{e}_k^{\top} \boldsymbol{A} \boldsymbol{\phi} - \boldsymbol{\pi}^{\top} \boldsymbol{A} \boldsymbol{\phi})$$

However, due to the exploration mechanism of PG, the agents' choices of actions a_i and a_j, and thus the parameter dynamics $\partial_t \boldsymbol{y}$ are *not* deterministic. The mean square distance of the parameter dynamics to the expected gradient is evaluated as

$$\mathbb{E}[(\partial_t \boldsymbol{y} - \boldsymbol{\mu})^2] = \Sigma(\boldsymbol{y}, t) = [\sigma_{kl}] \in \Re^{d \times d} \tag{5}$$

in which

$$\begin{aligned}
\sigma_{kk} =& Var(\partial_t y_k) \\
=& \alpha^2 \pi_k [\boldsymbol{e}_k^{\top} \boldsymbol{A} \circ \boldsymbol{A} \boldsymbol{\phi} - 2b \boldsymbol{e}_k^{\top} \boldsymbol{A} \boldsymbol{\phi} + b^2] + \alpha^2 \pi_k^2 [-(\boldsymbol{e}_k^{\top} \boldsymbol{A} \boldsymbol{\phi})^2 \\
& + 2(\boldsymbol{e}_k^{\top} \boldsymbol{A} \boldsymbol{\phi})(\boldsymbol{\pi}^{\top} \boldsymbol{A} \boldsymbol{\phi}) - 2 \boldsymbol{e}_k^{\top} \boldsymbol{A} \circ \boldsymbol{A} \boldsymbol{\phi} - (\boldsymbol{\pi}^{\top} \boldsymbol{A} \boldsymbol{\phi})^2 + \boldsymbol{\pi}^{\top} \boldsymbol{A} \circ \boldsymbol{A} \boldsymbol{\phi} \\
& + b(4 \boldsymbol{e}_k^{\top} \boldsymbol{A} \boldsymbol{\phi} - 2 \boldsymbol{\pi}^{\top} \boldsymbol{A} \boldsymbol{\phi}) - b^2] \\
\sigma_{kl} =& Cov(\partial_t y_k, \partial_t y_l) \\
=& \alpha^2 \pi_k \pi_l [-(\boldsymbol{e}_k^{\top} \boldsymbol{A} \boldsymbol{\phi})(\boldsymbol{e}_l^{\top} \boldsymbol{A} \boldsymbol{\phi}) + (\boldsymbol{e}_k^{\top} \boldsymbol{A} \boldsymbol{\phi})(\boldsymbol{\pi}^{\top} \boldsymbol{A} \boldsymbol{\phi}) + (\boldsymbol{e}_l^{\top} \boldsymbol{A} \boldsymbol{\phi})(\boldsymbol{\pi}^{\top} \boldsymbol{A} \boldsymbol{\phi}) \\
& - \boldsymbol{e}_k^{\top} \boldsymbol{A} \circ \boldsymbol{A} \boldsymbol{\phi} - \boldsymbol{e}_l^{\top} \boldsymbol{A} \circ \boldsymbol{A} \boldsymbol{\phi} - (\boldsymbol{\pi}^{\top} \boldsymbol{A} \boldsymbol{\phi})^2 + \boldsymbol{\pi}^{\top} \boldsymbol{A} \circ \boldsymbol{A} \boldsymbol{\phi} + b(2 \boldsymbol{e}_k^{\top} \boldsymbol{A} \boldsymbol{\phi} \\
& + 2 \boldsymbol{e}_l^{\top} \boldsymbol{A} \boldsymbol{\phi} - 2 \boldsymbol{\pi}^{\top} \boldsymbol{A} \boldsymbol{\phi}) - b^2]
\end{aligned}$$

[2] Note that the baseline does not affect the expected value of the approximated gradient.

where $k \neq l$. Here, $\boldsymbol{X} \circ \boldsymbol{Y}$ represents the element-wise multiplication of matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Considering the random factors as a stochastic perturbation of the dynamical system [22], we model the parameter dynamics as a Wiener process as below:

$$\partial_t \boldsymbol{y} = \boldsymbol{\mu} dt + \sqrt{\Sigma} d\boldsymbol{W}_t, \tag{6}$$

where $\boldsymbol{W}_t := \boldsymbol{W}(t) = (W_1, ..., W_d)$ is the standard d-dimensional Wiener process, and $\Sigma := \Sigma(\boldsymbol{y}, t)$. Note that the model effectively assumes the sequence of parameter vector $\boldsymbol{y}(t)$ is strongly mixed so that the central limit theorem [18] applies.

Note that the main difference between our model and that of the previous model is that the second term of equation (6) is completely missing in the previous model. Consequently, previous studies failed to model the diffusion of the system, which is now captured in our new equation. The discrepancy between the mean dynamics (without diffusion) and the stochastic dynamics (with diffusion) will accumulate over time. As we shall see, experimental results (see Figure 1 and 2) show that such discrepancies indeed accumulate and lead to results inconsistent with the true dynamics. This can even lead to counter-intuitive conclusions when we consider the convergence behaviour of the system (see Theorem 4.5 and 4.8).

4.2 Policy Dynamics of PG

For ease of presentation, we simplify the expression of the covariances in $\partial_t \boldsymbol{y}$ as follows:

$$\sigma_{kk} := \alpha^2(\pi_k \Lambda_k + \pi_k^2 \chi_{kk}) := \alpha^2 \Psi_{kk} \qquad \sigma_{kl} := \alpha^2 \pi_k \pi_l \chi_{kl} := \alpha^2 \Psi_{kl}$$

Applying the Itô's lemma [17], we derive the change in the agent's policy over time (equation (7)) from the process of equation (6). The policy dynamics $\partial_t \boldsymbol{\pi}$ is evaluated as

$$\partial_t \boldsymbol{\pi} = \boldsymbol{\mu}_{\boldsymbol{\pi}} dt + \boldsymbol{G}_{\boldsymbol{\pi}} d\boldsymbol{W}_t := (\boldsymbol{\mu}_{\boldsymbol{\pi}A} + \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{\pi}B}) dt + \boldsymbol{G}_{\boldsymbol{\pi}} d\boldsymbol{W}_t \tag{7}$$

where $\boldsymbol{\mu}_{\boldsymbol{\pi}A} = (\mu_{\pi_1 A}, ..., \mu_{\pi_d A})$, $\boldsymbol{\mu}_{\boldsymbol{\pi}B} = (\mu_{\pi_1 B}, ..., \mu_{\pi_d B})$, and $\Sigma_{\boldsymbol{\pi}} = \boldsymbol{G}_{\boldsymbol{\pi}} \boldsymbol{G}_{\boldsymbol{\pi}}^\top = [\sigma_{\pi_{kl}}]$, in which

$$\mu_{\pi_k A} = \alpha \pi_k [\pi_k (\boldsymbol{e}_k^\top \boldsymbol{A}\boldsymbol{\phi} - \boldsymbol{\pi}^\top \boldsymbol{A}\boldsymbol{\phi}) - \sum_i \pi_i^2 (\boldsymbol{e}_i^\top \boldsymbol{A}\boldsymbol{\phi} - \boldsymbol{\pi}^\top \boldsymbol{A}\boldsymbol{\phi})]$$

$$\begin{aligned}
\mu_{\pi_k B} =& \alpha^2 \pi_k [\Psi_{kk} + 2\boldsymbol{\pi}^\top \Psi \boldsymbol{\pi} - \boldsymbol{\pi}^\top diag(\Psi) - 2\boldsymbol{\pi}^\top (\Psi_{k\cdot})] \\
=& \alpha^2 \pi_k [2 \sum_i \sum_j \pi_i^2 \pi_j^2 \chi_{ij} - \sum_i \pi_i^3 \chi_{ii} - 2\pi_k \sum_i \pi_i^2 \chi_{ki} + \pi_k^2 \chi_{kk} \\
& + \sum_i (2\pi_i^3 - \pi_i^2) \Lambda_i + (\pi_k - 2\pi_k^2) \Lambda_k]
\end{aligned}$$

$$\begin{aligned}
\sigma_{\pi_{kk}} =& \alpha^2 \pi_k^2 [\Psi_{kk} + \boldsymbol{\pi}^\top \Psi \boldsymbol{\pi} - 2\boldsymbol{\pi}^\top (\Psi_{k\cdot})] \\
=& \alpha^2 \pi_k^2 [\sum_i \sum_j \pi_i^2 \pi_j^2 \chi_{ij} - 2\pi_k \sum_i \pi_i^2 \chi_{ki} + \pi_k^2 \chi_{kk} + \sum_i \pi_i^3 \Lambda_i \\
& + (\pi_k - 2\pi_k^2) \Lambda_k]
\end{aligned}$$

$$\begin{aligned}
\sigma_{\pi_{kl}} =& \alpha^2 \pi_k \pi_l [\Psi_{kl} + \boldsymbol{\pi}^\top \Psi \boldsymbol{\pi} - \boldsymbol{\pi}^\top (\Psi_{k\cdot}) - \boldsymbol{\pi}^\top (\Psi_{l\cdot})] \\
=& \alpha^2 \pi_k \pi_l [\sum_i \sum_j \pi_i^2 \pi_j^2 \chi_{ij} - \pi_k \sum_i \pi_i^2 \chi_{ki} - \pi_l \sum_i \pi_i^2 \chi_{li} \\
& + \pi_k \pi_l \chi_{kl} + \sum_i \pi_i^3 \Lambda_i - \pi_k^2 \Lambda_k - \pi_l^2 \Lambda_l]
\end{aligned}$$

where $\Psi = [\Psi_{kl}] \in \mathfrak{R}^{d \times d}$, $diag(\Psi) = (\Psi_{11}, \ldots, \Psi_{dd})$, $(\Psi_{k\cdot}) = (\Psi_{k1}, \ldots, \Psi_{kd})$ and $k \neq l$.

Looking at the policy dynamics, we see that the random factor Σ in the parameter dynamics $\partial_t \boldsymbol{y}$ contributes to the drift term $\boldsymbol{\mu}_{\boldsymbol{\pi}} dt$ in policy dynamics $\partial_t \boldsymbol{\pi}$ through $\boldsymbol{\mu}_{\boldsymbol{\pi}B}$. If we ignore the randomness in $\partial_t \boldsymbol{y}$, the drift term of $\partial_t \boldsymbol{\pi}$ will be (incorrectly) equal to $\boldsymbol{\mu}_{\boldsymbol{\pi}A}$ only, which is the result calculated in [2] under the mean dynamics, and that is equivalent to the replicator dynamics associate with a non-linear transformation to the original game. Therefore, the stochastic policy dynamics are far more complicated than the mean policy dynamics, and previous analysis on the asymptotic behaviour of the learning would have to be re-evaluated.

4.3 Stochastic Stability Analysis of PG under Self-Play

Learning by self-play is a popular training scheme for multi-agent games. By self-play, the agent plays against a copy of itself during training, which corresponds to $\boldsymbol{\phi} = \boldsymbol{\pi}$ in the equation (6). In the following, we study the convergence behaviour of a PG agent training under self-play.

Stability is one of the most important characteristics of a dynamical system, which refers to the insensitivity of the stationary points of a system under perturbation. For a stochastic process $\partial_t \boldsymbol{x}(t) = \boldsymbol{\mu}_{\boldsymbol{x}} dt + \boldsymbol{G}_{\boldsymbol{x}} dW_t$, we say that it has a *zero solution* if and only if zero is a stationary point, that is, $\partial_t \boldsymbol{x} \equiv \boldsymbol{0}$ when $\boldsymbol{x} = \boldsymbol{0}$. Hence, stationary points are conventionally synonymous with zero solutions. The definitions for various types of stochastic stability of a zero solution are given as follows [31][Definition 4.2.1].

Definition 4.1. *Let $\boldsymbol{x}(t)$ be a continuous time stochastic process with $\partial_t \boldsymbol{x} = \boldsymbol{\mu}_{\boldsymbol{x}} dt + \boldsymbol{G}_{\boldsymbol{x}} dW_t$ that has a zero solution.*

i. The zero solution is said to be stochastically stable *(SS) if for every pair of $\epsilon \in (0,1)$ and $r > 0$, there exists a $\delta > 0$ such that*

$$Pr\{|\boldsymbol{x}(t)| < r \quad \forall t \geq 0\} \geq 1 - \epsilon$$

whenever $|\boldsymbol{x}(0)| < \delta$.

ii. The zero solution is said to be stochastically asymptotically stable *(SAS) if it is stochastically stable and, moreover, for every $\epsilon \in (0,1)$, there exists a $\delta_0 > 0$ such that*

$$Pr\{\lim_{t \to \infty} \boldsymbol{x}(t) = 0\} \geq 1 - \epsilon$$

whenever $|\boldsymbol{x}(0)| < \delta_0$.

iii. The zero solution is said to be globally stochastically asymptotically stable *(GSAS) if it is stochastically stable and, moreover, for all $\boldsymbol{x}(0) \in \mathfrak{R}^d$*

$$Pr\{\lim_{t \to \infty} \boldsymbol{x}(t) = 0\} = 1$$

The stochastic stabilities (SS, SAS, GSAS) are important properties that define the local convergence behaviour to the zero solution of a stochastic process, which can be shown with the following stochastic Lyapunov theorem [31][Theorems 4.2.2, 4.2.3, 4.2.4].

Lemma 4.2. *Let $\boldsymbol{x}(t)$ be a continuous time stochastic process with $\partial_t \boldsymbol{x} = \boldsymbol{\mu}_{\boldsymbol{x}} dt + \boldsymbol{G}_{\boldsymbol{x}} dW_t$ that has a zero solution. Let $S_h = \{\boldsymbol{x} \in \mathfrak{R}^d : |\boldsymbol{x}| < h\}$. Let there be a positive definite function $V(\boldsymbol{x}) : S_h \mapsto \mathfrak{R}^+$ such that*

$$LV(\boldsymbol{x}) := \sum_{i=1}^d \frac{\partial V}{\partial x_i} \mu_{x_i} + \frac{1}{2} \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2 V}{\partial x_i \partial x_j} [\boldsymbol{G}_{\boldsymbol{x}} \boldsymbol{G}_{\boldsymbol{x}}^\top]_{ij}$$

i. *If $LV(\boldsymbol{x}) \leq 0$ for all $\boldsymbol{x} \in S_h$, then the zero solution of $\boldsymbol{x}(t)$ is stochastically stable (SS).*
ii. *If $LV(\boldsymbol{x}) < 0$ for all $\boldsymbol{x} \in S_h$, then the zero solution of $\boldsymbol{x}(t)$ is stochastically asymptotically stable (SAS).*
iii. *If $LV(\boldsymbol{x}) < 0$ for all $\boldsymbol{x} \in \Re^d$, then the zero solution of $\boldsymbol{x}(t)$ is globally stochastically asymptotically stable (GSAS).*

The above lemma specifies that if one can identify a function $V(x)$ that is decreasing in $x \in S_h$, then the stability of $\boldsymbol{x}(t)$ at $\boldsymbol{x} = 0$ will follow. We consider the stochastic stability of the Nash equilibria (NE) of the game. Due to the exploration behaviour of the PG algorithm, the dynamics $\partial_t \boldsymbol{y}$ of parameter $\boldsymbol{y}$ of any non-pure strategy $\boldsymbol{\pi}$ must be nonzero. This is because if $\boldsymbol{\pi}$ is non-pure, then π_k is nonzero for at least an action a_k, hence by equation (5), Σ will be nonzero, and consequently by equation (6) the dynamics $\partial_t \boldsymbol{y}$ is nonzero. We can express this result formally in the following Theorem.

Theorem 4.3. *All non-pure Nash equilibrium strategies are not stochastically stable under PG dynamics.*

Theorem 4.3 above shows that non-pure Nash equilibrium strategies can never be stochastically stable. So we consider the stochastic stability of pure NE strategies, i.e. $\boldsymbol{\pi} \in \{\boldsymbol{e}_1, ..., \boldsymbol{e}_d\}$. We find that the sufficient condition for pure Nash equilibrium strategy to be stochastically stable is that it is an evolutionary stable strategy (ESS) [36] and that the learning rate used in PG must not be too large. The definition of an ESS is as follows:

Definition 4.4. *In symmetric game $\boldsymbol{A}$, policy $\boldsymbol{\pi}^*$ is an* evolutionary stable strategy *(ESS) if there exists a neighbourhood $\mathcal{O}(\boldsymbol{\pi}^*)$ of $\boldsymbol{\pi}^*$ such that*

$$(\boldsymbol{\pi} - \boldsymbol{\pi}^*)^\top \boldsymbol{A}\boldsymbol{\pi} < 0 \quad \forall \boldsymbol{\pi} \in \mathcal{O}(\boldsymbol{\pi}^*) \setminus \boldsymbol{\pi}^*$$

We formally summarize the sufficient conditions of various types of stochastic stability of pure Nash equilibrium strategy in the following theorem.

Theorem 4.5. *Let $\boldsymbol{e}_i$ (action a_i) be an ESS. Define α_0 as*

$$\begin{aligned}\alpha_0 &= \frac{2[\pi_{0i}\boldsymbol{e}_i^\top \boldsymbol{A}\boldsymbol{\pi}_0 - \pi_{0i}\boldsymbol{\pi}_0^\top \boldsymbol{A}\boldsymbol{\pi}_0 - \boldsymbol{\pi}_0^{2\top} \boldsymbol{A}\boldsymbol{\pi}_0 + (\boldsymbol{\pi}_0^\top \boldsymbol{\pi}_0)\boldsymbol{\pi}_0^\top \boldsymbol{A}\boldsymbol{\pi}_0]}{\boldsymbol{\pi}_0^\top diag(\Psi) - \boldsymbol{\pi}_0^\top \Psi \boldsymbol{\pi}_0} \\ &= \frac{\begin{array}{l} 2\sum_{k\neq i}[\pi_{0i}(1-\pi_{0i})\pi_{0k} \\ \quad + \pi_{0k}^2](A_{ii} - A_{ki})] + \sum_{j,k,l\neq i} O(\pi_{0j}\pi_{0k}\pi_{0l}) \end{array}}{\begin{array}{l} \sum_{k\neq i}[\pi_{0d}^3(1-\pi_{0d})\pi_{0k} + \pi_{0k}^2 \\ \quad - 2\pi_{0d}^2\pi_{0k}^2](A_{ki} - b)^2 + \sum_{j,k,l\neq i} O(\pi_{0j}\pi_{0k}\pi_{0l}) \end{array}}\end{aligned} \tag{8}$$

where $\boldsymbol{\pi}_0 = (\pi_{01}, ..., \pi_{0d})$ for $\boldsymbol{\pi}_0 \in \mathcal{O}(\boldsymbol{e}_i)$, and $\boldsymbol{\pi}_0^2 = (\pi_{01}^2, ..., \pi_{0d}^2)$.

i. *If $\alpha \leq \alpha_0$, then the pure strategy of choosing a_i is stochastically stable (SS).*
ii. *If $\alpha < \alpha_0$, then the pure strategy of choosing a_i is stochastically asymptotically stable (SAS).*
iii. *If $\alpha < \alpha_0$ and a_i is a strictly dominant strategy, then the pure strategy of choosing a_i is globally stochastically asymptotically stable (GSAS).*

Therefore, a pure ESS with a small enough learning rate will satisfy the stochastic stability conditions. Looking into equation (8), for the pure strategy a_i to be stochastically stable, the magnitude of the learning rate is restricted by a value α_0 that can be intuitively interpreted as the weighted sum of losses $(A_{ki} - A_{ii})$ of performing alternative actions a_k, $k \neq i$ when the opponent uses action a_i and the baseline is b. A larger value of the learning rate can be used (to facilitate a potentially higher speed of learning, for example) if we can carefully choose a suitable baseline b so that the major term in the denominator is minimized.

We note that in the work of Bernasconi *et al.* [2], the authors show that if only the mean dynamics is considered, a fully mixed strategy is asymptotically stable for PG if it is an ESS for a symmetric game. That is to say, the stability is fully determined by the type of strategy, but it is independent of the learning rate and the baseline. Such a conclusion is counterintuitive and generally flawed. In real-life situations, the learning rate should be low enough to ensure the learning converges, together with an appropriate baseline. This is reflected in the conclusions in Theorem 4.5.

4.4 Existence of Stationary Distribution of PG under Self-Play

We have shown that all non-pure NE strategies are not stochastically stable. However, we can still investigate other convergent behaviours such as the convergence to a stationary (invariant) distribution. The existence of a stationary distribution indicates that the stochastic process will finally stabilize to some distribution, which can be proven with the following theorem [4][Theorem 5.31].

Lemma 4.6. *Let $\boldsymbol{x}(t)$ be a continuous time stochastic process with $\partial_t \boldsymbol{x}(t) = \boldsymbol{\mu}_{\boldsymbol{x}} dt + \boldsymbol{G}_{\boldsymbol{x}} dW_t$, $\Sigma = [\sigma_{ij}] = \boldsymbol{G}_{\boldsymbol{x}} \boldsymbol{G}_{\boldsymbol{x}}^\top$, assume that there exist bounded domains $D \subset \Re^d$ with a smooth boundary, such that*

i *for a suitable $M > 0$, $\sum_{ij=1}^d \sigma_{ij}\xi_i\xi_j \geq M|\boldsymbol{\xi}|^2$ for all $\boldsymbol{x} \in D$, $\boldsymbol{\xi} \in \Re^d$;*
ii *there exist a non-negative function V such that $inf_{|\boldsymbol{x}|>R} V(\boldsymbol{x}) \to \infty$ as $R \to \infty$ and $LV(\boldsymbol{x}) \leq -C$ for all $\boldsymbol{x} \in \Re^d \setminus D$, for a suitable $C > 0$.*

Then there exist an invariant distribution $\tilde{P}$ such that for any function f integrable with respect to $\tilde{P}$:

$$\int p(t, \boldsymbol{x}, d\boldsymbol{y}) f(\boldsymbol{y}) \xrightarrow{t\to\infty} \int \tilde{P}(d\boldsymbol{y}) f(\boldsymbol{y})$$

Our conditions for the existence of stationary distribution require the following concept of stable games [36].

Definition 4.7. *The symmetric game $\boldsymbol{A}$ is a stable game if*

$$(\boldsymbol{\pi} - \boldsymbol{\phi})^\top \boldsymbol{A}(\boldsymbol{\pi} - \boldsymbol{\phi}) \leq 0 \quad \forall \boldsymbol{\pi}, \boldsymbol{\phi} \in [0, 1]$$

If the inequality holds strictly whenever $\boldsymbol{\pi} \neq \boldsymbol{\phi}$, then $\boldsymbol{A}$ is a strictly stable game. If the inequality always binds, then $\boldsymbol{A}$ is a null stable game.

A strictly stable game possesses certain good properties such as having a unique global ESS [36], which is useful in proving the following result. The following theorem summarises the sufficient conditions for the existence of stationary distribution.

Theorem 4.8. *If the game is strictly stable. Define α_0 as*

$$\alpha_0 = \frac{2[\boldsymbol{\pi}_0^\top(\boldsymbol{\pi}_0 - \boldsymbol{\pi}^*)]\boldsymbol{\pi}_0^\top \boldsymbol{A}\boldsymbol{\pi}_0 - 2[\boldsymbol{\pi}_0 \circ (\boldsymbol{\pi}_0 - \boldsymbol{\pi}^*)]^\top \boldsymbol{A}\boldsymbol{\pi}_0}{-\boldsymbol{\pi}_0^\top \Psi \boldsymbol{\pi}_0 + \boldsymbol{\pi}_0^\top diag(\Psi)} \tag{9}$$

where $\boldsymbol{\pi}^*$ is the unique Nash equilibrium of the game and $\boldsymbol{\pi}_0 \in \Delta$, $\Delta = \{\boldsymbol{\pi} \in [0,1]^d : \sum_i \pi_i = 1\}$.

If $\alpha < \alpha_0$, then the policy dynamics will converge to a stationary distribution.

In case of the non-existence of stationary distribution, the policy $\boldsymbol{\pi}$ will diverge to the boundary as $t \to \infty$, therefore we will expect the policy distribution split and the probability mass will be scattered among the pure strategies.

Again, the above result is different from the previous result with mean dynamics approach [2]. Under the mean dynamics model, any fully mixed ESS is asymptotically stable. If we consider the stochastic dynamics model, a fully mixed ESS will possess a stationary distribution, under the condition that the game is strictly stable and the learning rate is small enough.

4.5 Time Evolution of the Probability Density Function

As the learning dynamics of PG are stochastic, we can study the parameter dynamics in its distribution. The Fokker–Planck equation (FPE) [6, 35] is a partial differential equation that describes the dynamics of the probability density function (PDF) of a diffusion process. For the parameter dynamics $\partial_t \boldsymbol{y}$ in (6), the corresponding FPE writes

$$\frac{\partial p(\boldsymbol{y})}{\partial t} = -\sum_{i=1}^{d} \frac{\partial}{\partial y_i}[\mu_i p(\boldsymbol{y})] + \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial^2}{\partial y_i \partial y_j}[\frac{1}{2}\sigma_{ij} p(\boldsymbol{y})] \quad (10)$$

Given the initial parameter distribution $p(\boldsymbol{y}(0))$ of the agent, the parameter distribution over time $p(\boldsymbol{y}(t))$ can be obtained by solving the equations (6) and (10) numerically.

5 EXPERIMENTS

In this section, we apply the model to six different 2-player symmetric games. In Section 5.1, we describe the game configurations and the agent settings. In Section 5.2, we validate our model by comparing the expected value of the agent's strategy with the results obtained from agent-based simulation. We demonstrate that our stochastic dynamics model is far more accurate than the mean dynamics model. In Section 5.3, we discuss the effect of the expected policy in the existence of randomness. We also investigate the agent's convergence behaviour under stable and unstable games.

5.1 Games and Agents Settings

We consider 6 different 2-player symmetric games: Prisoner's Dilemma (PD) game, Stag Hunt (SH) game, Hawk-Dove (HD) game, standard Rock, Paper, Scissors (standard-RPS) game, good Rock, Paper, Scissors (good-RPS) game, and bad Rock, Paper, Scissors (bad-RPS) game. Note that in the standard-RPS, the reward for winning the game is equal to the punishment for losing the game. In the good-RPS, the reward of winning is larger than the punishment of losing. In the bad-RPS, the reward of winning is smaller than the punishment of losing. The payoff bi-matrices of these games are shown in Table 1 and the summary for Nash equilibria is given in Table 2.

For the PD, SH and HD games, we consider the learning dynamics through random matching. Whereas for the standard-RPS, good-RPS, and bad-RPS games, we adopt self-play as the training scheme.

	C	D
C	$2,2$	$0,3$
D	$3,0$	$1,1$

(a) Prisoner's Dilemma

	S	H
S	$3,3$	$0,1.5$
H	$1.5,0$	$1,1$

(b) Stag Hunt

	H	D
H	$-2,-2$	$2,0$
D	$0,2$	$1,1$

(c) Hawk–Dove

	R	P	S
R	$0,0$	$-1,1$	$1,-1$
P	$1,-1$	$0,0$	$-1,1$
S	$-1,1$	$1,-1$	$0,0$

(d) standard Rock, Paper, Scissors

	R	P	S
R	$0,0$	$-1,2$	$2,-1$
P	$2,-1$	$0,0$	$-1,2$
S	$-1,2$	$2,-1$	$0,0$

(e) good Rock, Paper, Scissors

R	P	S
$0,0$	$-2,1$	$1,-2$
$1,-2$	$0,0$	$-2,1$
$-2,1$	$1,-2$	$0,0$

(f) bad Rock, Paper, Scissors

Table 1: Payoff bi-matrices of the games considered in our experiments.

Game	Pure strategy NE	Non-pure strategy NE (β, β)
PD	(D, D)	-
SH	(S, S), (H, H)	$\beta = (S(2/5), H(3/5))$
HD	(H, D), (D, H)	$\beta = (H(1/3), D(2/3))$
standard-RPS, good-RPS, bad-RPS	-	$\beta = (R(1/3), P(1/3), S(1/3))$

Table 2: Nash equilibria of the games considered in our experiments. The non-pure strategy Nash equilibrium (β, β) means that both players take the same strategy β.

The agent-based simulation is the ground truth in our experiment. In the case of random matching, we conduct 100 simulations of a population of $1,000$ agents training over 500 iterations. In the case of self-play, we conduct $1,000$ simulations of self-play agents, within each simulation, an agent is trained over $1,000$ iterations. The average policy at iteration t is evaluated by taking the average from the simulations.

We set the initial parameter value $\boldsymbol{y}(0) = (0,0)$ for PD, SH, HD game, and $\boldsymbol{y}(0) = (1,0,0)$ for 3 RPS games. The learning rate is set as $\alpha = 0.1$. The baseline is set as $b = 0$.

5.2 Stochastic Dynamics versus Mean Dynamics

We compare the expected value of the agent's strategy over time $E[\boldsymbol{\pi}(t)]$ from 3 different methods. Figure 1 and 2 present the results of our experiments. The blue line represents the results obtained with our stochastic dynamics model, the green line plots the results using the mean dynamics approach, and the red scatter line plots the results of the agent-based simulation.

For the stochastic dynamics model, we first obtain the PDF $p(\boldsymbol{y}(t))$ by solving equations (6) and (10) with the finite volume method [47], then we can evaluate the term $E[\boldsymbol{\pi}(t)]$ with (2). For the mean dynamics model, we use the same model as in previous studies [2, 12, 39].

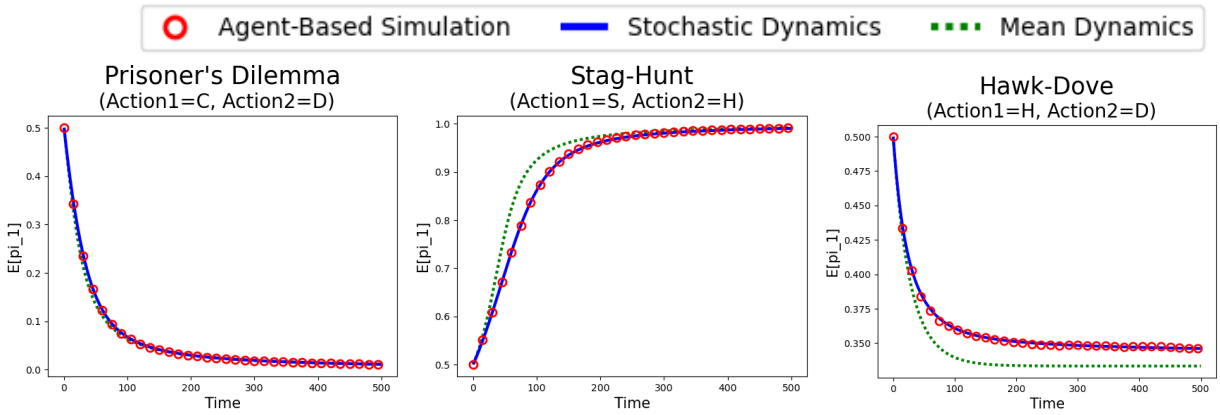

Figure 1: Expected value of policy 1 $E[\pi_1]$ over time, initial parameter value $\boldsymbol{y}(0) = (0, 0)$.

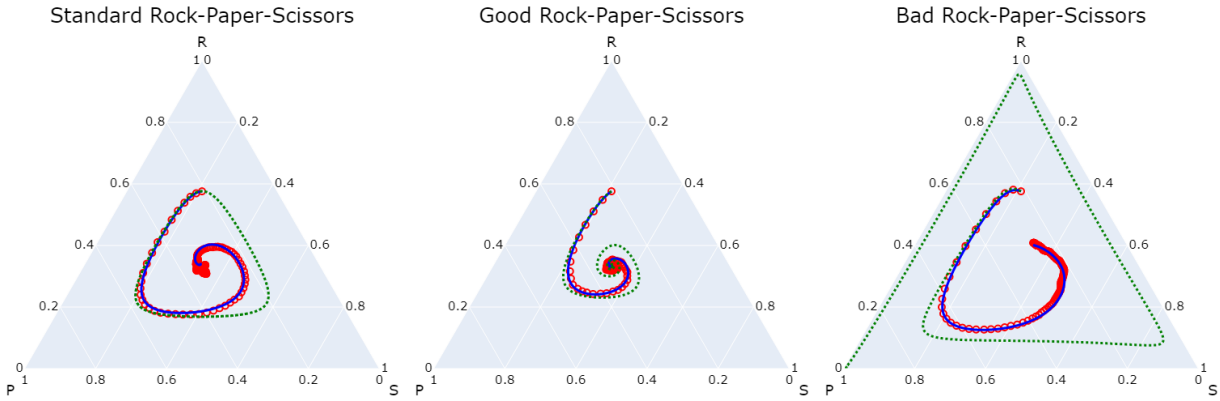

Figure 2: Ternary plot of expected value of policy $E[\boldsymbol{\pi}]$, initial parameter value $\boldsymbol{y}(0) = (1, 0, 0)$.

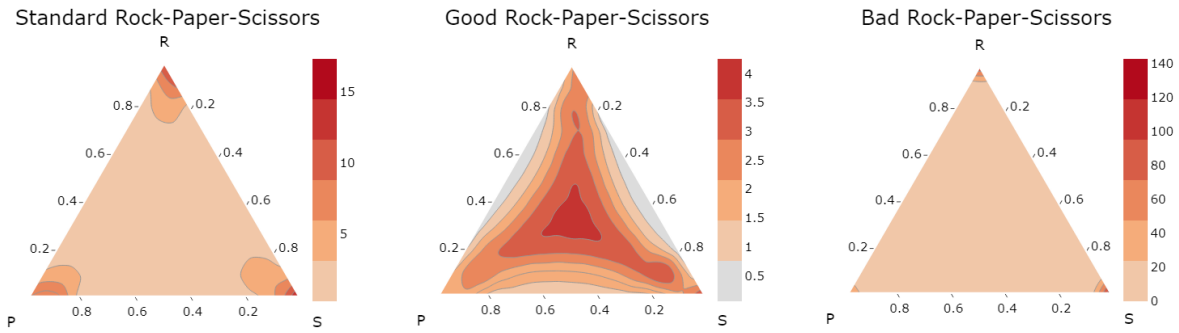

Figure 3: Ternary heat-map of probability density of strategy $p(\boldsymbol{\pi})$ at $t = 1,000$, initial parameter value $\boldsymbol{y}(0) = (1, 0, 0)$

Looking at the results, we can see that the blue and red lines almost overlapped with each other, meaning that our model can capture the behaviour of the true dynamics well. On the other hand, we can see substantial differences between the mean dynamics approach (green line) and the agent-based simulation, which confirms our claims in the earlier discussion.

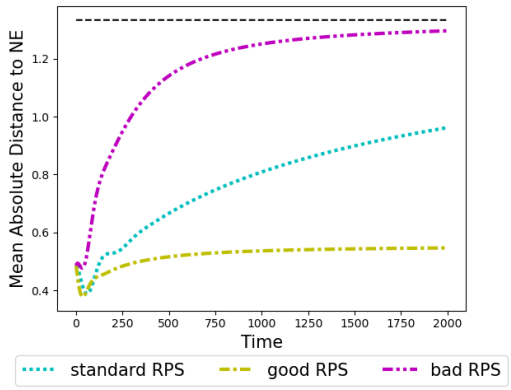

Figure 4: Mean absolute distance to NE over time for RPS games

5.3 Discussions on Convergence Behavior

Expected policy in PD, SH and HD games. Looking at Figure 1, we can see there is a substantial discrepancy between the mean dynamics model and the actual result in the HD game. Failing to consider the random effects, the mean dynamics model has predicted the agent's expected policy converges to the mixed Nash equilibrium strategy. The reason is that the mean dynamics have only focused on the expected payoff. However, the payoff variance of a policy will also affect the agent during learning, which in turn affects the equilibrium of the dynamical system. The result is aligned with the previous findings in EGT [7], where the asymptotic behaviour of the population can be affected by the existence of randomness.

For the PD and SH games, the expected policy of the population and the mean dynamics model have converged to the same pure Nash equilibrium strategy. However, we can still see some discrepancies in their policy dynamics, especially in the SH game. The existence of randomness hinders the agents from obtaining the correct information about the population, and this affects the speed of convergence in learning. As the mean dynamics model fails to consider the random effects, it will generally predict a faster convergence speed compared to the stochastic dynamics model.

Convergence behaviours among RPS games. Looking at figure 2, we can see for the RPS games, the expected behaviours vary in different payoff settings. For standard-RPS and good-RPS, the agent's expected policy converges to the Nash equilibrium, whereas it converges to some other point in bad-RPS. Note that the good-RPS game is a strictly stable game, the standard-RPS game is a null stable game, and the bad-RPS is an unstable game.

According to Theorem 4.8, the policy $\boldsymbol{\pi}$ in good-RPS will converge to a stationary distribution (and thus a stable expected policy) since good-RPS is strictly stable and $\alpha = 0.1 < 0.11 = \alpha_0$. This verifies the correctness of the theorem.

However, convergence is not guaranteed in standard-RPS and bad-RPS. In fact, when we examine the probability distribution of the agent's policy, we can see the policy distribution of standard-RPS and bad-RPS diverge to the boundary (pure strategy). Figure 3 plot the heat-map of $p(\boldsymbol{\pi})$ at $t = 1,000$ for 3 RPS games. We can see for good-RPS, the density is concentrated around the NE strategy $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. For standard-RPS, the density has split into 3 parts and scattered around the pure strategies, although its expected policy converges to the NE strategy. This illustrates that in standard-RPS, an agent is unlikely to learn a strategy that is close to the NE strategy. The divergent behaviour is more intense when we look at the heat map for bad-RPS. This shows that the policy distribution $p(\boldsymbol{\pi})$ does not converge to a stationary distribution in standard-RPS and bad-RPS.

Figure 4 shows the mean absolute distance to NE (MAD_{NE}) over time for 3 RPS games. The MAD_{NE} is evaluated as

$$MAD_{NE}(\boldsymbol{\pi}) = E_{\boldsymbol{y}}[\sum_{k=1}^{3} |\pi_k - \frac{1}{3}|]$$

The black dash line ($MAD_{NE} = \frac{4}{3}$) is the maximum possible value of MAD_{NE}, which corresponds to the situation where agents adopt only pure strategies. We can see for bad-RPS, this value quickly approaches the maximum value during training, which illustrates the divergence behaviour of $\boldsymbol{\pi}$. For standard-RPS, the value goes up steadily. For good-RPS, the value is stabilized around 0.55, which confirms the existence of stationary distribution of $\boldsymbol{\pi}$.

6 CONCLUSIONS

In this paper, we study the learning dynamics of the policy gradient algorithm in multi-agent environments. We propose a stochastic dynamics model to analyse the learning dynamics of PG under symmetric games. We model the parameter dynamics of a learning agent as a multidimensional Wiener process. Applying the Itô's lemma, we obtain the corresponding policy dynamics for the agent. From that, we study the convergence behaviour of the policy dynamics under a self-play training scheme for learning in games. We work out the sufficient conditions for the stochastic stability of the pure Nash equilibrium strategy and evaluate the sufficient conditions for the existence of stationary distribution for strictly stable games. Moreover, we express the dynamics of the parameter distribution with the Fokker-Planck equation. In the experiments, we demonstrate that our stochastic dynamics model always provides a significantly more accurate description of the actual learning dynamics than the mean dynamics model across different games and settings. Future directions include applying the analysis to other learning algorithms such as Q-learning and SARSA, as well as extending the analysis to multi-state environments and with stochastic rewards.

ACKNOWLEDGMENTS

This study is supported by the Leverhulme Trust for the Research Grant RPG-2023-050. The work presented in this paper is partially supported by a research grant from the Research Grants Council, Hong Kong, China (RGC Ref. No. CUHK 14206820).

REFERENCES

[1] Wolfram Barfuss, Jonathan F Donges, and Jürgen Kurths. 2019. Deterministic limit of temporal difference reinforcement learning for stochastic games. *Physical Review E* 99, 4 (2019), 043305.

[2] Martino Bernasconi, Federico Cacciamani, Simone Fioravanti, Nicola Gatti, and Francesco Trovo. 2022. The Evolutionary Dynamics of Soft-Max Policy Gradient in Games. *AAAI* (2022).

[3] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. 2015. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research* 53 (2015), 659–697.

[4] Vincenzo Capasso and David Bakstein. 2021. *Introduction to Continuous-Time Stochastic Processes.* Springer.

[5] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. 2020. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems* 33 (2020), 5527–5540.

[6] Adriaan Daniël Fokker. 1914. Die mittlere Energie rotierender elektrischer Dipole im Strahlungsfeld. *Annalen der Physik* 348, 5 (1914), 810–820.

[7] Dean Foster and Peyton Young. 1990. Stochastic evolutionary game dynamics. *Theoretical population biology* 38, 2 (1990), 219–232.

[8] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55, 1 (1997), 119–139.

[9] Drew Fudenberg and Christopher Harris. 1992. Evolutionary dynamics with aggregate shocks. *Journal of Economic Theory* 57, 2 (1992), 420–441.

[10] Eduardo Rodrigues Gomes and Ryszard Kowalczyk. 2009. Dynamic analysis of multiagent Q-learning with ε-greedy exploration. In *ICML*. 369–376.

[11] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. PMLR, 1861–1870.

[12] Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Rémi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duéñez-Guzmán, et al. 2020. Neural replicator dynamics: Multiagent learning via hedging policy gradients. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 492–501.

[13] Josef Hofbauer and Lorens A Imhof. 2009. Time averages, recurrence and transience in the stochastic replicator dynamics. (2009).

[14] Shuyue Hu, Chin-wing Leung, and Ho-fung Leung. 2019. Modelling the Dynamics of Multiagent Q-Learning in Repeated Symmetric Games: a Mean Field Theoretic Approach. In *NeurIPS*. 12125–12135.

[15] Shuyue Hu, Chin-Wing Leung, Ho-fung Leung, and Harold Soh. 2022. The Dynamics of Q-learning in Population Games: A Physics-inspired Continuity Equation Model. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 615–623.

[16] Lorens A Imhof. 2005. The long-run behavior of the stochastic replicator dynamics. (2005).

[17] Kiyosi Itô. 1944. Stochastic integral. *Proceedings of the Imperial Academy* 20, 8 (1944), 519 – 524. https://doi.org/10.3792/pia/1195572786

[18] Galin L Jones. 2004. On the Markov chain central limit theorem. (2004).

[19] Michael Kaisers, Daan Bloembergen, and Karl Tuyls. 2012. A common gradient in multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*. International Foundation for Autonomous Agents and Multiagent Systems, 1393–1394.

[20] Michael Kaisers and Karl Tuyls. 2010. Frequency adjusted multi-agent Q-learning. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. 309–316.

[21] Ardeshir Kianercy and Aram Galstyan. 2012. Dynamics of Boltzmann Q learning in two-player two-action games. *Physical Review E* 85, 4 (2012), 041145.

[22] Yuri Kifer. 1988. Random perturbations of dynamical systems. *Nonlinear Problems in Future Particle Accelerators* 189 (1988).

[23] Tomas Klos, Gerrit Jan van Ahee, and Karl Tuyls. 2010. Evolutionary dynamics of regret minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 82–96.

[24] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. 2017. A unified game-theoretic approach to multiagent reinforcement learning. In *NeurIPS*. 4190–4203.

[25] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent reinforcement learning in sequential social dilemmas. In *AAMAS*. 464–473.

[26] Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. 2021. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969* (2021).

[27] Stefanos Leonardos and Georgios Piliouras. 2022. Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory. *Artificial Intelligence* 304 (2022), 103653.

[28] Chin-wing Leung, Shuyue Hu, and Ho-fung Leung. 2022. Modelling the Dynamics of Multi-Agent Q-learning: The Stochastic Effects of Local Interaction and Incomplete Information. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 384–390.

[29] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[30] Nick Littlestone and Manfred K Warmuth. 1994. The weighted majority algorithm. *Information and computation* 108, 2 (1994), 212–261.

[31] Xuerong Mao. 2007. *Stochastic differential equations and applications.* Elsevier.

[32] Panayotis Mertikopoulos and Aris L Moustakas. 2010. The emergence of rational behavior in the presence of stochastic perturbations. (2010).

[33] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. PMLR, 1928–1937.

[34] Liviu Panait, Karl Tuyls, and Sean Luke. 2008. Theoretical advantages of lenient learners: An evolutionary game theoretic perspective. *Journal of Machine Learning Research* 9, Mar (2008), 423–457.

[35] M. Planck. 1917. *Über einen Satz der statistischen Dynamik und seine Erweiterung in der Quantentheorie.* Reimer. https://books.google.com.hk/books?id=Sf4wGwAACAAJ

[36] William H Sandholm. 2010. *Population games and evolutionary dynamics.* MIT press.

[37] JMPGR Smith and George R Price. 1973. The logic of animal conflict. *Nature* 246, 5427 (1973), 15–18.

[38] John Maynard Smith. 1982. *Evolution and the Theory of Games.* Cambridge university press.

[39] Sriram Srinivasan, Marc Lanctot, Vinicius Zambaldi, Julien Pérolat, Karl Tuyls, Rémi Munos, and Michael Bowling. 2018. Actor-critic policy optimization in partially observable multiagent environments. *Advances in neural information processing systems* 31 (2018).

[40] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.

[41] Peter D Taylor and Leo B Jonker. 1978. Evolutionary stable strategies and game dynamics. *Mathematical biosciences* 40, 1-2 (1978), 145–156.

[42] Karl Tuyls, Pieter Jan'T Hoen, and Bram Vanschoenwinkel. 2006. An evolutionary dynamical analysis of multi-agent learning in iterated games. *AAMAS* 12, 1 (2006), 115–153.

[43] Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. 2003. A selection-mutation model for q-learning in multi-agent systems. In *AAMAS*. 693–700.

[44] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3-4 (1992), 279–292.

[45] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3 (1992), 229–256.

[46] Michael Wunder, Michael L Littman, and Monica Babes. 2010. Classes of multiagent q-learning dynamics with epsilon-greedy exploration. In *ICML*. Citeseer, 1167–1174.

[47] Maarten Wyns and Jacques Du Toit. 2017. A finite volume–alternating direction implicit approach for the calibration of stochastic local volatility models. *International Journal of Computer Mathematics* 94, 11 (2017), 2239–2267.

[48] Runyu Zhang, Zhaolin Ren, and Na Li. 2021. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198* (2021).