The Dynamics of Policy Gradient in Social Dilemmas with Partner Selection

Benedict Russell
University of Warwick
benedict.i.russell@warwick.ac.uk

Chin-wing Leung
Queen's University Belfast
c.leung@qub.ac.uk

Paolo Turrini
University of Warwick
p.turrini@warwick.ac.uk

Abstract

In social dilemmas self-interested learning agents face the choice between the societal benefit of cooperation and the immediate reward of defection. Significant evidence exists on the benefits of assortment mechanisms such as partner selection for the emergence of cooperation, but this is largely available through agent-based simulations. In this paper, we provide an analytical solution to the problem, studying the policy-gradient dynamics in the repeated Prisoner's Dilemma with partner selection. We show how partner selection changes the opponent distribution and hence the reward landscape, and prove this promotes cooperation under simple rules known from the literature. In particular, we find that population variance is a necessary condition for cooperation to emerge under the mean dynamics. Using a two-dimensional Wiener process, we extend the dynamics to capture the stochastic effects of partner selection and the resulting opponent distribution. We derive a sufficient condition for the population to be cooperation-promoting and prove the existence of a stationary distribution. Simulations confirm that the stochastic model accurately captures the policy-gradient dynamics and clarifies how the learning rate affects the emergence of cooperation.

1 Introduction

Social dilemmas capture the tension faced by self-interested agents when given the option to contribute to a common goal, or exploit the contributions of others. Ensuring that autonomous systems can work together for the greater good is currently one of the biggest challenges in the field of Artificial Intelligence. Recent research has explored the cooperation of LLMs [51, 30], multi-agent reinforcement learning (MARL) [19], and evolutionary game theory [31]; each providing unique insight into how cooperation can be sustained.

In MARL, this problem is especially challenging, as agents experience a non-stationary environment induced by other learning agents [1]. Consequently, defection is a common outcome unless the agents are equipped with an additional mechanism such as behavioural signals of the opponent [35, 40] or dynamic interactions [2]. In his seminal work, Nowak [31] identified network reciprocity as a key mechanism for the emergence of cooperation. This has led to extensive research on cooperation on dynamic networks [2, 49, 9], as well as pairwise partner selection [1, 19].

Researchers in this area commonly find that partner selection rules which increase interaction frequency or duration between cooperative agents are key to the emergence of cooperation. A popular mechanism, coined Out-for-Tat (OFT), enforces cooperative pairs to stay together; whilst re-wiring pairs with a defector [16, 52]. Recent contributions have found allowing agents to learn partner

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

selection is sufficient for cooperation to emerge [39, 8]; the agents adopt OFT as a mechanism to avoid exploitation. When random matching is enforced, defection takes hold of the population.

Despite the empirical success, the current understanding of partner selection on emergent behaviour remains heavily reliant on agent-based simulations [23, 22]. While these simulations provide robust evidence that cooperation can emerge, they often fail to provide a formal theoretical justification for how these selection rules reshape the underlying learning dynamics. Specifically, there is a lack of analytical work that maps policy gradient (PG) updates onto the non-stationary reward landscapes induced by shifting opponent distributions. Without this theoretical foundation, it remains unclear which population conditions are strictly necessary for cooperation to take hold, or how stochasticity in action and partner selection influences long-term stability.

Contribution In this paper, we provide a formal theoretical analysis of policy gradient dynamics in the repeated Prisoner's Dilemma with partner selection. Using mean-field theory, we derive the conditional partner distribution to capture how partner selection rules alter the reward structure for learning agents. We prove that partner selection promotes cooperation under well-known rules (OFT and ROFT) and identify that population variance is a necessary condition for cooperation to emerge. We extend the mean dynamics to a stochastic model using a two-dimensional Wiener process to capture the randomness inherent in action selection and partner selection. We derive a sufficient condition for the population to be cooperation-promoting and prove the existence of a stationary distribution under the derived model. Simulations corroborate our theoretical findings, demonstrating that our stochastic model accurately captures the evolution of strategies and clarifies the pivotal role of the learning rate in fostering cooperative clusters.

Related Literature *Partner selection.* In evolutionary and network models, enabling cooperators to avoid defectors alters the payoff landscape, generating more assortative interaction patterns than a well-mixed population [32, 41, 33]. This can take many forms, including active linking [32, 34, 2], partner switching [10, 25], and conditional dissociation [16, 17]. Rewiring acts as a weak form of ostracism; defectors cannot repeatedly exploit cooperators, but cooperators have continued access to each other. These results carry over to human populations, where experimental work shows that dynamic partner updates can increase cooperation [36, 9, 52]. Recent research in MARL has implemented these ideas into learning environments. Anastassacos et al. [1] show that partner selection can induce cooperation amongst selfish reinforcement learning agents. This has been extended to enable agents to learn partner selection rules themselves, with behavioural signals [19, 22] and without any prior information on opponents [39].

Learning Dynamics. The complexity associated with non-stationary environments has led researchers to model learning dynamics using tools from dynamical systems and evolutionary game theory. Prior work has linked learning to coupled replicator equations [42, 12], selection-mutation models [47, 5], and mean field approximations [13]. Learning dynamics have been considered in population games, using mean-field theory to derive the dynamics under the concurrent learning protocol [14, 13]. This has been extended to agents acting on graphs [7], and considering stochastic interactions under the social learning protocol [20]. In a closely related work, Zheng et al. [53] use a pair approximation in an evolutionary setting to show the emergence of cooperation under a partner selection rule; our work differs by analysing a learning population.

Policy Gradient Dynamics. In 2-player symmetric games, Srinivasan et al. [45] provide the mean dynamics of the policy gradient and compare it against the replicator dynamics. Bernasconi et al. [3] formally establish the connection between the replicator dynamics and the soft-max policy gradient, and prove asymptotic stability for evolutionarily stable strategies (ESS) in symmetric games. This setting has been extended to analyse the stochastic effects of exploration [21], with stochastic stability shown under self-play. Stochastic stability has also been analysed when the payoff faces aggregate shocks [11] and random perturbations [28]. Recent work on stochastic gradient bandits establishes convergence with constant learning rates [26, 27]; our setting additionally introduces population-dependent rewards and noise through partner selection.

Whilst there has been extensive work on symmetric games, including the stochastic effects of shocks and action selection, there has not been theoretical research which shows the role of partner selection on the behaviour of reinforcement learning (RL) agents. This adds an additional layer of complexity, where the opponent distribution will influence both the mean and stochastic dynamics.

2 Background

In this section, we provide the necessary background on social dilemmas with partner selection and the policy gradient algorithm in the repeated Prisoner's Dilemma.

2.1 Repeated Prisoner's Dilemma

In the Prisoner's Dilemma (PD), two players simultaneously choose whether to cooperate (C) or defect (D), receiving payoffs based on their combined actions. Whilst mutual cooperation yields the greatest collective payoff, each player is incentivised to defect and exploit their opponent. Consequently, mutual defection remains the only Nash equilibrium [29] and ESS in the one-shot game [44]. The general payoff matrix is shown in Table 1; the payoff is such that $b > c > 0$.

Table 1: Prisoner's Dilemma payoff matrix.

	C	D
C	b	0
D	$b+c$	c

This paper examines a repeated Prisoner's Dilemma, where for each episode, an agent plays H rounds of PD with their opponents. The opponent changes based on the actions of the pair, according to a predefined partner selection rule, denoted by $\mathcal{R}$. We will consider 4 commonly studied rules [19, 39]:

- Out-for-Tat (OFT): stay if and only if both cooperate
- Reverse Out-for-Tat (ROFT): stay if and only if both defect
- Always Stay (Stay): stay independent of the actions
- Always Switch (Switch): switch independent of the actions

These provide a baseline to analyse how partner selection can be used to promote cooperation in a population of self-interested learning agents.

2.2 Policy Gradient

We study independent reinforcement learners where each agent aims at optimising their expected return. The policy gradient (PG) algorithm searches over the parameter space $\boldsymbol{\psi} \in \mathbb{R}^m$ and updates the policy through stochastic gradient ascent to find parameters that optimise the agent's decision-making. We consider softmax parameterisation for each agent's policy, which enables balancing exploration and exploitation. The policy for action a_k is given by

$$\pi(a_k|\boldsymbol{\psi}(t)) := \frac{\exp(\psi_k)}{\sum_{l=1}^{m}\exp(\psi_l)}. \tag{1}$$

The policy update follows the REINFORCE algorithm [50]. If action a_h is selected in round h of an episode, the reward-to-go with discount factor 1 is $R_h = \sum_{\ell=h}^{H-1} r_\ell$ and the approximated gradient of the score-function is $(R_h - \beta)(\boldsymbol{e}_{a_h} - \boldsymbol{\pi})$, where $\beta \in \mathbb{R}$ is a baseline [46] which we set to 0. The parameter update across an episode of length H is

$$\partial_t \boldsymbol{\psi} := \boldsymbol{\psi}(t+1) - \boldsymbol{\psi}(t) = \alpha \sum_{h=0}^{H-1} R_h(\boldsymbol{e}_{a_h} - \boldsymbol{\pi}). \tag{2}$$

where a_h is the action sampled at round h, and $\boldsymbol{e}_{a_h} = (0, ...1, ...0)$ is the basis vector. In the PD, the actions are limited to $a \in \{C, D\}$. Denoting the probability of cooperating $x := \pi(C|\boldsymbol{\psi}(t))$ and defecting $1 - x := \pi(D|\boldsymbol{\psi}(t))$, the *parameter* mean dynamics are given by

$$\partial_t \psi_C = \alpha x(1-x) \sum_{h=0}^{H-1} (G_C^h - G_D^h), \qquad \partial_t \psi_D = -\partial_t \psi_C,$$

where $G_C^h := \mathbb{E}[R_h|a_h = C]$ and $G_D^h := \mathbb{E}[R_h|a_h = D]$. Applying the chain rule to equation (1), the *policy* mean dynamics are

$$\frac{dx}{dt} = 2\alpha x^2(1-x)^2 \sum_{h=0}^{H-1} (G_C^h - G_D^h), \tag{3}$$

which is a one-dimensional non-linear ODE, where the sign of the episodic reward difference between cooperating and defecting will determine the evolution. Defining $G_a = \sum_{h=0}^{H-1} G_a^h$, we denote the overall difference by $\Delta G = G_C - G_D$.

3 Model Formulation

In this section, we introduce the agent-based model and derive how partner selection alters the opponent distribution and resulting reward structure.

3.1 Setup

We consider a population of agents, where the distribution of strategies is denoted by $\rho(x)$. Each strategy, x, corresponds to the probability of cooperating by that agent, and as such is bounded between 0 and 1. Therefore, if $x = 0$ the agent always defects, whilst if $x = 1$, the agent always cooperates. Policies are memoryless: each agent uses the same cooperation probability in each round, independently of the interaction history. This isolates the effect of partner selection from reciprocity.

For each episode, an agent i is randomly selected to be the focal agent. At round $h = 0$, sample an opponent j from the population; the pair interact by playing the PD game, and the payoff is given according to Table 1. The opponent j is then redrawn according to the partner selection rule. This repeats for a fixed number of rounds, H. At the end of the episode, the focal agent i updates their strategy according to the REINFORCE algorithm. Algorithm 1 provides the exact pseudocode for the interactions. We use the population timescale $t = e/N$, where e is the number of episodes and N is the population size. On average, each agent updates once per unit time.

The partner selection rule chosen will play a significant role in the evolutionary dynamics. For example, under the Out-For-Tat mechanism, an agent will continue to play with their partner if and only if they both cooperate. Otherwise, a new opponent is sampled (with replacement) from the population. The distribution of these opponents is inherently different: if the opponent stayed, they are more likely to have a higher cooperation rate x than a new draw from the population.

3.2 Conditional Opponent Distribution & Reward Structure

Consider an arbitrary agent i in the population, where the strategy of this agent is denoted by x. At step h of the episode, the opponent distribution given agent i's strategy is $q^h(y|x)$, where $q^0(y|x) = \rho(y)$. For any partner selection rule, the distribution at the next step is

$$q^{h+1}(y|x) = \int_0^1 q^{h+1}(y|x,z) q^h(z|x) dz.$$

The one-step change between the distributions is subject to the chosen Markovian partner selection rule. For example, under the OFT rule, the pair will stay if and only if both agents cooperate. Hence,

$$\begin{aligned} q^{h+1}(y|x) &= \int_0^1 \Big(\underbrace{\delta_z(y) xz}_{\text{both cooperate}} + \underbrace{q^{h+1}(y|x, \text{switch})(1-xz)}_{\text{otherwise, sample new partner}} \Big) q^h(z|x) dz \\ &= x \int_0^1 \delta_z(y) z q^h(z|x) dz + q^{h+1}(y|x, \text{switch}) \int_0^1 (1-xz) q^h(z|x) dz \\ &= xy q^h(y|x) + \rho(y)(1 - x m^h(x)) \end{aligned}$$

where, upon breaking a connection, a new opponent is sampled from the population. Here, δ_z is a point mass at the current partner's strategy z. The pair stays with probability xz; otherwise a new

partner is drawn from ρ. For the 4 partner selection rules, the recursion is given in Table 2. Given the focal agent's strategy x, the mean cooperation frequency of the opponents at step h is denoted by

$$m^h(x) = \mathbb{E}^h[Y|x] = \int_0^1 yq^h(y|x)dy.$$

For $X \sim \rho$, the ℓth moment of population is

$$\mu_\ell = \mathbb{E}_\rho[X^\ell] = \int_0^1 y^\ell \rho(y)dy, \quad \ell \geq 1.$$

In particular, $\text{Var}_\rho(X) = \mu_2 - \mu_1^2$. The opponent distribution provides an explicit way to capture the reward at each step, h. We can utilise this structure to derive the conditional rewards across the episode. Following equation (2), the expected policy update sums the reward-to-go contributions conditioned on the action adopted at each step. Consequently, we denote the agent's strategy, conditioned on its action in round k, as x, C^k and x, D^k, for cooperating and defecting respectively. Likewise, the opponent distribution, expected cooperation, and reward at time h are denoted by $q^h(y|x, C^k)$, $m_C^{k,h}(x)$, and $r_C^{k,h}$. At step $h = k$, the expected difference between cooperation and defection is always $-c$. The reward difference for $h \geq k+1$ is given by

$$\begin{aligned}
\Delta r^{k,h}(x) = r_C^{k,h} - r_D^{k,h} &= \int [by + (1-x)c](q^h(y|x, C^k) - q^h(y|x, D^k))dy \\
&= b \int y(q^h(y|x, C^k) - q^h(y|x, D^k))dy \\
&= b(m_C^{k,h}(x) - m_D^{k,h}(x)) \\
&= b\Delta m^{k,h}(x)
\end{aligned}$$

where given an agent has a policy x and cooperated in the kth round, $m_C^{k,h}(x)$ is explicitly given by

$$m_C^{k,h}(x) = \mathbb{E}^h[Y|x, C^k] = \int_0^1 yq^h(y|x, C^k)dy. \tag{4}$$

The reward differences for the first two steps in the episode are shown in Table 2 (a full derivation is provided in Appendix B). Regardless of the partner selection rule, the expected reward difference between cooperation and defection in the kth round is $-c$. Therefore, for cooperation to emerge the future rewards must compensate by pairing the focal agent with more cooperative agents which can build sustained partnerships. Interestingly, both OFT and ROFT provided the same expected reward difference across the first two rounds: keeping defective pairs together is as effective as keeping cooperative ones.

Table 2: Conditional partner distribution update step and per round reward difference for episode length $H = 2$. In each case, the initial distribution is $\rho(y)$.

$\mathcal{R}$	$q^{h+1}(y\|x)$	$\Delta r^{h,h}(x)$	$\Delta r^{0,1}(x)$
OFT	$xyq^h(y\|x) + \rho(y)(1 - xm^h(x))$	$-c$	$b(\mu_2 - \mu_1^2)$
ROFT	$(1-x)(1-y)q^h(y\|x) + \rho(y)(1 - (1-x)(1-m^h(x)))$	$-c$	$b(\mu_2 - \mu_1^2)$
Stay	$\rho(y)$	$-c$	0
Switch	$\rho(y)$	$-c$	0

We will now provide a more general result for when $H > 2$. The conditional reward difference for Stay and Switch is zero for all steps $h \geq k+1$. Here, we aim to show that cooperating has a non-negative effect on future rewards for the OFT and ROFT mechanisms.

Proposition 3.1. *For the OFT and ROFT partner selection rules, the reward difference* $\Delta r^{k,h}(x) \geq 0$ *for all* $x \in [0, 1]$ *and* $h \geq k+1$.

Proof. All proofs are provided in Appendix C. □

Corollary 3.1. *For the OFT and ROFT partner selection rules, any population distribution ρ, and any finite $H \geq 2$,*

$$\Delta G_H[\rho](x) \geq (H-1)b\,\mathrm{Var}_\rho(Y) - Hc, \qquad x \in [0,1].$$

Consequently, if $b\,\mathrm{Var}_\rho(Y) > 2c$, then $\Delta G_H[\rho](x) > 0$ for every $H \geq 2$ and every $x \in [0,1]$.

Proposition 3.1 identifies the future reward advantage created by partner selection. Corollary 3.1 gives a sufficient variance condition for this advantage to outweigh the immediate costs of cooperation. Applied to an initial population with positive mass in $(0,1)$, this condition ensures that mean cooperation increases over an initial time interval, for every finite $H \geq 2$. For $H = 2$, Theorem 3.2 further establishes that cooperation continues to increase and characterises the limiting distribution.

3.3 Evolution of the Mean Dynamics

In this section, we show how the above reward structure can be embedded within the policy gradient update. We then prove that Stay and Switch converge to the pure defection equilibrium. Finally, we show that both OFT and ROFT increase cooperation when the initial population has sufficient variance.

Under the Stay and Switch mechanisms, the reward difference simplifies to $-Hc$ for any episode length H. The policy update is given by

$$\frac{dx}{dt} = -2\alpha H c x^2 (1-x)^2$$

which generates the mean-field continuity PDE

$$\partial_t \rho + \partial_x(-2\alpha H c x^2 (1-x)^2 \rho) = 0.$$

Theorem 3.1. *Let $\rho_0 \in L^1(0,1), \rho_0 \geq 0$, and $\int_0^1 \rho_0 = 1$. Under the Stay and Switch rules, the population converges to pure defection.*

Proof Sketch. Under Stay and Switch, the current action does not improve the distribution of future partners. Cooperation therefore incurs an immediate cost of c without any future benefit, giving $\Delta G = -Hc < 0$. The negative drift moves all strategies towards defection. □

Consequently, we know that always-switch and always-stay will, under the mean dynamics, cause mass defection to take over the population. Therefore, we seek to show that partner selection rules such as OFT and ROFT can induce a more cooperative population. Under the OFT mechanism, the continuity equation is

$$\partial_t \rho + \partial_x(u[\rho](x)\rho) = 0, \qquad u[\rho](x) = 2\alpha x^2(1-x)^2 \Delta G[\rho](x),$$

where for $H = 2$, we have

$$\begin{aligned}
\Delta G[\rho] &= \sum_{h=0}^{H-1} \Delta G^h[\rho] \\
&= \Delta r^{0,0}(x) + \Delta r^{0,1}(x) + \Delta r^{1,1}(x) \\
&= -c + b(\mu_2 - \mu_1^2) - c \\
&= b\mathrm{Var}_\rho(X) - 2c.
\end{aligned}$$

For the same population distribution ρ, Corollary 3.1 implies that $b\mathrm{Var}_\rho(X) > 2c$ also gives a positive drift for every finite $H \geq 2$. Denote $\mathcal{P}([0,1])$ as the set of all probability measures on the domain.

Proposition 3.2. *Let $\rho_0 \in \mathcal{P}([0,1])$, then there exist a unique solution to the continuity equation. Furthermore, the unique solution can be expressed as*

$$\rho(t) = (\Phi_t)_{\#}\rho_0$$

where $(\Phi_t)_{\#}\rho_0$ is the push-forward of ρ_0 by the characteristic flow.

For the remainder of the section, we fix $H = 2$ and derive the limiting dynamics. Define the time integral of the non-local velocity induced by the episodic reward,

$$K(t) = \int_0^t \Delta G[\rho(s)]\, ds$$

Then the continuity PDE depends on time through K, and hence satisfies

$$\rho(t, \cdot) = (\Phi_{K(t)})_{\#}\rho_0. \tag{5}$$

See Appendix C for details. This enables the flow of the mass to be uniquely determined by the behaviour of K.

Theorem 3.2. *Fix $H = 2$ and let $\rho_0 \in L^1(0,1), \rho_0 \geq 0$, and $\int_0^1 \rho_0 = 1$. If $\Delta G[\rho_0] > 0$, then $K(t) \nearrow K^*$ where $K^* \in \mathbb{R}$ and $\rho(t) \rightharpoonup (\Phi_{K^*})_{\#}\rho_0$. That is, if the initial population variance is high enough, the limiting distribution is shifted towards higher cooperation.*

Proof Sketch. For $H = 2$, the reward advantage is common to all strategies and so the population dynamics reduce to the equation $K' = h(K)$. As $K \to \infty$, the variance vanishes and $h(K) \to -2c$. Since $h(0) > 0$, the flow approaches a finite root K^*. □

Here K^* is the finite limiting cumulative reward advantage. The characteristic flow is injective, so a non-Dirac initial distribution cannot collapse to a single strategy. Moreover, $x < \Phi_{K^*}(x) < 1$ for $x \in (0,1)$, so the limit has a higher cooperation mean but does not reach pure cooperation. The positivity of the reward difference $\Delta G[\rho_0]$ constrains the payoff; the variance on the domain $[0,1]$ is bounded above by $1/4$, meaning $b > 8c$ is a necessary condition when $H = 2$. This constraint relaxes as H increases, as the long term benefits of OFT and ROFT increase the reward difference.

4 Stochastic Dynamics

To capture the stochasticity induced by action exploration and the opponent distribution, we model the episodic REINFORCE update as a random increment in the parameter space, $\boldsymbol{\psi}$. To do this we follow the approach in [18, 21], characterising the random increment by the first two moments. In particular, we approximate the discrete stochastic updates by a continuous-time diffusion process:

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

where $\boldsymbol{\mu}$ is the expected update, $\boldsymbol{W}_t$ is a two-dimensional standard Wiener process, and $\Sigma := \Sigma(\boldsymbol{\psi}, t)$ denotes the covariance matrix. In 2-action repeated games, we can exploit the logit difference to simplify the stochastic dynamics. Applying Itô's lemma [15], we derive the policy evolution as

$$dx = [2\alpha x^2(1-x)^2\Delta G[\rho] + 2x(1-x)(1-2x)\Sigma_{CC}]\, dt + 2x(1-x)\sqrt{\Sigma_{CC}}\, dW_t \tag{7}$$

where Σ_{CC} denotes the variance of the parameter update. Denote the hth update step of the REINFORCE algorithm as $U^h := (R^h_{a_h} - \beta)(1\{a_h = C\} - x)$. The covariance is

$$\Sigma_{CC} := \alpha^2 \sum_{h=0}^{H-1} \Big(x(1-x)^2 S^h_C + x^2(1-x) S^h_D - x^2(1-x)^2 (\Delta G^h[\rho])^2 \Big) + 2\alpha^2 \sum_{h<l} \mathrm{Cov}(U^h, U^l),$$

where $S^h_C = \mathbb{E}[(R_h - \beta)^2 | a_h = C]$ and $S^h_D = \mathbb{E}[(R_h - \beta)^2 | a_h = D]$ are the second moments of the reward conditioned on actions C and D at round h. For any action a, this is given by

$$S^h_a = \sum_{l=h}^{H-1} \mathrm{Var}(r_l | a_h = a) + 2 \sum_{h \leq l < k} \mathrm{Cov}(r_l, r_k | a_h = a) + (G^h_a - \beta)^2.$$

Details of the variance and covariance terms are provided in Appendix B.

4.1 Population Evolution

By propagating the stochastic updates across the population, we can study the strategy evolution at a population scale. The Fokker-Planck equation (FPE) describes the evolution of the probability density function as a diffusion process [37]. For the strategy space x, the FPE corresponding to (7) is

$$\partial_t \rho(t,x) = -\partial_x[A_\rho(t,x)\rho(t,x)] + \frac{1}{2}\partial_{xx}[B^2_\rho(t,x)\rho(t,x)] \tag{8}$$

where

$$A_\rho(t,x) = 2\alpha x^2(1-x)^2\Delta G[\rho] + 2x(1-x)(1-2x)\Sigma_{CC}, \quad B^2_\rho(t,x) = 4x^2(1-x)^2\Sigma_{CC}.$$

Given an initial strategy distribution $\rho(0,x) = \rho_0(x)$, the time evolution can be solved numerically. These numerical solutions will enable us to analyse the long-term dynamics; for the short-term we can provide an equivalent sufficient condition to the mean dynamics for an increasing level of cooperation. In particular, the next result shows that given a sufficiently low learning rate, the cooperation levels in the population will initially increase for sufficiently high variance in the initial population. Define the critical learning rate threshold by

$$I_0 = \int_0^1 x^2(1-x)^2\Delta G[\rho_0](x)d\rho_0(x), \qquad \alpha^* = \frac{4I_0}{H^2(H(b+c)+|\beta|)^2}.$$

Proposition 4.1. *Suppose ρ_0 has non-zero interior mass and $I_0 > 0$. For every $0 < \alpha < \alpha^*$, there exists $T > 0$ such that the mean cooperation level is increasing on $[0,T]$.*

It does not, however, guarantee that a stable cooperative equilibrium will be reached. When $H = 2$, the high population variance would provide a sufficient condition initially, but the evolution of the variance depends upon higher-order moments. To study the stochastic case, we prove existence of a stationary measure and turn to numerical solutions to analyse the long-term behaviour.

The stationary distribution occurs at the point when the time derivative is zero. Applying no-flux boundary conditions, this reduces to finding a solution to the ODE given by

$$\partial_x[B^2_\rho(x)\rho(x)] = 2A_\rho(x)\rho(x). \tag{9}$$

Theorem 4.1. *There exists at least one weak solution $\rho \in \mathcal{P}([0,1])$ which solves the steady-state equation* (9)*.*

Proof Sketch. Add a positive regularisation to the diffusion coefficient to give a continuous, compact mapping from trial populations to stationary densities. Schauder's fixed-point theorem yields a stationary density; as the regularisation is removed, a convergent subsequence yields a stationary probability measure. □

These stationary measures can contain atoms at the boundary. In particular, every mix of δ_0 and δ_1 is stationary as both the drift and diffusion vanish. Consequently, simulations may exhibit mass accumulation at pure defection and cooperation. Moreover, non-uniqueness of the stationary measures explains why the dynamics are strongly influenced by the initial population and parameter regimes.

5 Experiments

In this section, we show how the theoretical analysis and model capture the dynamics displayed in the agent-based simulations. To do so, we solve equation (8) for the 4 different partner selection rules.

5.1 Game Setting

We use a finite volume method over the domain $[0,1]$ to capture the time and spatial evolution. The agent-based simulations act as a ground truth in the experiments; we conduct 30 simulations of a population of $N = 1,000$ training over $E = 5 \times 10^6$ episodes. The population distribution at time t is obtained by averaging the simulations. Simulation time is scaled such that $t = e/N$.

The initial strategy values are sampled from a specified distribution ρ_0, with the parameter $\boldsymbol{\psi}(0)$ obtained through the inverse of the softmax function (1). The payoff is given by Table 1, where $b = 3$ and $c = 0.1$. Unless specified, the parameters are $\alpha = 0.01, H = 2$. Evolution under a wide range of initial conditions is shown in Appendix D.

5.2 Results

We compare the distribution of strategies in the population over time. Figure 1 presents the results of both the theoretical solution (solid line) and the simulations (histogram). We clearly see that the derived FPE captures the distribution of strategies as it evolves through time. This close alignment is present for all 4 partner selection rules, where the behaviour of the two cooperation-inducing rules (OFT and ROFT) displays very similar dynamics which lead to the emergence of a defecting and cooperating cluster. On the other hand, Stay and Switch induce defection to take over the population.

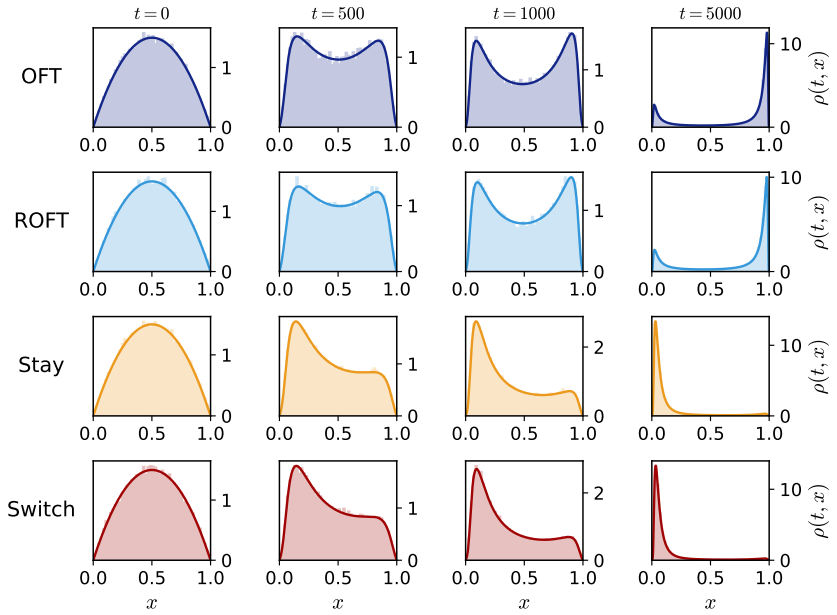

Figure 1: Evolution of the strategy distribution where the population is initialised with $X \sim \text{Beta}(2, 2)$ for the 4 rules. Simulation results are shown with a histogram, whilst theoretical evolution is plotted with solid lines; the theory matches the simulations almost exactly.

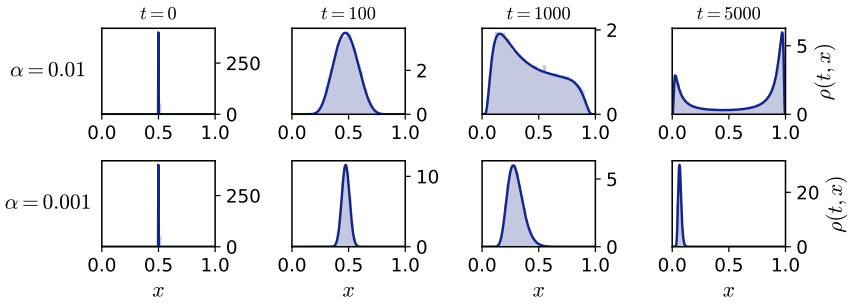

Figure 2: Evolution of the strategy distribution under OFT where the population is initialised with a Dirac at 0.5. Simulation results are shown with a histogram, whilst theoretical evolution is plotted with solid lines. The effect of the learning rate is clear: higher values induce additional variance in the population which in turn induces cooperation. Time is scaled with the learning rate for comparison.

As highlighted in Section 3.3, the underlying distribution requires sufficient variance for cooperation to emerge under the mean dynamics. This begs the question of how and why cooperation can emerge when the population is initialised with a common policy [19, 39]. The answer lies in how the learning rate can induce population variance. With low learning rates, the stochastic effects are diminished and therefore policies are updated close to the underlying expectation. A higher value of α enables

sampled trajectories to influence the dynamics and therefore induce a faster-growing population variance. Figure 2 shows how for an initially unbiased population, a cooperation cluster can be supported through only increasing the learning rate. Whilst a higher learning rate can induce greater variance and support cooperation, the FPE approximation loses fidelity.

Table 3: The effect of episode length on cooperation. Results show the mean and variance at $t = 5000$, averaged over 30 simulations with $N = 1000$. Population is initialised such that $X \sim \text{Unif}(0, 1)$ and the parameters are $b = 1$, $c = 0.1$, $\alpha = 0.01$, and $\beta = 0$.

	OFT		ROFT		Stay		Switch	
H	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
2	0.21	0.063	0.21	0.063	0.13	0.044	0.13	0.043
3	0.45	0.14	0.26	0.072	0.087	0.034	0.086	0.033
4	0.69	0.15	0.49	0.097	0.064	0.026	0.064	0.027
5	0.76	0.15	0.67	0.095	0.051	0.022	0.051	0.022

Table 3 shows that longer episodes can support greater cooperation, despite the initial population not satisfying the sufficient condition $b\text{Var}_\rho(X) > 2c$. The longer episode lengths coupled with the learning rate can induce greater strategic variety through unique trajectories.

The episode length distinguishes the behaviour between OFT and ROFT; $H = 2$ yields the same cooperation level for each rule, whilst OFT outperforms ROFT for $H > 2$. This captures the difference between staying with a cooperative partner - incentivising pro-social behaviour, and staying with a defecting partner - disincentivising anti-social behaviour. In contrast, Stay and Switch provide similar levels of cooperation for all episode lengths.

6 Conclusion

In this paper, we study the learning dynamics of the policy gradient in a multi-agent environment when partner selection rules are enforced. We show how the opponent distribution alters the reward landscape, and prove that the underlying dynamics promote cooperation for both the OFT and ROFT partner selection rules. In particular, we find that population variance is a fundamental requirement for the emergence of cooperation under partner selection. We extend the model to capture the stochastic dynamics using a 2-dimensional Wiener process which models the stochasticity induced by action and opponent selection. In this setting, we show how the same variance requirement is sufficient for an initial increase in cooperation with a sufficiently small learning rate, before proving the existence of a stationary distribution. Numerical solutions exhibit increasing concentration near the pure defection and pure cooperation strategies, consistent with the stationary measures supported by the model. The experiments further demonstrate that the stochastic model accurately describes the evolution of the policy gradient and can be used to understand the role of the learning rate in inducing cooperation.

The analysis assumes memoryless policies and asynchronous agent updates. It therefore excludes history-dependent reciprocity and the fast learning that could arise when frequently selected opponents also update. Extensions to reactive policies, synchronous updates, and other learning algorithms remain future work.

Acknowledgments and Disclosure of Funding

BR was supported by the Engineering and Physical Sciences Research Council through the Mathematics of Systems II Centre for Doctoral Training at the University of Warwick (reference EP/S022244/1). CL and PT acknowledge the support of the Leverhulme Trust for the Research Grant RPG 2023-050.

References

[1] N. Anastassacos, S. Hailes, and M. Musolesi. Partner Selection for the Emergence of Cooperation in Multi-Agent Systems Using Reinforcement Learning. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, Feb. 2020. doi:10.1609/aaai.v34i05.6190. URL https://doi.org/10.1609/aaai.v34i05.6190.

[2] J. Bara, P. Turrini, and G. Andrighetto. Enabling imitation-based cooperation in dynamic social networks. *Autonomous Agents and Multi-Agent Systems*, 36(2):34, May 2022. ISSN 1573-7454. doi:10.1007/s10458-022-09562-w. URL https://doi.org/10.1007/s10458-022-09562-w.

[3] M. Bernasconi, F. Cacciamani, S. Fioravanti, N. Gatti, and F. Trovò. The evolutionary dynamics of soft-max policy gradient in multi-agent settings. *Theoretical Computer Science*, 1027: 115011, Feb. 2025. ISSN 0304-3975. doi:10.1016/j.tcs.2024.115011. URL https://www.sciencedirect.com/science/article/pii/S0304397524006285.

[4] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., second edition, 1999. ISBN 0-471-19745-9. A Wiley-Interscience Publication.

[5] D. Bloembergen, K. Tuyls, D. Hennes, and M. Kaisers. Evolutionary Dynamics of Multi-Agent Learning: A Survey. *Journal of Artificial Intelligence Research*, 53:659–697, Aug. 2015. doi:10.1613/jair.4818.

[6] B. Bonnet and F. Rossi. The pontryagin maximum principle in the wasserstein space. *Calculus of Variations and Partial Differential Equations*, 58(1):11, Dec. 2018. ISSN 0944-2669, 1432-0835. doi:10.1007/s00526-018-1447-2. URL http://arxiv.org/abs/1711.07667.

[7] C. Chu, Y. Li, J. Liu, S. Hu, X. Li, and Z. Wang. A Formal Model for Multiagent Q-Learning Dynamics on Regular Graphs. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 194–200, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-00-3. doi:10.24963/ijcai.2022/28. URL https://www.ijcai.org/proceedings/2022/28.

[8] X. Fan, C.-w. Leung, and P. Turrini. Co-learning of strategy and structure achieves full cooperation in complex networks with dynamical linking. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 72–80. International Joint Conferences on Artificial Intelligence Organization, Sep. 2025. doi:10.24963/ijcai.2025/9.

[9] K. Fehl, D. J. van der Post, and D. Semmann. Co-evolution of behaviour and social network structure promotes human cooperation. *Ecology Letters*, 14(6):546–551, June 2011. ISSN 1461-0248. doi:10.1111/j.1461-0248.2011.01615.x.

[10] F. Fu, T. Wu, and L. Wang. Partner switching stabilizes cooperation in coevolutionary prisoner's dilemma. *Physical Review E*, 79(3):036101, Mar. 2009. ISSN 1539-3755, 1550-2376. doi:10.1103/PhysRevE.79.036101. URL https://link.aps.org/doi/10.1103/PhysRevE.79.036101.

[11] D. Fudenberg and C. Harris. Evolutionary dynamics with aggregate shocks. *Journal of Economic Theory*, 57(2):420–441, Aug. 1992. ISSN 00220531. doi:10.1016/0022-0531(92)90044-I. URL https://linkinghub.elsevier.com/retrieve/pii/002205319290044I.

[12] A. Galstyan. Continuous strategy replicator dynamics for multi-agent Q-learning. *Autonomous Agents and Multi-Agent Systems*, 26(1):37–53, Jan. 2013. ISSN 1573-7454. doi:10.1007/s10458-011-9181-6. URL https://doi.org/10.1007/s10458-011-9181-6.

[13] S. Hu, C.-w. Leung, and H.-f. Leung. Modelling the dynamics of multiagent q-learning in repeated symmetric games: a mean field theoretic approach. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://dl.acm.org/doi/10.5555/3454287.3455374.

[14] S. Hu, C.-W. Leung, H.-f. Leung, and H. Soh. 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*, AAMAS '22, page 615–623, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450392136. URL https://dl.acm.org/doi/10.5555/3535850.3535920.

[15] K. Itô. Stochastic integral. *Proceedings of the Imperial Academy*, 20(8): 519–524, Jan. 1944. ISSN 0369-9846. doi:10.3792/pia/1195572786. URL https://projecteuclid.org/journals/proceedings-of-the-imperial-academy/volume-20/issue-8/Stochastic-integral/10.3792/pia/1195572786.full.

[16] L. R. Izquierdo, S. S. Izquierdo, and F. Vega-Redondo. Leave and let leave: A sufficient condition to explain the evolutionary emergence of cooperation. *Journal of Economic Dynamics and Control*, 46:91–113, Sept. 2014. ISSN 0165-1889. doi:10.1016/j.jedc.2014.06.007. URL https://www.sciencedirect.com/science/article/pii/S0165188914001456.

[17] S. S. Izquierdo, L. R. Izquierdo, and F. Vega-Redondo. The option to leave: Conditional dissociation in the evolution of cooperation. *Journal of Theoretical Biology*, 267(1):76–84, Nov. 2010. ISSN 0022-5193. doi:10.1016/j.jtbi.2010.07.039. URL https://www.sciencedirect.com/science/article/pii/S0022519310003966.

[18] Y. Kifer. Random Perturbations of Dynamical Systems, 1988. URL https://link.springer.com/book/10.1007/978-1-4615-8181-9.

[19] C.-w. Leung and P. Turrini. Learning partner selection rules that sustain cooperation in social dilemmas with the option of opting out. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, page 1110–1118, Richland, SC, 2024. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400704864. doi:10.65109/jtrb7413. URL http://dx.doi.org/10.65109/jtrb7413.

[20] C.-w. Leung, S. Hu, and H.-f. Leung. 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*, pages 384–390, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-00-3. doi:10.24963/ijcai.2022/55. URL https://www.ijcai.org/proceedings/2022/55.

[21] C.-w. Leung, S. Hu, and H.-f. Leung. The stochastic evolutionary dynamics of softmax policy gradient in games. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, page 1101–1109, Richland, SC, 2024. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400704864. URL https://dl.acm.org/doi/10.5555/3635637.3662966.

[22] C.-w. Leung, P. Turrini, and A. Nowé. Curiosity-driven partner selection accelerates convention emergence in language games. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, page 1282–1290, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269. URL https://dl.acm.org/doi/10.5555/3709347.3743760.

[23] C.-w. Leung, P. Turrini, F. P. Santos, and M. Musolesi. Learning to Cooperate with Minimal Observability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(35):29521–29529, Mar. 2026. ISSN 2374-3468. doi:10.1609/aaai.v40i35.40194. URL https://ojs.aaai.org/index.php/AAAI/article/view/40194.

[24] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2002. URL https://doi.org/10.1017/CBO9780511791253.005.

[25] Z. Li, Z. Yang, T. Wu, and L. Wang. Aspiration-Based Partner Switching Boosts Cooperation in Social Dilemmas. *PLOS ONE*, 9(6):e97866, June 2014. ISSN 1932-6203. doi:10.1371/journal.pone.0097866. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0097866.

[26] J. Mei, Z. Zhong, B. Dai, A. Agarwal, C. Szepesvari, and D. Schuurmans. Stochastic gradient succeeds for bandits. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023. URL https://dl.acm.org/doi/10.5555/3618408.3619421.

[27] J. Mei, B. Dai, A. Agarwal, S. Vaswani, A. Raj, C. Szepesvári, and D. Schuurmans. Small steps no more: global convergence of stochastic gradient bandits for arbitrary learning rates. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385. URL https://dl.acm.org/doi/10.5555/3737916.3740287.

[28] P. Mertikopoulos and A. L. Moustakas. The emergence of rational behavior in the presence of stochastic perturbations. *The Annals of Applied Probability*, 20(4), Aug. 2010. ISSN 1050-5164. doi:10.1214/09-AAP651. URL http://arxiv.org/abs/0906.2094. arXiv:0906.2094 [math].

[29] J. Nash. Non-Cooperative Games. *Annals of Mathematics*, 54(2):286–295, 1951. ISSN 0003-486X. doi:10.2307/1969529. URL https://www.jstor.org/stable/1969529.

[30] D. Nguyen, H. Le, K. Do, S. Gupta, S. Venkatesh, and T. Tran. Navigating social dilemmas with llm-based agents via consideration of future consequences. In J. Kwok, editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 223–231. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi:10.24963/ijcai.2025/26. URL https://doi.org/10.24963/ijcai.2025/26. Main Track.

[31] M. A. Nowak. Five rules for the evolution of cooperation. *Science (New York, N.y.)*, 314 (5805):1560–1563, Dec. 2006. ISSN 0036-8075. doi:10.1126/science.1133755. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3279745/.

[32] J. M. Pacheco, A. Traulsen, and M. A. Nowak. Active linking in evolutionary games. *Journal of Theoretical Biology*, 243(3):437–443, Dec. 2006. ISSN 0022-5193. doi:10.1016/j.jtbi.2006.06.027. URL https://www.sciencedirect.com/science/article/pii/S0022519306002736.

[33] J. M. Pacheco, A. Traulsen, and M. A. Nowak. Coevolution of Strategy and Structure in Complex Networks with Dynamical Linking. *Physical Review Letters*, 97(25):258103, Dec. 2006. ISSN 0031-9007, 1079-7114. doi:10.1103/PhysRevLett.97.258103. URL https://link.aps.org/doi/10.1103/PhysRevLett.97.258103.

[34] J. M. Pacheco, A. Traulsen, H. Ohtsuki, and M. A. Nowak. Repeated games and direct reciprocity under active linking. *Journal of Theoretical Biology*, 250(4):723–731, Feb. 2008. ISSN 00225193. doi:10.1016/j.jtbi.2007.10.040. URL https://linkinghub.elsevier.com/retrieve/pii/S0022519307005450.

[35] T. Priklopil, K. Chatterjee, and M. Nowak. Optional interactions and suspicious behaviour facilitates trustful cooperation in prisoners dilemma. *Journal of Theoretical Biology*, 433: 64–72, Nov. 2017. ISSN 0022-5193. doi:10.1016/j.jtbi.2017.08.025. URL https://www.sciencedirect.com/science/article/pii/S0022519317303995.

[36] D. G. Rand, S. Arbesman, and N. A. Christakis. Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*, 108(48): 19193–19198, Nov. 2011. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1108243108. URL https://pnas.org/doi/full/10.1073/pnas.1108243108.

[37] H. Risken. Fokker-Planck Equation. In H. Risken, editor, *The Fokker-Planck Equation: Methods of Solution and Applications*, pages 63–95. Springer, Berlin, Heidelberg, 1996. ISBN 978-3-642-61544-3. doi:10.1007/978-3-642-61544-3_4. URL https://doi.org/10.1007/978-3-642-61544-3_4.

[38] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 3 edition, 1976. ISBN 978-0070856134.

[39] B. Russell, C.-w. Leung, and P. Turrini. Defection at First Sight: Learning Partner Selection in Optional Social Dilemmas without Prior Information. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '26, pages 264–272, Richland, SC, May 2026. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 979-8-4007-2317-9. doi:10.65109/IBSZ1473. URL https://dl.acm.org/doi/10.65109/IBSZ1473.

[40] J. Sabater-Mir and C. Sierra. Reputation and social network analysis in multi-agent systems. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, pages 475–482, July 2002. doi:10.1145/544741.544854.

[41] F. C. Santos, J. M. Pacheco, and T. Lenaerts. Cooperation Prevails When Individuals Adjust Their Social Ties. *PLoS Computational Biology*, 2(10):e140, Oct. 2006. ISSN 1553-7358. doi:10.1371/journal.pcbi.0020140. URL https://dx.plos.org/10.1371/journal.pcbi.0020140.

[42] Y. Sato and J. P. Crutchfield. Coupled Replicator Equations for the Dynamics of Learning in Multiagent Systems. *Physical Review E*, 67(1):015206, Jan. 2003. ISSN 1063-651X, 1095-3787. doi:10.1103/PhysRevE.67.015206. URL http://arxiv.org/abs/nlin/0204057. arXiv:nlin/0204057.

[43] J. Shapiro. *A Fixed-Point Farrago*. Springer, 01 2016. ISBN 978-3-319-27976-3. doi:10.1007/978-3-319-27978-7.

[44] J. M. Smith. *Evolution and the Theory of Games*. Cambridge University Press, 1982. URL https://www.jstor.org/stable/27847040.

[45] S. Srinivasan, M. Lanctot, V. Zambaldi, J. Pérolat, K. Tuyls, R. Munos, and M. Bowling. Actor-critic policy optimization in partially observable multiagent environments. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3426–3439, Red Hook, NY, USA, 2018. Curran Associates Inc. URL https://dl.acm.org/doi/abs/10.5555/3327144.3327261.

[46] R. S. Sutton and A. G. Barto. Reinforcement learning - an introduction, 2nd edition. 2018. URL https://api.semanticscholar.org/CorpusID:277058247.

[47] K. Tuyls, K. Verbeeck, and T. Lenaerts. A selection-mutation model for q-learning in multi-agent systems. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, AAMAS '03, pages 693–700, New York, NY, USA, July 2003. Association for Computing Machinery. ISBN 978-1-58113-683-8. doi:10.1145/860575.860687. URL https://dl.acm.org/doi/10.1145/860575.860687.

[48] C. Villani. *Optimal transport: old and new*, volume 338. Springer, Berlin, 2009. ISBN 9781281851208;3540710507;1281851205;9783540710509;. doi:10.1007/978-3-540-71050-9.

[49] J. Wang, S. Suri, and D. J. Watts. Cooperation and assortativity with dynamic partner updating. *Proceedings of the National Academy of Sciences*, 109(36):14363–14368, Sept. 2012. doi:10.1073/pnas.1120867109. URL https://www.pnas.org/doi/10.1073/pnas.1120867109.

[50] R. J. Williams. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, 8(3-4):229–256, May 1992. ISSN 0885-6125, 1573-0565. doi:10.1023/A:1022672621406. URL https://link.springer.com/10.1023/A:1022672621406.

[51] R. Willis, Y. Du, J. Z. Leibo, and M. Luck. Will Systems of LLM Agents Cooperate: An Investigation into a Social Dilemma, Jan. 2025. URL https://arxiv.org/abs/2501.16173v1.

[52] B.-Y. Zhang, S.-J. Fan, C. Li, X.-D. Zheng, J.-Z. Bao, R. Cressman, and Y. Tao. Opting out against defection leads to stable coexistence with cooperation. *Scientific Reports*, 6:35902, Oct. 2016. ISSN 2045-2322. doi:10.1038/srep35902. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5075917/.

[53] X.-D. Zheng, C. Li, J.-R. Yu, S.-C. Wang, S.-J. Fan, B.-Y. Zhang, and Y. Tao. A simple rule of direct reciprocity leads to the stable coexistence of cooperation and defection in the Prisoner's Dilemma game. *Journal of Theoretical Biology*, 420:12–17, May 2017. ISSN 00225193. doi:10.1016/j.jtbi.2017.02.036. URL https://linkinghub.elsevier.com/retrieve/pii/S0022519317300991.

A Derivation of the Population Equations

A.1 Finite population model

Algorithm 1 Policy gradient with partner selection

Require: N, H, E, α, β, $\mathcal{R}$.

Initialise policy parameters $\psi_1, \ldots, \psi_N$
for $e = 0, \ldots, E-1$ **do**
  Sample focal agent i uniformly from $\{1, \ldots, N\}$
  Sample opponent j from the population
  $\pi_i \leftarrow \mathrm{softmax}(\psi_i)$, $\pi_j \leftarrow \mathrm{softmax}(\psi_j)$
  for $h = 0, \ldots, H-1$ **do**
    Sample $a_h \sim \pi_i$ and $\tilde{a}_h \sim \pi_j$
    Record focal reward $r_h = r(a_h, \tilde{a}_h)$
    if $h < H-1$ and the $\mathcal{R}$ prescribes switching **then**
      Resample opponent j
      $\pi_j \leftarrow \mathrm{softmax}(\psi_j)$
    end if
  end for
  $R_h \leftarrow \sum_{\ell=h}^{H-1} r_\ell$ for each $h = 0, \ldots, H-1$
  $\psi_i \leftarrow \psi_i + \alpha \sum_{h=0}^{H-1} (R_h - \beta)(e_{a_h} - \pi_i)$
end for

A.2 Continuous Mean-Dynamics

Let $x_i(t) \in [0,1]$ denote the cooperation probability of agent i, and define the empirical measure

$$\rho_t^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i(t)}.$$

For a probability measure η and a test function φ, write

$$\langle \eta, \varphi \rangle := \int_0^1 \varphi(x)\, d\eta(x).$$

We write $\Delta G[\eta](x)$ for the episodic reward difference experienced by an agent with strategy x in population η. The continuous mean dynamics give

$$\frac{dx_i}{dt} = u[\rho_t^N](x_i), \qquad u[\eta](x) := 2\alpha x^2 (1-x)^2 \Delta G[\eta](x).$$

For $\varphi \in C^1([0,1])$, the chain rule yields

$$\begin{aligned}
\frac{d}{dt}\langle \rho_t^N, \varphi \rangle &= \frac{1}{N} \sum_{i=1}^{N} \varphi'(x_i) \frac{dx_i}{dt} \\
&= \frac{1}{N} \sum_{i=1}^{N} \varphi'(x_i) u[\rho_t^N](x_i) \\
&= \int_0^1 \varphi'(x) u[\rho_t^N](x)\, d\rho_t^N(x).
\end{aligned}$$

Thus ρ_t^N is an exact weak solution of the continuity equation associated with this continuous particle system. Formally passing to a limit $\rho_t^N \rightharpoonup \rho_t$ gives

$$\frac{d}{dt}\langle \rho_t, \varphi \rangle = \int_0^1 \varphi'(x) u[\rho_t](x)\, d\rho_t(x).$$

When ρ_t has a sufficiently regular density, integration by parts yields

$$\partial_t \rho(t,x) + \partial_x \left(2\alpha x^2 (1-x)^2 \Delta G[\rho_t](x) \rho(t,x) \right) = 0.$$

Here $u[\eta](0) = u[\eta](1) = 0$, so the pure strategies are fixed under the mean dynamics.

A.3 Stochastic Dynamics

Conditioned on the current parameters $\boldsymbol{\psi}$ and population distribution ρ_t, define the mean update vector and covariance matrix by

$$\mu(\boldsymbol{\psi}, t) := \mathbb{E}[\Delta\boldsymbol{\psi}(t)], \qquad \Sigma(\boldsymbol{\psi}, t) = \mathbb{E}[(\Delta\boldsymbol{\psi}(t) - \mu)(\Delta\boldsymbol{\psi}(t) - \mu)^T].$$

In the two-action case, since $\Delta\psi_C = -\Delta\psi_D$, we have

$$\mu = (\mu_C, -\mu_C)^T, \qquad \Sigma = \begin{pmatrix} \Sigma_{CC} & -\Sigma_{CC} \\ -\Sigma_{CC} & \Sigma_{CC} \end{pmatrix}$$

where $\mu_C = \alpha x(1-x)\Delta G[\rho_t](x)$, and $\Delta G[\rho_t] = G_C - G_D$. Equation (6) can be represented using a single Wiener process

$$d\psi_C = \mu_C dt + \sqrt{\Sigma_{CC}} dW_t, \quad d\psi_D = -\mu_C dt - \sqrt{\Sigma_{CC}} dW_t.$$

Let $z = \psi_C - \psi_D$ and $x = s(z)$ where s is the sigmoid function. We have

$$dz = 2\mu_C dt + 2\sqrt{\Sigma_{CC}} dW_t, \quad s'(z) = x(1-x), \quad s''(z) = x(1-x)(1-2x).$$

Applying Itô's formula,

$$dx = \underbrace{[2x(1-x)\mu_C + 2x(1-x)(1-2x)\Sigma_{CC}]}_{A_{\rho_t}(x)} dt + \underbrace{2x(1-x)\sqrt{\Sigma_{CC}}}_{B_{\rho_t}(x)} dW_t$$

Substituting in the expression for μ_C gives equation (7). Under the mean-field approximation, let X_t denote the strategy of a representative agent satisfying the above SDE, with $\rho_t = \mathrm{Law}(X_t)$. Let $\varphi \in C^2([0,1])$. Applying Itô's formula to the above and taking expectations gives

$$\begin{aligned} \frac{d}{dt}\int_0^1 \varphi(x)\, d\rho_t(x) &= \frac{d}{dt}\mathbb{E}[\varphi(X_t)] \\ &= \mathbb{E}\left[A_{\rho_t}(X_t)\varphi'(X_t) + \frac{1}{2}B^2_{\rho_t}(X_t)\varphi''(X_t)\right] \\ &= \int_0^1 \left[A_{\rho_t}(x)\varphi'(x) + \frac{1}{2}B^2_{\rho_t}(x)\varphi''(x)\right] d\rho_t(x). \end{aligned}$$

This is the weak form of equation (8). The differential form follows by integration by parts.

B Conditional Rewards and Moments

B.1 Partner Distributions and Rewards

For additional clarity, we provide the explicit iterative form for the first two steps of the episode with the 4 partner selection rules. All steps are computed using the iterative formulas provided in Table 2. For each, the initial opponent distribution is set as $\rho(y)$.

Table 4: Conditional partner distribution and reward difference for each rule.

$\mathcal{R}$	h	$q^h(y\|x, C^0)$	$q^h(y\|x, D^0)$	$m_C^{0,h}(x)$	$m_D^{0,h}(x)$	$\Delta r^{0,h}(x)$
OFT	0	ρ	ρ	μ_1	μ_1	$-c$
	1	$(y+1-\mu_1)\rho$	ρ	$\mu_2 - \mu_1^2 + \mu_1$	μ_1	$\mathsf{b}(\mu_2 - \mu_1^2)$
ROFT	0	ρ	ρ	μ_1	μ_1	$-c$
	1	ρ	$(1+\mu_1-y)\rho$	μ_1	$\mu_1 + \mu_1^2 - \mu_2$	$\mathsf{b}(\mu_2 - \mu_1^2)$
Stay/	0	ρ	ρ	μ_1	μ_1	$-c$
Switch	≥ 1	ρ	ρ	μ_1	μ_1	0

When $H = 2$, the only remaining reward difference to calculate is $\Delta r^{1,1}(x) = -c$ for any partner selection rule. For larger values of H, the rewards depend on higher-order moments and can be calculated using the iterative formulas. Taking the sum across these rewards gives ΔG for each rule.

B.2 Moment and Reward Identities

To compute the covariance, note that

$$\Sigma_{CC} = \text{Var}(\Delta\psi_C) = \alpha^2 \sum_{h=0}^{H-1} \text{Var}(U^h) + 2\alpha^2 \sum_{0\le h<k\le H-1} \text{Cov}(U^h, U^k).$$

For each round h, the variance term is

$$\text{Var}(U^h) = x(1-x)^2 S_C^h + x^2(1-x) S_D^h - x^2(1-x)^2 (\Delta G^h)^2$$

where $S_a^h = \mathbb{E}[(R_h - \beta)^2 | a_h = a]$ and $\Delta G^h = G_C^h - G_D^h$. For $h < k$, define the conditional moment

$$M_{a,b}^{h,k} = \mathbb{E}[(R_h - \beta)(R^k - \beta) | a_h = a, a_k = b],$$

then each covariance term is

$$\text{Cov}(U^h, U^k) = x^2(1-x)^2 (M_{C,C}^{h,k} - M_{D,C}^{h,k} - M_{C,D}^{h,k} + M_{D,D}^{h,k} - \Delta G^h \Delta G^k).$$

For action $a \in \{C, D\}$, the second moment of the episodic reward is

$$\begin{aligned} S_a^h &= \mathbb{E}[(R_h - \beta)^2 | a_h = a] \\ &= \text{Var}(R_h - \beta | a) + \mathbb{E}[R_h - \beta | a_h = a]^2 \\ &= \text{Var}(R_h | a_h = a) + (G_a^h - \beta)^2 \\ &= \sum_{l=h}^{H-1} \text{Var}(r_l | a_h = a) + 2 \sum_{h\le l<k} \text{Cov}(r_l, r_k | a_h = a) + (G_a^h - \beta)^2. \end{aligned}$$

Let the opponent's strategy at round l be Y_l, and denote the opponent's and focal agent's actions by ζ_l and z_l, respectively. At each round, the actions are sampled independently from the current policies:

$$\zeta_l \sim \text{Ber}(Y_l), \quad z_l \sim \text{Ber}(x).$$

The focal agent's actions are independently sampled across rounds. The reward at round l is

$$r_l = b\zeta_l + c(1 - z_l)$$

At the conditioned round $l = h$, the action of the focal agent is fixed, so

$$\begin{aligned} \mathbb{E}[r_h | a_h = a] &= b\mathbb{E}[\zeta_h | a_h = a] + c \cdot 1\{a_h = D\} \\ &= bm^h(x) + c \cdot 1\{a_h = D\}. \end{aligned}$$

For all $l > h$, the conditional mean for the focal agent taking action a at time h is

$$\mathbb{E}[\zeta_l | a_h = a] = \mathbb{E}[Y_l | a_h = a] = m_a^{h,l}(x)$$

which means the expected conditional reward is

$$\begin{aligned} \mathbb{E}[r_l | a_h = a] &= b\mathbb{E}[\zeta_l | a_h = a] + c\mathbb{E}[1 - z_l] \\ &= bm_a^{h,l}(x) + c(1 - x) \end{aligned}$$

Likewise, the conditional variance is

$$\begin{aligned} \text{Var}(r_h | a_h = a) &= b^2 \text{Var}(\zeta_h | a_h = a) \\ &= b^2 m^h(x)(1 - m^h(x)), \\ \text{Var}(r_l | a_h = a) &= b^2 \text{Var}(\zeta_l | a_h = a) + c^2 \text{Var}(1 - z_l) \\ &= b^2 m_a^{h,l}(x)(1 - m_a^{h,l}(x)) + c^2 x(1 - x). \end{aligned}$$

The conditional means can be computed using repeated substitution of the iterative partner distribution update. For the covariance terms, let $h \le l < k \le H - 1$. Since the focal action z_k is independent of the previous variables,

$$\begin{aligned} \text{Cov}(r_l, r_k | a_h = a) &= \text{Cov}(b\zeta_l + (1 - z_l)c, b\zeta_k + c(1 - z_k) | a_h = a) \\ &= b^2 \text{Cov}(\zeta_l, \zeta_k | a_h = a) + bc\text{Cov}(Y_k, 1 - z_l | a_h = a). \end{aligned}$$

B.3 Computing the Moments for $H = 2$

For either initial action $a \in \{C, D\}$, the marginal reward variances are

$$\text{Var}(r_0|a_0 = a) = b^2\mu_1(1-\mu_1)$$
$$\text{Var}(r_1|a_0 = a) = b^2\mathbb{E}[Y_1|a_0 = a](1-\mathbb{E}[Y_1|a_0 = a]) + c^2x(1-x)$$

Conditioned on $a_0 = a$,

$$\begin{aligned}\text{Cov}(r_0, r_1|a_0 = a) &= b^2\text{Cov}(\zeta_0, \zeta_1|a_0 = a)\\ &= b^2(\mathbb{E}[\zeta_0\zeta_1|a_0 = a] - \mu_1\mathbb{E}[Y_1|a_0 = a]).\end{aligned}$$

For $R_0 = r_0 + r_1$,

$$\text{Var}(R_0|a_0 = a) = \text{Var}(r_0|a_0 = a) + \text{Var}(r_1|a_0 = a) + 2\text{Cov}(r_0, r_1|a_0 = a)$$

For the second round reward-to-go,

$$G_a^1 = bm^1(x) + c\mathbb{1}[a = D]$$
$$\text{Var}(R_1|a_1 = a) = b^2m^1(x)(1-m^1(x))$$

The between-round covariance is

$$\begin{aligned}M_{a,b}^{0,1} &= \mathbb{E}[(r_0 + r_1 - \beta)(r_1 - \beta)|a_0 = a, a_1 = b]\\ &= \mathbb{E}[r_0r_1|a_0 = a, a_1 = b] + \mathbb{E}[r_1^2|a_0 = a, a_1 = b]\\ &\quad - \beta\mathbb{E}[r_0|a_0 = a, a_1 = b] - 2\beta\mathbb{E}[r_1|a_0 = a, a_1 = b] + \beta^2.\end{aligned}$$

OFT The expected cooperation of the next opponent, conditioned on the current opponent and focal agent's action is

$$\mathbb{E}[Y_1|Y_0, a_0 = C] = Y_0^2 + (1-Y_0)\mu_1,$$
$$\mathbb{E}[Y_1|Y_0, a_0 = D] = \mu_1.$$

Computing the conditional mean we have

$$\mathbb{E}[Y_1|a_0 = C] = \mu_2 + \mu_1 - \mu_1^2,$$
$$\mathbb{E}[Y_1|a_0 = D] = \mu_1.$$

For the joint moment, if $a_0 = C$ and $\zeta_0 = 1$ then the opponent is the same, and

$$\mathbb{E}[\zeta_0\zeta_1|Y_0, a_0 = C] = Y_0^2,$$
$$\mathbb{E}[\zeta_0\zeta_1|a_0 = C] = \mu_2.$$

For $a_0 = D$ and $\zeta_0 = 1$, a new opponent is sampled:

$$\mathbb{E}[\zeta_0\zeta_1|Y_0, a_0 = D] = Y_0\mu_1,$$
$$\mathbb{E}[\zeta_0\zeta_1|a_0 = D] = \mu_1^2.$$

The covariance of the rewards is then

$$\begin{aligned}\text{Cov}(r_0, r_1|a_0 = C) &= b^2[\mu_2 - \mu_1(\mu_2 + \mu_1 - \mu_1^2)]\\ &= b^2(1-\mu_1)(\mu_2 - \mu_1^2),\\ \text{Cov}(r_0, r_1|a_0 = D) &= b^2(\mu_1^2 - \mu_1^2) = 0.\end{aligned}$$

Finally, $m^1(x) = \mu_1 + x(\mu_2 - \mu_1^2)$ which can be substituted into expression for $\text{Var}(R_1|a_1 = a)$.

Table 5: Second moments (S_a^h) of episodic reward when $H = 2$ under OFT.

h	a	$\text{Var}(R_h\|a_h = a)$	$(G_a^h - \beta)^2$
0	C	$b^2[\mu_1(1-\mu_1)+(\mu_2+\mu_1-\mu_1^2)(1-\mu_2-\mu_1+\mu_1^2)$ $+ 2(1-\mu_1)(\mu_2-\mu_1^2)] + c^2x(1-x)$	$[b(\mu_2 + 2\mu_1 - \mu_1^2) + c(1-x) - \beta]^2$
	D	$2b^2\mu_1(1-\mu_1) + c^2x(1-x)$	$[2b\mu_1 + c(2-x) - \beta]^2$
1	C	$b^2(\mu_1 + x(\mu_2-\mu_1^2))(1-\mu_1-x(\mu_2-\mu_1^2))$	$[b(\mu_1 + x(\mu_2-\mu_1^2)) - \beta]^2$
	D	$b^2(\mu_1 + x(\mu_2-\mu_1^2))(1-\mu_1-x(\mu_2-\mu_1^2))$	$[b(\mu_1 + x(\mu_2-\mu_1^2)) + c - \beta]^2$

ROFT The expected cooperation of the next opponent, conditioned on the current opponent and focal agent's action is

$$\mathbb{E}[Y_1|Y_0, a_0 = C] = \mu_1,$$
$$\mathbb{E}[Y_1|Y_0, a_0 = D] = Y_0(1 - Y_0) + Y_0\mu_1.$$

Computing the conditional mean we have

$$\mathbb{E}[Y_1|a_0 = C] = \mu_1,$$
$$\mathbb{E}[Y_1|a_0 = D] = \mu_1 + \mu_1^2 - \mu_2.$$

For the joint moment, if $\zeta_0 = 1$ then the opponent is redrawn regardless of the focal agent's action:

$$\mathbb{E}[\zeta_0\zeta_1|Y_0, a_0 = C] = Y_0\mu_1,$$
$$\mathbb{E}[\zeta_0\zeta_1|a_0 = C] = \mu_1^2.$$

For $a_0 = D$, a new opponent is sampled and so

$$\mathbb{E}[\zeta_0\zeta_1|Y_0, a_0 = D] = Y_0\mu_1,$$
$$\mathbb{E}[\zeta_0\zeta_1|a_0 = D] = \mu_1^2.$$

The covariance of the rewards is then

$$\begin{aligned}\text{Cov}(r_0, r_1|a_0 = C) &= b^2(\mu_1^2 - \mu_1^2) = 0,\\ \text{Cov}(r_0, r_1|a_0 = D) &= b^2(\mu_1^2 - \mu_1(\mu_1 + \mu_1^2 - \mu_2))\\ &= b^2\mu_1(\mu_2 - \mu_1^2).\end{aligned}$$

Finally, $m^1(x) = \mu_1 - (1-x)(\mu_2 - \mu_1^2)$ which can be substituted into expression for $\text{Var}(R_1|a_1 = a)$.

Table 6: Second moments (S_a^h) of episodic reward when $H = 2$ under ROFT.

h	a	$\text{Var}(R_h\|a_h = a)$	$(G_a^h - \beta)^2$
0	C	$2b^2\mu_1(1-\mu_1) + c^2x(1-x)$	$[2b\mu_1 + c(1-x) - \beta]^2$
	D	$b^2[(\mu_1 + \mu_1^2 - \mu_2)(1 - \mu_1 - \mu_1^2 + \mu_2)$ $+ \mu_1(1-\mu_1) + 2\mu_1(\mu_2 - \mu_1^2)] + c^2x(1-x)$	$[b(2\mu_1 + \mu_1^2 - \mu_2) + c(2-x) - \beta]^2$
1	C	$b^2(\mu_1 - (1-x)(\mu_2 - \mu_1^2)) \cdot$ $(1 - \mu_1 + (1-x)(\mu_2 - \mu_1^2))$	$[b(\mu_1 - (1-x)(\mu_2 - \mu_1^2)) - \beta]^2$
	D	$b^2(\mu_1 - (1-x)(\mu_2 - \mu_1^2)) \cdot$ $(1 - \mu_1 + (1-x)(\mu_2 - \mu_1^2))$	$[b(\mu_1 - (1-x)(\mu_2 - \mu_1^2)) + c - \beta]^2$

Stay The opponent is retained regardless of the sampled actions. For $a \in \{C, D\}$,

$$\mathbb{E}[Y_1|Y_0, a_0 = a] = Y_0,$$
$$\mathbb{E}[Y_1|a_0 = a] = \mu_1.$$

For the joint moment, the opponent's actions are independent draws conditional on the fixed strategy, so

$$\mathbb{E}[\zeta_0\zeta_1|Y_0, a_0 = C] = Y_0^2,$$
$$\mathbb{E}[\zeta_0\zeta_1|a_0 = C] = \mu_2.$$

The covariance of the rewards is then

$$\text{Cov}(r_0, r_1|a_0 = a) = b^2(\mu_2 - \mu_1^2)$$

Finally, $m^1(x) = \mu_1$.

Table 7: Second moments (S_a^h) of episodic reward when $H = 2$ under Stay.

h	a	$\text{Var}(R_h\|a_h = a)$	$(G_a^h - \beta)^2$
0	C	$2b^2(\mu_1 + \mu_2 - 2\mu_1^2) + c^2x(1-x)$	$[2b\mu_1 + c(1-x) - \beta]^2$
	D	$2b^2(\mu_1 + \mu_2 - 2\mu_1^2) + c^2x(1-x)$	$[2b\mu_1 + c(2-x) - \beta]^2$
1	C	$b^2\mu_1(1-\mu_1)$	$[b\mu_1 - \beta]^2$
	D	$b^2\mu_1(1-\mu_1)$	$[b\mu_1 + c - \beta]^2$

Switch The opponent is redrawn regardless of the sampled actions. For $a \in \{C, D\}$,

$$\mathbb{E}[Y_1|Y_0, a_0 = a] = \mu_1,$$
$$\mathbb{E}[Y_1|a_0 = a] = \mu_1.$$

For the joint moment, the opponent's actions are independent draws so

$$\mathbb{E}[\zeta_0\zeta_1|Y_0, a_0 = C] = Y_0\mu_1,$$
$$\mathbb{E}[\zeta_0\zeta_1|a_0 = C] = \mu_1^2.$$

The covariance of the rewards is then

$$\text{Cov}(r_0, r_1|a_0 = a) = 0$$

Finally, $m^1(x) = \mu_1$.

Table 8: Second moments (S_a^h) of episodic reward when $H = 2$ under Switch.

h	a	$\text{Var}(R_h\|a_h = a)$	$(G_a^h - \beta)^2$
0	C	$2b^2\mu_1(1-\mu_1) + c^2x(1-x)$	$[2b\mu_1 + c(1-x) - \beta]^2$
	D	$2b^2\mu_1(1-\mu_1) + c^2x(1-x)$	$[2b\mu_1 + c(2-x) - \beta]^2$
1	C	$b^2\mu_1(1-\mu_1)$	$[b\mu_1 - \beta]^2$
	D	$b^2\mu_1(1-\mu_1)$	$[b\mu_1 + c - \beta]^2$

C Proofs of the Main Results

C.1 Conditional Rewards

We begin by analysing the base case of $k = 0$. Defining $\Delta q^{0,h+1}(y) = q^{h+1}(y|x, C^0) - q^{h+1}(y|x, D^0)$, the system reduces to analysing the following recursion:

$$\Delta q^{0,h+1}(y) = xy\Delta q^{0,h}(y) - x\rho(y)\Delta m^{0,h}$$
$$\Delta m^{0,h} = \int_0^1 y\Delta q^{0,h}(y)dy$$

Define the operator $\mathcal{T}$ as the recursive update on a function g. Formally,

$$\mathcal{T}g := Yg - \mathbb{E}[Yg]$$

then $g^{h+1} = \mathcal{T}g^h$. For $h > 1$

$$g^h(y) = yg^{h-1}(y) - \mathbb{E}[Yg^{h-1}(Y)], \quad g^1(y) = y - \mu_1. \tag{10}$$

Then we can isolate the partner dependence y, and from the policy x.

Lemma C.1. *Let $\rho(y)$ be the initial distribution, and g^h be defined by equation* (10)*. Then for all $h \geq 1$, the difference in distribution and mean of the opponent is given by*

$$\Delta q^{0,h}(y) = x^{h-1}g^h(y)\rho(y), \quad \Delta m^{0,h} = x^{h-1}\mathbb{E}[Yg^h(Y)].$$

Proof. We proceed by induction. Let $h = 1$, then

$$\Delta q^{0,1}(y) = (y - \mu_1)\rho(y) = x^0 g^1(y)\rho(y)$$
$$\Delta m^{0,1} = \text{Var}_\rho(Y) = \mathbb{E}[Y^2 - \mu_1 Y] = x^0\mathbb{E}[Yg^1(Y)]$$

Assume the form holds at time step h, then for $h + 1$,

$$\begin{aligned}
\Delta q^{0,h+1}(y) &= xy\Delta q^{0,h}(y) - x\rho(y)\Delta m^{0,h} \\
&= xy\Big(x^{h-1}g^h\rho(y)\Big) - x\rho(y)x^{h-1}\mathbb{E}[Yg^h(Y)] \\
&= x^h\Big(yg^h - \mathbb{E}[Yg^h(Y)]\Big)\rho(y) \\
&= x^h g^{h+1}(y)\rho(y) \\
\Delta m^{0,h+1} &= x^h\int_0^1 yg^{h+1}(y)\rho(y)dy \\
&= x^h\mathbb{E}[Yg^{h+1}(Y)]
\end{aligned}$$

□

This form enables us to provide a simple condition for the future rewards to be non-negative. Specifically, for $\Delta m^{0,h} \geq 0$ we need $\mathbb{E}[Yg^h(Y)] \geq 0$. This can be analysed through the properties of the operator $\mathcal{T}$, by splitting the steps of the update. Define the inner product of two functions as

$$\langle f, g \rangle := \mathbb{E}[Yf(Y)g(Y)]. \tag{11}$$

Lemma C.2. *The operator $\mathcal{T}$ is self-adjoint with respect to the inner product in equation* (11)*.*

Proof.

$$\begin{aligned}
\langle \mathcal{T}f, g \rangle &= \langle Yf - \mathbb{E}[Yf], g \rangle \\
&= \mathbb{E}[Y(Yf - \mathbb{E}[Yf])g] \\
&= \mathbb{E}[Y^2 fg] - \mathbb{E}[Yf]\mathbb{E}[Yg] \\
&= \mathbb{E}[Yf(Yg - \mathbb{E}[Yg])] \\
&= \langle f, Yg - \mathbb{E}[Yg] \rangle \\
&= \langle f, \mathcal{T}g \rangle.
\end{aligned}$$

□

Lemma C.3. *Let g^h be defined by equation* (10)*. For any $m, n \in \mathbb{N}$,*

$$\mathbb{E}[Yg^m g^n] = \mathbb{E}[Yg^{m+n}].$$

Proof. By Lemma C.2, we have

$$\begin{aligned}
\langle g^m, g^n \rangle &= \langle \mathcal{T}^{m-1} g^1, \mathcal{T}^{n-1} g^1 \rangle \\
&= \langle g^1, \mathcal{T}^{m+n-2} g^1 \rangle \\
&= \langle g^1, g^{m+n-1} \rangle
\end{aligned}$$

and using the definition of the inner product, for any $k \in \mathbb{N}$

$$\begin{aligned}
\langle g^1, g^k \rangle &= \mathbb{E}[Yg^1 g^k] \\
&= \mathbb{E}[Y(Y - \mu_1)g^k] \\
&= \mathbb{E}[Y^2 g^k] - \mu_1 \mathbb{E}[Yg^k]
\end{aligned}$$

From the recursion, we can take expectation over Yg^{k+1}, giving the identity

$$\mathbb{E}[Yg^{k+1}] = \mathbb{E}[Y^2 g^k] - \mu_1 \mathbb{E}[Yg^k]$$

Therefore,

$$\langle g^1, g^k \rangle = \mathbb{E}[Yg^1 g^k] = \mathbb{E}[Yg^{k+1}]$$

Finally,

$$\mathbb{E}[Yg^m g^n] = \langle g^m, g^n \rangle = \langle g^1, g^{m+n-1} \rangle = \mathbb{E}[Yg^{m+n}].$$

□

Lemma C.4. *Under the OFT rule, for all $h \geq 1$ the reward difference $\Delta r^{0,h} = b\Delta m^{0,h} \geq 0$.*

Proof. We will separate into odd and even cases. First, let $m = n = l$ in Lemma C.3, then

$$\Delta m^{0,2l} = x^{2l-1}\mathbb{E}[Yg^{2l}] = x^{2l-1}\mathbb{E}[Yg^l g^l] = x^{2l-1}\mathbb{E}[Y(g^l)^2] \geq 0$$

Likewise, let $m = l$ and $n = l + 1$. By Lemma C.3

$$\begin{aligned}
\Delta m^{0,2l+1} &= x^{2l}\mathbb{E}[Yg^{2l+1}] \\
&= x^{2l}\mathbb{E}[Yg^l g^{l+1}] \\
&= x^{2l}\mathbb{E}[Yg^l(Yg^l - \mathbb{E}[Yg^l])] \\
&= x^{2l}\Big(\mathbb{E}[(Yg^l)^2] - \mathbb{E}[Yg^l]^2\Big) \geq 0
\end{aligned}$$

Hence, since $b > 0$, $\Delta r^{0,h} \geq 0$ for all $h \geq 1$. □

Here, we show how the same result applies to ROFT.

Lemma C.5. *Under the ROFT rule, for all $h \geq 1$ the reward difference $\Delta r^{0,h} = b\Delta m^{0,h} \geq 0$.*

Proof. Under the ROFT partner selection rule,

$$\Delta q^{0,h+1}(y) = (1-x)(1-y)\Delta q^{0,h}(y) + (1-x)\rho(y)\Delta m^{0,h}(x)$$

and using the same inductive proof in Lemma C.1, we get

$$\begin{aligned}\Delta q^{0,h}(y) &= (1-x)^{h-1} f^h(y)\rho(y)\\ \Delta m^{0,h} &= (1-x)^{h-1}\mathbb{E}[Y f^h(Y)]\end{aligned}$$

where $f^h(y) = (1-y)f^{h-1}(y) + \mathbb{E}[Y f^{h-1}(Y)]$, with $f^1(y) = y - \mu_1$ and $\mathbb{E}[f^h(Y)] = 0$. This means we can write

$$f^h(y) = (1-y)f^{h-1}(y) - \mathbb{E}[(1-Y)f^{h-1}(Y)].$$

Now consider the transformation $Z = 1 - Y$, and the function $\phi^h(z) = -f^h(1-z)$. Then,

$$\phi^h(z) = z\phi^{h-1}(z) - \mathbb{E}[Z\phi^{h-1}(Z)], \quad \phi^1 = z - \mathbb{E}[Z]$$

this is the exact same recursion defined for the OFT partner selection rule, and therefore the result of Lemma C.4 follows. Concretely,

$$\begin{aligned}\Delta m^{0,h} &= (1-x)^{h-1}\mathbb{E}[Y f^h(Y)]\\ &= (1-x)^{h-1}\mathbb{E}[Z\phi^h(Z)] \geq 0.\end{aligned}$$

Since $b > 0$, the result follows. □

We now consider the general case of conditioning on the action at step k. Note that by dividing through by $\rho(y)$, the one-step update for OFT becomes

$$\varrho^{h+1}(y|x) = xy\varrho^h(y|x) + (1 - xm^h(x)), \quad m^h(x) = \mathbb{E}[Y\varrho^h(Y|x)]$$

This distribution can then be expressed in terms of the recursive function, $g^h(y)$.

Lemma C.6. *For all h, the density ratio $\varrho^h(y|x)$ is given by*

$$\varrho^h(y|x) = 1 + \sum_{j=1}^{h} x^j g^j(y)$$

Proof. We proceed by induction. Let $h = 0$, then

$$\varrho^0(y|x) = \frac{q^0(y|x)}{\rho(y)} = 1.$$

Assume that the statement holds for $h = k$. Then for $h = k+1$,

$$\begin{aligned}\varrho^{k+1}(y|x) &= xy\varrho^k(y|x) + 1 - xm^k(x)\\ &= xy\Big(1 + \sum_{j=1}^{k} x^j g^j(y)\Big) + 1 - x\mathbb{E}\Big[Y\Big(1 + \sum_{j=1}^{k} x^j g^j(Y)\Big)\Big]\\ &= 1 + x(y - \mu_1) + \sum_{j=1}^{k} x^{j+1}\Big(yg^j(y) - \mathbb{E}[Y g^j(Y)]\Big)\\ &= 1 + xg^1(y) + \sum_{j=1}^{k} x^{j+1} g^{j+1}(y)\\ &= 1 + \sum_{j=1}^{k+1} x^j g^j(y)\end{aligned}$$

□

Proof of Proposition 3.1

Proof. Consider the kth step of the REINFORCE algorithm. Then the subsequent update under the OFT rule is,

$$\begin{aligned}
\Delta q^{k,k+1} &= yq^k(y|x) - m^k(x)\rho(y) \\
g^{k,k+1} &= y\varrho^k(y|x) - m^k(x) \\
&= y\left(1 + \sum_{j=1}^{k} x^j g^j(y)\right) - \left(\mu_1 + \sum_{j=1}^{k} x^j \mathbb{E}[Y g^j(Y)]\right) \\
&= y - \mu_1 + \sum_{j=1}^{k} x^j \left(y g^j(y) - \mathbb{E}[Y g^j(Y)]\right) \\
&= y - \mu_1 + \sum_{j=1}^{k} x^j g^{j+1}(y) \\
&= \sum_{j=0}^{k} x^j g^{j+1}(y)
\end{aligned}$$

Then for $h > k+1$, we have

$$\Delta q^{k,h+1}(y) = xy\Delta q^{k,h}(y) - x\rho(y)\Delta m^{k,h}$$

and therefore the same operator $\mathcal{T}$ applies. Consequently,

$$\begin{aligned}
g^{k,h}(y) &= \mathcal{T}^{h-k-1} g^{k,k+1}(y) \\
&= \sum_{j=0}^{k} x^j \mathcal{T}^{h-k-1} g^{j+1}(y) \\
&= \sum_{j=0}^{k} x^j g^{h-k+j}(y)
\end{aligned}$$

The difference in the conditional means at this step is then

$$\begin{aligned}
\Delta m^{k,h} &= x^{h-k-1}\mathbb{E}[Y g^{k,h}(Y)] \\
&= x^{h-k-1}\mathbb{E}[Y \sum_{j=0}^{k} x^j g^{h-k+j}(Y)] \\
&= \sum_{j=0}^{k} x^{h-k+j-1}\mathbb{E}[Y g^{h-k+j}(Y)] \\
&= \sum_{j=0}^{k} \Delta m^{0,h-k+j} \geq 0
\end{aligned}$$

where the final line follows by Lemma C.4. Note that the proof has an identical structure for the ROFT rule, invoking Lemma C.5 in the final step. □

Proof of Corollary 3.1

Proof. For each step k of the REINFORCE algorithm, we obtain a reward difference of $\Delta r^{k,k} = -c$. The future rewards must compensate for this. Consider the actions conditioned at step k. By the proof of Proposition 3.1, at step $k+1$

$$\begin{aligned}
\Delta r^{k,k+1}(x) &= b\Delta m^{k,k+1}(x) \\
&= b\sum_{j=0}^{k} \Delta m^{0,j+1} \\
&\geq b\Delta m^{0,1}.
\end{aligned}$$

Since we have $\Delta m^{0,1} = \text{Var}_\rho(X)$, and for all $h > k+1$ the reward difference $\Delta r^{k,h}(x) \geq 0$, then

$$\begin{aligned}
\sum_{h=0}^{H-1} \Delta G^h[\rho] &= \sum_{k=0}^{H-1} \sum_{h=k}^{H-1} \Delta r^{k,h} \\
&\geq \sum_{k=0}^{H-2} \Big(\Delta r^{k,k} + \Delta r^{k,k+1} \Big) + \Delta r^{H-1,H-1} \\
&\geq \sum_{k=0}^{H-2} \Big(-c + b\text{Var}_\rho(X) \Big) - c \\
&\geq (H-1)(b\text{Var}_\rho(X) - c) - c.
\end{aligned}$$

If $\Delta G[\rho] = b\text{Var}_\rho(X) - 2c > 0$, the final bound is positive for every $H \geq 2$. □

C.2 Mean Dynamics

Proof of Theorem 3.1

Proof. The characteristic flow is generated by

$$\frac{d}{dt}\Phi_t(x) = -2\alpha H c \Phi_t^2(x)(1-\Phi_t(x))^2, \quad \Phi_0(x) = x_0,$$

This is a separable ODE, which integrating gives

$$\begin{aligned}
\int \frac{1}{z^2(1-z)^2} dz &= \int -2\alpha H c \, dt \\
\frac{1}{1-z} - \frac{1}{z} + 2\ln|\frac{z}{1-z}| &= -2\alpha H c t + C.
\end{aligned}$$

Let

$$F(x) = \frac{1}{1-x} - \frac{1}{x} + 2\ln|\frac{x}{1-x}|$$

then the solution at time t is given by

$$F(\Phi_t) = F(x) - 2\alpha H c t.$$

For $c > 0$ and $\alpha > 0$, then as $t \to \infty$ we have $F(\Phi_t) \to -\infty$ for every $x \in (0,1)$. Since $F(x) \to -\infty$ as $x \to 0$, we have that $\Phi_t \to 0$. Since the velocity field $u(x) = -2\alpha H c x^2(1-x)^2 \in C^1([0,1])$, the solution to the continuity equation is given by the pushforward of the initial point under the characteristic flow:

$$\rho(t,\cdot) = (\Phi_t)_\# \rho_0.$$

Then for any bounded test function $\varphi \in C([0,1])$,

$$\int_0^1 \varphi(y)\rho(t,y)\, dy = \int_0^1 \varphi(\Phi_t)\rho_0(x) dx$$

Since $X(t,x) \to 0$ pointwise for all $x \in (0,1)$, then $\varphi(\Phi_t) \to \varphi(0)$. Moreover,

$$|\varphi(\Phi_t)\rho_0(x)| \leq \|\varphi\|_\infty |\rho_0(x)|$$

By the dominated convergence theorem,

$$\int_0^1 \varphi(\Phi_t)\rho_0(x) dx \to \varphi(0) \int_0^1 \rho_0(x) dx = \varphi(0)$$

which means

$$\int_0^1 \varphi(y)\rho(t,y)\, dy \to \varphi(0)$$

and therefore $\rho(t) \rightharpoonup \delta_0$. □

Note that for fixed H, each action and partner-selection history has a polynomial probability. Integrating over the oponent draws yields a product of moments of at most H. As such, write

$$\Delta G[\rho](x) = P(x, \mu_1(\rho), \dots, \mu_H(\rho)),$$

where P is a polynomial. We can now choose a constant C_H such that

$$\sup_{[0,1]^{H+1}} |P| \leq C_H, \qquad \sup_{[0,1]^{H+1}} \|\nabla P\|_\infty \leq C_H.$$

Lemma C.7. *The non-local advection is spatially Lipschitz and sub-linear. That is, there exists an $L_1, M \in \mathbb{R}$ such that*

$$|u[\rho](x) - u[\rho](y)| \leq L_1|x-y|, \qquad |u[\rho](x)| \leq M(1+|x|).$$

Proof. The spatial derivative is

$$\partial_x u[\rho](x) = 4\alpha x(1-x)(1-2x)\Delta G[\rho](x) + 2\alpha x^2(1-x)^2\partial_x \Delta G[\rho](x).$$

Therefore,

$$\|\partial_x u[\rho]\|_\infty \leq \alpha C_H + \alpha C_H/8 = \frac{9\alpha C_H}{8}.$$

By the mean value theorem,

$$\begin{aligned} |u[\rho](x) - u[\rho](y)| &\leq \sup_z |\partial_x u[\rho](z)|\ |x-y| \\ &\leq \frac{9\alpha C_H}{8}|x-y| \end{aligned}$$

The condition for sub-linearity follows from

$$\begin{aligned} |u[\rho](x)| &= |2\alpha \Delta G[\rho](x) x^2(1-x)^2| \\ &\leq \frac{\alpha C_H}{8} \\ &\leq \frac{\alpha C_H}{8}(1+|x|). \end{aligned}$$

□

Lemma C.8. *There exists $L_2 > 0$ such that for every $\rho_1, \rho_2 \in \mathcal{P}$,*

$$\|u[\rho_1](\cdot) - u[\rho_2](\cdot)\|_{\mathcal{C}([0,1])} \leq L_2 W_1(\rho_1, \rho_2).$$

Proof.

$$\begin{aligned} |u[\rho_1](x) - u[\rho_2](x)| &= |2\alpha\Delta G[\rho_1]x^2(1-x)^2 - 2\alpha\Delta G[\rho_2]x^2(1-x)^2| \\ &= 2\alpha|x^2(1-x)^2(\Delta G[\rho_1] - \Delta G[\rho_2])| \\ &\leq \frac{\alpha}{8}|\Delta G[\rho_1](x) - \Delta G[\rho_2](x)| \end{aligned}$$

Therefore,

$$\|u[\rho_1] - u[\rho_2]\|_\infty \leq \frac{\alpha}{8}\|\Delta G[\rho_1] - \Delta G[\rho_2]\|_\infty.$$

It suffices to show that

$$\|\Delta G[\rho_1] - \Delta G[\rho_2]\|_\infty \leq C W_1(\rho_1, \rho_2)$$

for some constant $C \in \mathbb{R}$. Using the Kantorovich-Rubinstein duality [48], the Wasserstein distance with $p = 1$ can be expressed as

$$W_1(\rho_1, \rho_2) = \sup\left\{ \int_{[0,1]} f(x) d(\rho_1 - \rho_2) \;\middle|\; f \in \mathcal{C}^1,\ f : [0,1] \to \mathbb{R},\ \text{Lip}(f) \leq 1 \right\}.$$

For any integer ℓ, the ℓth moment is

$$\mu_\ell = \int_0^1 x^\ell d\rho(x)$$

Let $f(x) = \frac{x^i}{i}$, then f is continuous and has Lipschitz constant 1. By the Wasserstein metric,

$$|\mu_\ell(\rho_1) - \mu_\ell(\rho_2)| \le \ell W_1(\rho_1, \rho_2).$$

Therefore,

$$\begin{aligned}\|\Delta G[\rho_1] - \Delta G[\rho_2]\|_\infty &\le C_H \sum_{\ell=1}^{N} |\mu_\ell(\rho_1) - \mu_\ell(\rho_2)| \\ &\le \frac{C_H H(H+1)}{2} W_1(\rho_1, \rho_2)\end{aligned}$$

Therefore the result holds with $L_2 = \frac{\alpha C_H H(H+1)}{16}$. □

C.2.1 Proof of Proposition 3.2

Proof. For any fixed finite H, Lemmas C.7 and C.8 imply the velocity field $u[\rho]$ is uniformly Lipschitz continuous in space, Lipschitz continuous in the distribution with respect to W_1, and bounded. The conditions in [6, Theorem 2] are satisfied, giving a unique soluion represented by the characteristic flow

$$\rho(t) = (\Phi_t)_\# \rho_0, \quad \frac{d}{dt}\Phi_t(x) = u[\rho_t](\Phi_t(x)).$$

Since $u[\rho](0) = u[\rho](1) = 0$, the flow remains in the domain $[0, 1]$. □

For $H = 2$, the characteristic ODE admits a solution Φ_t

$$\begin{aligned}\frac{d}{dt}\Phi_t &= 2\alpha \Delta G[\rho] \Phi_t^2 (1 - \Phi_t)^2 \\ \int \frac{1}{z^2(1-z)^2} dz &= 2\alpha \int \Delta G[\rho] dt \\ F(\Phi_t) &= F(x) + 2\alpha K(t)\end{aligned}$$

where $F(x)$ is defined in the proof of Theorem 3.1. Clearly, the time evolution is uniquely determined by K. Since $F'(x) > 0$ for all $x \in (0, 1)$, F is strictly increasing and therefore invertible on the domain. Therefore the solution can be expressed by

$$\Phi_K(x) = F^{-1}(F(x) + 2\alpha K).$$

The next result proves that under this flow, the cooperation levels increase when the initial population has sufficient variance.

Proof of Theorem 3.2

Proof. Define the autonomous equations

$$\begin{aligned}K' &= h(K), \\ h(K) &= b\text{Var}_{(\Phi_K)_\# \rho_0}(X) - 2c\end{aligned}$$

The initial condition means

$$h(0) = b\text{Var}_{\rho_0}(X) - 2c > 0$$

Note that for every $x > 0$

$$\lim_{K \to \infty} \Phi_K(x) \to 1, \quad \partial_K \Phi_K = 2\alpha \Phi_K^2 (1 - \Phi_K)^2 \le \alpha/8,$$

therefore Φ_K is uniformly Lipschitz in K and bounded on $[0, 1]$. In particular,

$$|\Phi_{K_1}(x) - \Phi_{K_2}(x)| \le \frac{\alpha}{8} |K_1 - K_2|, \quad \forall x \in [0, 1].$$

For $X \sim \rho_0$, the first moment is

$$\begin{aligned}|\mu_1(K_1) - \mu_1(K_2)| &= |\mathbb{E}[\Phi_{K_1}(X)] - \mathbb{E}[\Phi_{K_2}(X)]| \\ &\leq \mathbb{E}[|\Phi_{K_1}(X) - \Phi_{K_2}(X)]| \\ &\leq \frac{\alpha}{8}|K_1 - K_2|.\end{aligned}$$

It follows that the second moment is Lipschitz:

$$\begin{aligned}|\mu_2(K_1) - \mu_2(K_2)| &= |\mathbb{E}[\Phi^2_{K_1}(X)] - \mathbb{E}[X^2_{K_2}(X)]| \\ &= |\mathbb{E}[\Phi^2_{K_1}(X) - X^2_{K_2}(X)]| \\ &\leq \mathbb{E}[|\Phi^2_{K_1}(X) - X^2_{K_2}(X)|] \\ &\leq \mathbb{E}[2|\Phi_{K_1}(X) - \Phi_{K_2}(X)|] \\ &\leq \frac{\alpha}{4}|K_1 - K_2|\end{aligned}$$

Therefore the mean and variance are Lipschitz continuous in K, which means there is a unique solution to the IVP $K' = h(K), K(0) = 0$. Under the limit as $K \to \infty$, $\mathbb{E}[\Phi^2_K(X)] \to 1$, so $\text{Var}(K) = \mathbb{E}[\Phi^2_K(X)] - \mathbb{E}[\Phi_K(X)]^2 \to 0$. The left and right limits of h are

$$h(0) > 0, \qquad \lim_{K\to\infty} h(K) = -2c < 0,$$

so by the intermediate value theorem (h is Lipschitz continuous), there exists a $K^* = \inf\{K > 0 \,|h(K) = 0\}$. For $0 \leq K < K^*,\ h(K) > 0$ therefore K is increasing. Therefore, $K \to K^*$. □

Proof of Proposition 4.1

Proof. Let $f \in C^2([0,1])$ be an arbitrary test function, then

$$\frac{d}{dt}\mathbb{E}[f(X_t)] = \mathbb{E}[\mathcal{L}f(X_t)]$$

where $\mathcal{L}$ is the infinitesimal generator given by

$$\mathcal{L}f(x) = A_\rho(t,x)f'(x) + \frac{1}{2}B^2_\rho(t,x)f''(x).$$

In particular, letting $f(x) = x$ gives the evolution of the mean as

$$\frac{d}{dt}\mathbb{E}[X_t] = \mathbb{E}[A_\rho(t,x)]$$

which gives the ODE

$$\frac{d}{dt}\mu_1(t) = \int_0^1 A_\rho(t,x)\rho(t,x)dx = 2\alpha\int_0^1 x^2(1-x)^2\Delta G[\rho]\rho(t,x)dx + O(\alpha^2).$$

To show that this is increasing, we begin by showing the $O(\alpha^2)$ term is uniformly bounded by a constant C. Each variance term is bounded by

$$\text{Var}(U^h) \leq \mathbb{E}[(U^h)^2] \leq (H(b+c) + |\beta|)^2$$

and the covariance is bounded by

$$|\text{Cov}(U^h, U^k)| \leq \sqrt{\text{Var}(U^h)\text{Var}(U^k)} \leq (H(b+c) + |\beta|)^2.$$

Therefore, we have

$$\begin{aligned}|\Sigma_{CC}| &= \left|\alpha^2\sum_{h=0}^{H-1}\text{Var}(U^h) + 2\alpha^2\sum_{h<k}^{H-1}\text{Cov}(U^h, U^k)\right| \\ &\leq \alpha^2(H + H(H-1))(H(b+c) + |\beta|)^2 \\ &= \alpha^2H^2(H(b+c) + |\beta|)^2\end{aligned}$$

This gives a simple bound on the $O(\alpha^2)$ term as

$$|2x(1-x)(1-2x)\Sigma_{CC}| \leq \frac{1}{4}\alpha^2 H^2 (H(b+c)+|\beta|)^2.$$

Define

$$I(t) = \int_0^1 x^2(1-x)^2 \Delta G[\rho]\rho(t,x)dx$$

and let $\alpha^* = \frac{4I(0)}{H^2(H(b+c)+|\beta|)^2}$. Since $I(t)$ is continuous and $I(0) > 0$, then there exists a T such that

$$I(t) > I(0)/2 > 0 \qquad \forall t \in [0,T].$$

Hence on the interval $[0,T]$,

$$\begin{aligned}\frac{d}{dt}\mu_1(t) &= 2\alpha I(t) + O(\alpha^2) \\ &\geq \alpha I(0) - \frac{1}{4}\alpha^2 H^2 (H(b+c)+|\beta|)^2 \\ &> 0\end{aligned}$$

for all $\alpha < \alpha^*$. □

C.3 Stochastic Dynamics

A major obstacle in proving such an equation is well-defined is the behaviour at the boundary. At $x = 0$ and $x = 1$, the noise term $B^2(x) = 0$ which can prevent using standard techniques. We first consider an $\varepsilon-$regularised version, then extend the existence as $\varepsilon \to 0$. Fix $\eta \in L^\infty([0,1])$ as a trial density such that $A_\eta(x)$ and $B^2_\eta(x)$ are functions of x and η only. The regularised versions are $A^\varepsilon_\eta(x) = A_\eta(x)$ and $B^{2,\varepsilon}_\eta(x) = B^2_\eta(x) + \varepsilon$. Define the set

$$S^\varepsilon = \{\eta \in C([0,1]) : \eta \geq 0, \int_0^1 \eta(x)dx = 1, \|\eta\|_\infty \leq M^\varepsilon\}$$

For a given $\eta \in S^\varepsilon$, the unique solution $\mathcal{F}^\varepsilon[\eta](x)$ to the linear ODE is given by

$$\mathcal{F}^\varepsilon[\eta](x) = \frac{w^\varepsilon_\eta(x)}{\int_0^1 w^\varepsilon_\eta(x)dx}, \qquad w^\varepsilon_\eta(x) = \frac{1}{B^{2,\varepsilon}_\eta(x)} \exp(\int_0^x \frac{2A^\varepsilon_\eta(y)}{B^{2,\varepsilon}_\eta(y)}dy)$$

By updating the value of η using the operator $\mathcal{F}$, we aim to converge to a fixed point of the system, which will correspond to a stationary distribution of the PDE in (8). The required regularity of the operator and space is proven below. We begin by using the following Lipschitz continuity result on the moments:

$$|m_i(\eta) - m_i(\nu)| = |\int_0^1 y^i d(\eta-\nu)(y)| \leq iW_1(\eta,\nu).$$

Next, we use this moment regularity to show the conditional reward terms are also Lipschitz.

Lemma C.9. *The drift A^ε_η and diffusion $B^{2,\varepsilon}_\eta$ are Lipschitz continuous in both x and η. Specifically,*

$$\|A^\varepsilon_\eta - A^\varepsilon_\nu\|_\infty \leq L_A W_1(\eta,\nu), \quad \|B^{2,\varepsilon}_\eta - B^{2,\varepsilon}_\nu\|_\infty \leq L_B W_1(\eta,\nu),$$

and

$$|A^\varepsilon_\eta(x) - A^\varepsilon_\eta(y)| \leq L_A|x-y|, \quad |B^{2,\varepsilon}_\eta(x) - B^{2,\varepsilon}_\eta(y)| \leq L_B|x-y|,$$

where $L_A, L_B \in \mathbb{R}^+$ and are independent of ε.

Proof. Fix H. Summing over the finitely many actions and partner-selection histories expresses the conditional reward moments as polynomials in x and $\mu_1(\eta),\ldots,\mu_H(\eta)$. Therefore, $\Delta G[\rho](x)$ and $\Sigma_{CC}[\eta](x)$ are polynomials in x and $\mu_1(\eta),\ldots,\mu_H(\eta)$. Since the product of polynomials is a polynomial, we have A^ε_η and $B^{2,\varepsilon}_\eta$ are both polynomials. For either polynomial f, define

$$L_f = \sup_{[0,1]^{H+1}} \|\nabla f\|_\infty < \infty.$$

Hence,

$$|f(x,\mu_1(\eta),\dots,\mu_H(\eta)) - f(x,\mu_1(\nu)\dots,\mu_H(\nu))| \le L_f \sum_{\ell=1}^{H} |\mu_\ell(\eta) - \mu_\ell(\nu)|$$
$$\le \frac{H(H+1)}{2} L_f W_1(\eta,\nu).$$

For Lipschitz continuity in x, note that since f is a polynomial, its derivative is bounded. In particular,

$$\sup_{[0,1]^{H+1}} |\partial_x f| < \infty$$

and therefore for any η,

$$|f(x,\mu_1(\eta),\dots,\mu_H(\eta)) - f(y,\mu_1(\eta),\dots,\mu_H(\eta))| \le \sup_{[0,1]^{H+1}} |\partial_x f||x-y|$$

so A_η is Lipschitz in x. Taking L_A as the maximum of each Lipschitz bound gives the result. □

Note that since the drift and diffusion are Lipschitz and the domain is bounded, then both the drift and diffusion are bounded above such that $|A_\eta(x)| \le M_A$ and $0 \le B^2_\eta(x) \le M_B$. The next result shows the iterative operator is well-defined.

Lemma C.10. *The operator $\mathcal{F}^\varepsilon[\eta] : S^\varepsilon \to S^\varepsilon$ is well-defined.*

Proof. Note that

$$\left|\frac{2A^\varepsilon_\eta(x)}{B^{2,\varepsilon}_\eta(x)}\right| \le \left|\frac{2A^\varepsilon_\eta(x)}{\varepsilon}\right| \le \frac{2M_A}{\varepsilon}$$

so for every $x \in [0,1]$

$$\left|\int_0^x \frac{2A^\varepsilon_\eta(y)}{B^{2,\varepsilon}_\eta(y)}dy\right| \le \frac{2M_A}{\varepsilon}$$

and

$$e^{-\frac{2M_A}{\varepsilon}} \le \exp\left(\int_0^x \frac{2A^\varepsilon_\eta(y)}{B^{2,\varepsilon}_\eta(y)}dy\right) \le e^{\frac{2M_A}{\varepsilon}}$$

This means

$$0 < w^\varepsilon_\eta(x) \le \frac{e^{\frac{2M_A}{\varepsilon}}}{\varepsilon}$$

We have that $w^\varepsilon_\eta(x) \in C[0,1]$, $w^\varepsilon_\eta(x) > 0$ and the integral across the domain is finite and strictly positive. Moreover,

$$\int_0^1 \mathcal{F}^\varepsilon[\eta](x)dx = 1.$$

Therefore $\mathcal{F}^\varepsilon : S^\varepsilon \to S^\varepsilon$ is a well-defined mapping. Since $B^{2,\varepsilon}_\eta(x) \le M_B + \varepsilon$, we also have the bound

$$\frac{1}{M_B+\varepsilon}e^{-\frac{2M_A}{\varepsilon}} \le w^\varepsilon_\eta(x) \le \frac{e^{\frac{2M_A}{\varepsilon}}}{\varepsilon}$$
$$\frac{1}{M_B+\varepsilon}e^{-\frac{2M_A}{\varepsilon}} \le \int_0^1 w^\varepsilon_\eta(x)dx \le \frac{e^{\frac{2M_A}{\varepsilon}}}{\varepsilon}$$

hence $0 \le \mathcal{F}^\varepsilon[\eta](x) \le M^\varepsilon$ for some constant M^ε which is independent of η. □

Lemma C.11. *The set $\mathcal{F}^\varepsilon(S) := \{\mathcal{F}^\varepsilon[\eta] : \eta \in S^\varepsilon\} \subset C([0,1])$ is relatively compact.*

Proof. From the Arzela-Ascoli Theorem [38], it suffices to show that the set $\mathcal{F}^\varepsilon(S^\varepsilon)$ is equibounded and equicontinuous. Firstly, by Lemma C.10, the set is uniformly bounded by M^ε. Now let

$$\psi_\eta^\varepsilon(x) = \int_0^x \frac{2A_\eta^\varepsilon(y)}{B_\eta^{2,\varepsilon}(y)} dy$$

Then,

$$\begin{aligned} |\partial_x \psi_\eta^\varepsilon(x)| &= \left| \frac{2A_\eta^\varepsilon(x)}{B_\eta^{2,\varepsilon}(x)} \right| \\ &\leq \frac{2M_A}{\varepsilon} \end{aligned}$$

To pass the bound to the operator, we use that $B_\eta^{2,\varepsilon}(x)$ is uniformly Lipschitz in x and has a lower bound ε to get

$$|\partial_x \log(B_\eta^{2,\varepsilon})| = |\frac{\partial_x B_\eta^{2,\varepsilon}(x)}{B_\eta^{2,\varepsilon}(x)}| \leq \frac{L_B}{\varepsilon}.$$

then since $w_\eta^\varepsilon(x) = e^{\psi_\eta^\varepsilon(x)}/B_\eta^{2,\varepsilon}(x)$,

$$\log w_\eta^\varepsilon(x) = \psi_\eta^\varepsilon(x) - \log(B_\eta^{2,\varepsilon}(x))$$

has a uniform Lipschitz bound. By Lemma C.10, the normalisation is bounded away from zero. Therefore, $\mathcal{F}^\varepsilon[\eta](x)$ has a uniform Lipschitz bound. Hence, $\mathcal{F}^\varepsilon[\eta](x)$ is uniformly bounded and equicontinuous and therefore its image is relatively compact in $C([0,1])$. □

Proposition C.1. *The mapping $\mathcal{F}^\varepsilon : S^\varepsilon \to S^\varepsilon$ is Lipschitz continuous.*

Proof. Consider the integrand

$$\begin{aligned} \left| \frac{2A_\eta^\varepsilon}{B_\eta^2 + \varepsilon} - \frac{2A_\nu^\varepsilon}{B_\nu^2 + \varepsilon} \right| &\leq \left| \frac{2(A_\eta^\varepsilon - A_\nu^\varepsilon)}{B_\eta^2 + \varepsilon} \right| + \frac{2|A_\nu^\varepsilon||B_\eta^2 - B_\nu^2|}{|(B_\eta^2 + \varepsilon)(B_\nu^2 + \varepsilon)|} \\ &\leq \frac{2|A_\eta^\varepsilon - A_\nu^\varepsilon|}{\varepsilon} + \frac{2M_A|B_\eta^2 - B_\nu^2|}{\varepsilon^2} \\ &\leq (\frac{2L_A}{\varepsilon} + \frac{2M_A L_B}{\varepsilon^2}) W_1(\eta, \nu) \end{aligned}$$

where in the final line we take the supremum over x. This implies

$$\|\psi_\eta^\varepsilon - \psi_\nu^\varepsilon\|_\infty \leq (\frac{2L_A}{\varepsilon} + \frac{2M_A L_B}{\varepsilon^2}) W_1(\eta, \nu)$$

Now take

$$\begin{aligned} |w_\eta^\varepsilon(x) - w_\nu^\varepsilon(x)| &= \left| \frac{\exp(\psi_\eta^\varepsilon)}{B_\eta^{2,\varepsilon}} - \frac{\exp(\psi_\nu^\varepsilon)}{B_\nu^{2,\varepsilon}} \right| \\ &\leq \left| \frac{\exp(\psi_\eta^\varepsilon) - \exp(\psi_\nu^\varepsilon)}{B_\eta^{2,\varepsilon}} \right| + \frac{|\exp(\psi_\nu^\varepsilon)||B_\eta^{2,\varepsilon} - B_\nu^{2,\varepsilon}|}{|(B_\eta^{2,\varepsilon})(B_\nu^{2,\varepsilon})|} \\ &\leq \frac{\exp(\max\{\psi_\eta^\varepsilon, \psi_\nu^\varepsilon\})|\psi_\eta^\varepsilon - \psi_\nu^\varepsilon|}{\varepsilon} + \frac{\exp(\psi_\nu^\varepsilon) L_B W_1(\eta, \nu)}{\varepsilon^2} \\ &\leq e^{\frac{2M_A}{\varepsilon}} \Big(\frac{2L_A}{\varepsilon^2} + \frac{2M_A L_B}{\varepsilon^3} + \frac{L_B}{\varepsilon^2} \Big) W_1(\eta, \nu) \end{aligned}$$

where in the penultimate line we use mean-value theorem, and in the final line the uniform upper bound on ψ_η^ε and Lipschitz bound above. Denote this value by C_w, then normalisation has the same

Lipschitz bound. Hence

$$\begin{aligned}
\left|\mathcal{F}^\varepsilon[\eta] - \mathcal{F}^\varepsilon[\nu]\right| &= \left|\frac{w_\eta^\varepsilon(x)}{\int_0^1 w_\eta^\varepsilon(x)dx} - \frac{w_\nu^\varepsilon(x)}{\int_0^1 w_\nu^\varepsilon(x)dx}\right| \\
&\leq \frac{\left|w_\eta^\varepsilon(x) - w_\nu^\varepsilon(x)\right|}{\left|\int_0^1 w_\eta^\varepsilon(x)dx\right|} + \frac{|w_\nu^\varepsilon(x)|\left|\int_0^1 w_\eta^\varepsilon(x) - w_\nu^\varepsilon(x)dx\right|}{\left|\left(\int_0^1 w_\eta^\varepsilon(x)dx\right)\left(\int_0^1 w_\nu^\varepsilon(x)dx\right)\right|} \\
\left\|\mathcal{F}^\varepsilon[\eta] - \mathcal{F}^\varepsilon[\nu]\right\|_\infty &\leq (M_B + \varepsilon)e^{\frac{2M_A}{\varepsilon}} C_w W_1(\eta,\nu) + \frac{(M_B+\varepsilon)^2}{\varepsilon} e^{\frac{6M_A}{\varepsilon}} C_w W_1(\eta,\nu) \\
&:= L_{\mathcal{F}} W_1(\eta,\nu)
\end{aligned}$$

□

Proposition C.2. *Fix $\varepsilon > 0$, then there exists at least one fixed point of the mapping $\mathcal{F}^\varepsilon[\rho] = \rho$. That is, there is at least one solution to the regularised steady-state equations.*

Proof. First note S^ε is a non-empty, closed, bounded and convex subset of a Banach space. A solution to the steady state equation is uniquely expressed by the fixed point of the mapping $\mathcal{F}^\varepsilon : S^\varepsilon \to S^\varepsilon$. By Lemma C.10, this operator is well-defined and an invariant mapping. By Proposition C.1 it is continuous and by Lemma C.11 it is relatively compact in $C([0,1])$. Applying Schauder's Fixed Point Theorem [43], there exists at least one fixed point of the mapping $\mathcal{F}^\varepsilon : S^\varepsilon \to S^\varepsilon$, and therefore a solution to the regularised steady state equation. □

Remark C.1. *One can consider a PDE with additional noise not captured by the first two moments as an independent stochastic fluctuation in the policy space of size σ, with reflecting boundary conditions. The corresponding stationary distribution would be equivalent to the regularisation above, and therefore existence would follow. Moreover, uniqueness would follow for sufficiently high σ by analysing the Lipschitz constant of the operator, $\mathcal{F}$.*

We can now extend this to find an unregularised solution by sending the parameter $\varepsilon \to 0$.

Proof of Theorem 4.1

Proof. Let $\mu^\varepsilon(dx) = \rho^\varepsilon(x)dx$. Since $[0,1]$ is compact, the space of probability measures $\mathcal{P}([0,1])$ is sequentially compact under weak convergence. Along any sequence $\varepsilon_n \to 0$, there exists a subsequence and measure $\mu \in \mathcal{P}([0,1])$ such that [4]

$$\mu^{\varepsilon_{n_k}} \rightharpoonup \mu.$$

The weak form of the regularised steady-state equation at the stationary distribution μ^ε is

$$\int_0^1 \left[A_{\mu^\varepsilon}(x)\varphi'(x) + \frac{1}{2}(B^2_{\mu^\varepsilon}(x) + \varepsilon)\varphi''(x)\right] d\mu^\varepsilon(x) = 0$$

for all $\varphi \in C^2([0,1])$ such that $\varphi'(0) = \varphi'(1) = 0$. Along the subsequence, we have

$$\int_0^1 \left[A_{\mu^{\varepsilon_{n_k}}}(x)\varphi'(x) + \frac{1}{2}(B^2_{\mu^{\varepsilon_{n_k}}}(x) + \varepsilon_{n_k})\varphi''(x)\right] d\mu^{\varepsilon_{n_k}}(x) = 0.$$

Since the drift, $A_\mu(x)$, and diffusion terms $B^2_\mu(x)$ are continuous and only depend on finitely many bounded moments, weak convergence implies

$$A_{\mu^{\varepsilon_{n_k}}}(x)\varphi'(x) + \frac{1}{2}(B^2_{\mu^{\varepsilon_{n_k}}}(x) + \varepsilon_{n_k})\varphi''(x) \to A_\mu(x)\varphi'(x) + \frac{1}{2}B^2_\mu(x)\varphi''(x)$$

uniformly on $[0,1]$. We can take the limit as $k \to \infty$ and obtain

$$\int_0^1 \left[A_\mu(x)\varphi'(x) + \frac{1}{2}B^2_\mu(x)\varphi''(x)\right] d\mu(x) = 0$$

which is the weak form of the stationary unregularised problem. □

D Empirical Study

D.1 Finite Volume Scheme

We use a finite volume method, following the formulation in [24, Chapter 4]. In conservation form, equation (8) reads

$$\partial_t \rho + \partial_x F = 0, \quad F = A_\rho \rho - \frac{1}{2}\partial_x (B_\rho^2 \rho).$$

The domain $[0, 1]$ is divided into M uniform cells such that

$$C_i = [x_{i-1/2}, x_{i+1/2}], \quad \Delta x = \frac{1}{M}, \quad x_i = \big(i - \frac{1}{2}\big)\Delta x,$$

for $i = 1, ..., M$. At time t_n, we approximate the cell average by

$$\rho_i^n \approx \frac{1}{\Delta x} \int_{C_i} \rho(t_n, x) dx.$$

The first-order upwind scheme gives

$$\rho_i^{n+1} = \rho_i^n - \frac{\Delta t_n}{\Delta x}\big(F_{i+1/2}^n - F_{i-1/2}^n\big).$$

For advection, we approximate the velocity at the interface by

$$a = A_{i+1/2}^n = \frac{A_i^n + A_{i+1}^n}{2}.$$

Let $a^+ = \max\{a, 0\}$ and $a^- = \min\{a, 0\}$. The flux from advection is

$$F_{\text{adv},i+1/2}^n = a^+ \rho_i^n + a^- \rho_{i+1}^n.$$

Likewise for diffusion,

$$F_{\text{diff},i+1/2}^n = -\frac{1}{2}\frac{(B_{i+1}^n)^2 \rho_{i+1}^n - (B_i^n)^2 \rho_i^n}{\Delta x}.$$

The sum $F_{i+1/2}^n = F_{\text{adv},i+1/2}^n + F_{\text{diff},i+1/2}^n$ gives the flux at each cell interface. Moreover, the moments are approximated by

$$\mu_\ell(t_n) \approx \sum_i x_i^\ell \int_{C_i} \rho(t_n, x) dx \approx \Delta x \sum_i x_i^\ell \rho_i^n.$$

To ensure non-negative cell values, the time step is chosen such that

$$\Delta t_n \left(\frac{\max_i |A_i^n|}{\Delta x} + \frac{\max_i (B_i^n)^2}{\Delta x^2} \right) \leq 1.$$

D.2 Additional Figures

All simulations and numerical solutions have been run on a laptop with 12th Gen Intel(R) Core(TM) i7-1260P CPU. The total compute time for the experiments was approximately 40 hours. For the learning rate time scaling, the simulation and numerical solution for $\alpha = 0.001$ was run for ten times longer than for the $\alpha = 0.01$ case.

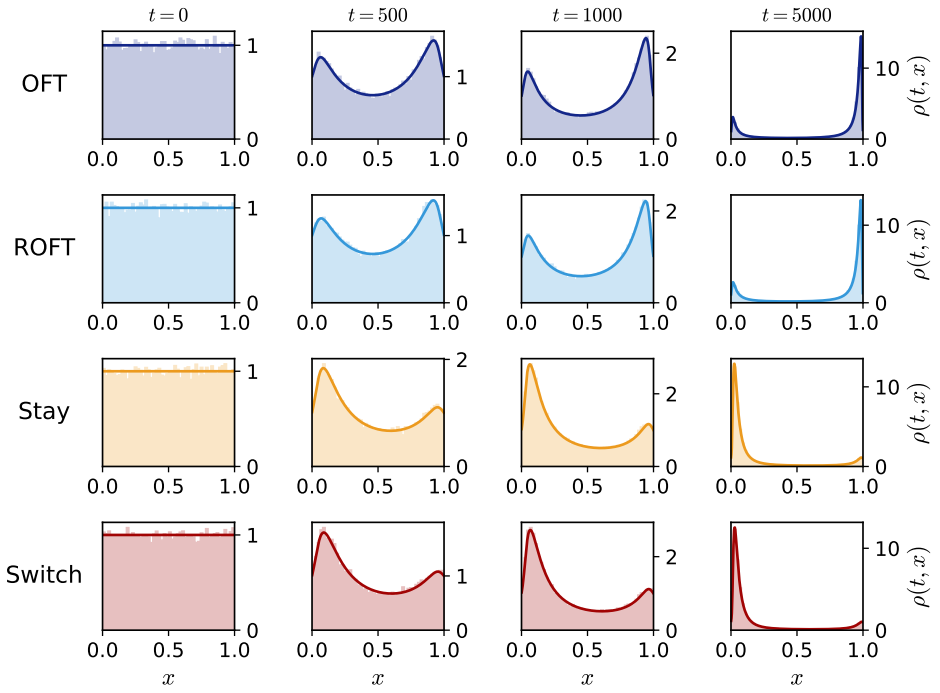

Figure 3: Evolution of the strategy distribution where the population is initialised with $X \sim \text{Unif}(0, 1)$ for the 4 rules. The theoretical solution (solid line) matches the simulations (histogram).

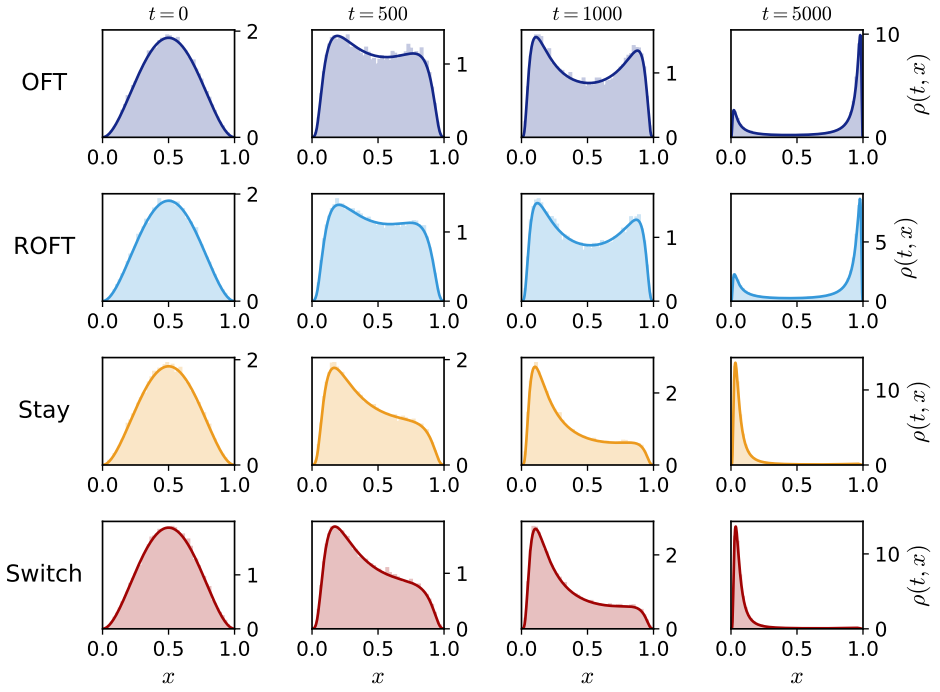

Figure 4: Evolution of the strategy distribution where the population is initialised with $X \sim$ Beta(3,3) for the 4 rules. The theoretical solution (solid line) matches the simulations (histogram).

NeurIPS Paper Checklist

1. **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The claims align with the theoretical and experimental results presented in the paper.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Limitations are discussed where assumptions are made. Moreover, these are highlighted as future directions to work on.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All proofs are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs providescoped in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We outline the exact methodology for the agent-based simulations and provide the explicit formulas to numerically solve the theoretical equations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Agent-based simulation code and finite volume code available in the supplementary materials.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All details of the agent-based simulations are provided in the paper. Multiple initial conditions are presented in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The histogram of agent strategy distribution is presented for agent-based simulations, averaged over 30 simulations. The confidence interval is not shown in the diagram to avoid any visual confusion.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details of the compute required are provided in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Paper follows the guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper looks at promoting cooperation in RL populations. There is no direct path to negative applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All code is original. Relevant sources for theoretical motivation are cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code is provided.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Paper does not involve crowdsourcing nor human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Paper does not involve crowdsourcing nor human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The paper does not make use of LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.