The Dynamics of Policy Gradient in Social Dilemmas with Partner Selection

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

Chin-wing Leung
Department of Computer Science
Warwick University
chin-wing.leung@warwick.ac.uk

Paolo Turrini
Department of Computer Science
Warwick University
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 a multi-agent environment 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. 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 [48, 27], multi-agent reinforcement learning (MARL) [19], and evolutionary game theory [28]; 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 [32, 37] or dynamic interactions [2]. In his seminal work, Nowak [28] identified partner selection as a key mechanism for the emergence of cooperation. This has led to extensive research on cooperation on dynamic networks [2, 46, 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

Preprint.

mechanism, coined Out-for-Tat (OFT), enforces cooperative pairs to stay together; whilst re-wiring pairs with a defector [16, 49]. Recent contributions have found allowing agents to learn partner selection is sufficient for cooperation to emerge [36, 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 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 social dilemmas 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 from an initially non-cooperative state. 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 [29, 38, 30]. This can take many forms, including active linking [29, 31, 2], partner switching [10, 24], 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 [33, 9, 49]. 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 [36].

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 [39, 12], selection-mutation models [44, 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. [50] uses 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. [42] provides the mean dynamics of policy gradient and compares against the replicator dynamics. Bernasconi et al. [3] formally establishes the connection between the replicator dynamics and the soft-max policy gradient, and proves 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 [25].

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 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 [26] and ESS in the one-shot game [41]. 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. We will consider 4 commonly studied rules [19, 36]:

- 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 [47]. If action a_h is selected in round h of an episode and the resulting accumulated rewards from step h onwards are $R^h_{a_h}$, where we assume the discount factor is 1, the approximated gradient of the score-function is $(R^h_{a_h} - \beta)(\boldsymbol{e}_{a_h} - \boldsymbol{\pi})$, where $\beta \in \mathbb{R}$ is a baseline [43] and we set it to 0. The parameter update across an episode of length H is

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

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

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

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 will always cooperate.

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.

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 distribution of the opponent given agent i's strategy is $\rho^h(y|x)$. For any partner selection rule, the distribution at the next step is

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

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

$$\begin{aligned}
\rho^{h+1}(y|x) &= \int_0^1 \Big(\delta_z(y)xz + \rho^{h+1}(y|x,\text{switch})(1-xz)\Big)\rho^h(z|x)dz \\
&= x\int_0^1 \delta_z(y)z\rho^h(z|x)dz + \rho^{h+1}(y|x,\text{switch})\int_0^1 (1-xz)\rho^h(z|x)dz \\
&= xy\rho^h(y|x) + \rho(y)(1-xm^h(x))
\end{aligned}$$

where, upon breaking a connection, a new opponent is sampled from the population. Given the focal agent's strategy x, the mean cooperation frequency of the opponents at step h is denoted by

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

For the 4 partner selection rules, the recursion is given in Table 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.

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

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

Following Equation (2), the episodic reward is the sum of all accumulated rewards conditioned on an action adopted at step $h = k$, followed by the original policy x in future interactions. 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 at k respectively. Likewise, the opponent distribution, expected cooperation, and reward at time h are denoted $\rho^h(y|x, C^k)$, $m_C^{k,h}(x)$, 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

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

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

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

Let $\mu_l = \int y^l \rho(y)dy$ denote the lth moment of the underlying population distribution $\rho(y)$; the reward differences for the first two steps in the episode are shown in Table 2 (a full derivation is provided in Appendix A).

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 across the first two rounds: keeping defective pairs together is as effective as keeping cooperative ones. We will now provide a more general result for when $H > 2$.

The conditional reward difference for Always stay and Always switch is zero for all steps $h \geq k + 1$. Here, we aim to show that the reward difference for all time steps $h \geq k + 1$ is non-negative 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 B. □

Corollary 3.1. *Under the OFT and ROFT partner selection rules, if $\Delta G[\rho] > 0$ for $H = 2$ then $\Delta G[\rho] > 0$ for all $H \geq 2$.*

The above result allows a much simpler reward structure to be studied, associated with the partner selection mechanisms. In particular, increasing the episodic length can only increase the marginal accumulated rewards for cooperation. Therefore, the results in Section 3.3 provide sufficient conditions for the emergence of cooperation for all $H \geq 2$.

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 always stay and always switch converge to the pure defection equilibrium, regardless of the initial distribution. Finally, we show that both OFT and R-OFT necessarily increase cooperation when the initial population has sufficient variance.

Under the random matching and always stay mechanisms, the reward difference simplifies to $-Hc$ for any episode length H. As such, when $H = 2$ the policy update is given by

$$\frac{dx}{dt} = -4\alpha c x^2(1-x)^2$$

which generates the mean-field continuity PDE

$$\partial_t \rho + \partial_x(-4\alpha x^2(1-x)^2\rho) = 0.$$

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.*

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 R-OFT can induce a more cooperative population. Under the OFT mechanism, the continuity equation is

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

where for $H = 2$, we have

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

By Corollary 3.1, for $H > 2$ this value is a lower bound for the sign of the drift. As a consequence, if cooperation emerges for $H = 2$, it provides a sufficient condition for longer episode lengths. 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 exists a unique solution to the continuity equation. Furthermore, the unique solution can be expressed as*

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

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

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

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

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

Theorem 3.2. *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 $\rho(t) \rightharpoonup (X_{K^*})_{\#}\rho_0$. That is, if the initial population variance is high enough, the limiting distribution is shifted towards higher cooperation.*

We highlight that a finite K^* is attained, which means $X_{K^*}(x) < 1$. This enforces that the population does not converge to a single strategy, and hence does not converge to 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 $\frac{1}{4}$, meaning $b > 4c$ 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 Ito'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

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

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

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 [34]. 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.

Proposition 4.1. *Suppose ρ_0 has non-zero interior mass and $\Delta G[\rho_0] > 0$. Then there exists a $T \in \mathbb{R}^+$ and α^* such that for all $\alpha < \alpha^*$, the mean level of cooperation is increasing on $[0, T]$.*

It does not, however, guarantee that a stable cooperative equilibrium will be reached. When $H = 2$, the population variance remaining sufficiently high would prove a sufficient condition, but the evolution of the variance depends upon higher moments. To analyse the equilibrium which is attained in the stochastic case, we prove the limiting distribution is well-posed and turn to numerical solutions to analyse the behaviour.

4.2 Stationary Distribution

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

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

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 $B_\eta^2(x)$ are a function of x and η only. The regularised versions are $A_\eta^\varepsilon(x) = A_\eta(x)$ and $B_\eta^{2,\varepsilon}(x) = B_\eta^2(x) + \varepsilon$. 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 $\eta \in S$, the unique solution $\mathcal{F}^\varepsilon[\eta](x)$ to the linear ODE is given by

$$\mathcal{F}^\varepsilon[\eta](x) = \frac{w_\eta^\varepsilon(x)}{\int_0^1 w_\eta^\varepsilon(x)dx}, \qquad w_\eta^\varepsilon(x) = \frac{1}{B_\eta^{2,\varepsilon}(x)} \exp(\int_0^x \frac{2A_\eta^\varepsilon(y)}{B_\eta^{2,\varepsilon}(y)}dy)$$

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 in Appendix B.

Proposition 4.2. *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.*

Remark 4.1. *One can consider a PDE with additional noise not captured by the first two moments as an independent stochastic fluctuation of size σ. 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 $\varepsilon \to 0$.

Theorem 4.1. *There exists at least one stationary probability measure $\rho \in \mathcal{P}([0,1])$ which solves* (9) *in the weak sense.*

The result implies that the steady state solutions can contain Dirac measures at the boundary. Consequently, we would expect in the long run for mass to accumulate in clusters of pure defection and cooperation. Moreover, the steady state solution is non-unique, explaining how 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 1,000 agents training over E =5e6 episodes. The population distribution at time t is obtained by averaging the simulations. For comparison, the simulation time when plotting is scaled such that $t = E/N$.

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

5.2 Result

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 FPE derived 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, the rules Stay and Switch induce defection to take over the population.

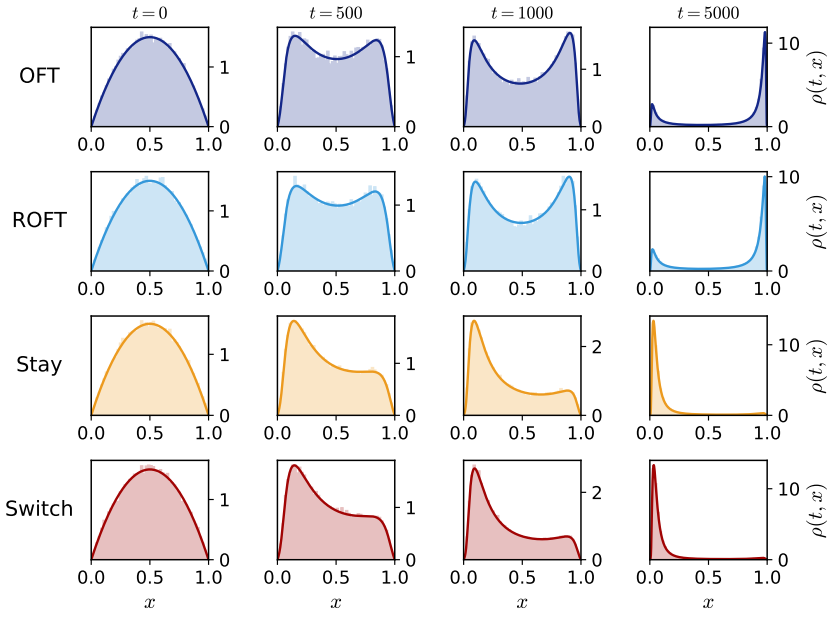

Figure 1: Evolution of the strategy distribution where the population is initialised with $X \sim \beta(2, 2)$ for the 4 rules. The theoretical solution (solid line) matches the simulations (histogram).

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 to a singular policy [19, 36]. 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 higher learning rate can induce greater variance and support cooperation, the FPE approximation loses fidelity.

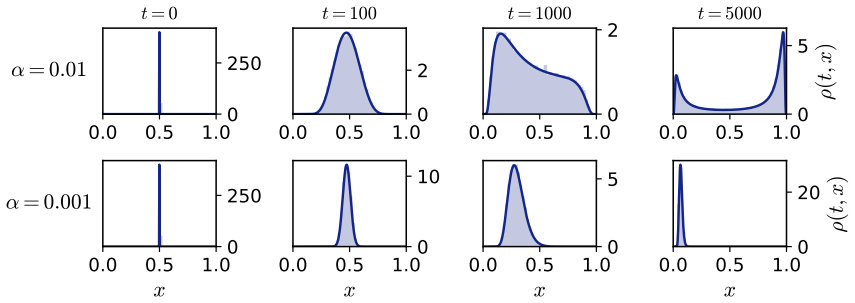

Figure 2: Evolution of the strategy distribution under OFT where the population is initialised with a Dirac at 0.5. The effect of the learning rate is clear: higher values induce additional variance in the population which in turn induces cooperation. Time has been scaled with the learning rate for comparison.

6 Conclusion

In this paper, we study the learning dynamics of 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. This analysis revealed that the structure of the long-term dynamics constitutes mass at the pure defection and pure cooperation strategies. In the experiments, we demonstrate that the stochastic model accurately describes the policy gradient dynamics and can be used to understand the role of the learning rate in inducing cooperation. Future directions include extending the model to other learning algorithms and considering a synchronous population partner selection model.

References

[1] N. Anastassacos, S. Hailes, and M. Musolesi. Partner Selection for the Emergence of Cooperation in Multi-Agent Systems Using Reinforcement Learning, Feb. 2020. URL https://aaai.org/Library/conferences-library.php. Conference Name: Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20) Meeting Name: Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20) Place: New York City, NY, USA.

[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. wing Leung, and H. fung Leung. Modelling the dynamics of multiagent q-learning in repeated symmetric games: a mean field theoretic approach. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:202768371.

[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. doi:10.65109/mlbh9101. URL http://dx.doi.org/10.65109/mlbh9101.

[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. doi:10.65109/mlsq9433. URL http://dx.doi.org/10.65109/mlsq9433.

[22] C.-w. Leung, P. Turrini, and A. Nowé. Curiosity-driven partner selection accelerates convention emergence in language games. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '25, page 1282–1290. International Foundation for Autonomous Agents and Multiagent Systems, May 2025. doi:10.65109/just1524. URL http://dx.doi.org/10.65109/just1524.

[23] C.-w. Leung, P. Turrini, F. P. Santos, and M. Musolesi. Learning to cooperate with minimal observability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29521–29529, 2026. doi:10.1609/aaai.v40i35.40194. URL https://doi.org/10.1609/aaai.v40i35.40194.

[24] 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.

[25] 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].

[26] 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.

[27] 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.

[28] 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/.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] B. Russell, C.-w. Leung, and P. Turrini. Defection at first sight : learning partner selection in optional social dilemmas without prior information. In *25th International Conference on Autonomous Agents and Multiagent Systems*. IFAAMAS; ACM Digital library, May 2026. doi:10.65109/IBSZ1473. URL https://doi.org/10.65109/IBSZ1473. In Press.

[37] 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.

[38] 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.

[39] 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.

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

[41] J. M. Smith. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, 1982. ISBN 978-0-521-28884-2. doi:10.1017/CBO9780511806292. URL https://www.cambridge.org/core/books/evolution-and-the-theory-of-games/A3BDF54AF5C6297E308AB15BBEF45E48.

[42] S. Srinivasan, M. Lanctot, V. Zambaldi, J. Pérolat, K. Tuyls, R. Munos, and M. Bowling. Actor-critic policy optimization in partially observable multiagent environments. *Advances in neural information processing systems*, 31, 2018.

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

[44] 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.

[45] C. Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[46] 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.

[47] 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.

[48] 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.

[49] 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/.

[50] 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 Derivations

A.1 Conditional Partner Distribution 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 3: Conditional partner distribution and reward difference for OFT

h	$\rho^h(y\|x,C^0)$	$\rho^h(y\|x,D^0)$	$m_C^{0,h}(x)$	$m_D^{0,h}(x)$	$\Delta r^{0,h}(x)$
0	$\rho(y)$	$\rho(y)$	μ_1	μ_1	$-c$
1	$y\rho(y)+(1-\mu_1)\rho(y)$	$\rho(y)$	$\mu_2-\mu_1^2+\mu_1$	μ_1	$\mathsf{b}(\mu_2-\mu_1^2)$

Table 4: Conditional partner distribution and reward difference for ROFT

h	$\rho^h(y\|x,C^0)$	$\rho^h(y\|x,D^0)$	$m_C^{0,h}(x)$	$m_D^{0,h}(x)$	$\Delta r^{0,h}(x)$
0	$\rho(y)$	$\rho(y)$	μ_1	μ_1	$-c$
1	$\rho(y)$	$(1+\mu_1-y)\rho(y)$	μ_1	$\mu_1+\mu_1^2-\mu_2$	$\mathsf{b}(\mu_2-\mu_1^2)$

Table 5: Conditional partner distribution and reward difference for Stay and Switch

h	$\rho^h(y\|x,C^0)$	$\rho^h(y\|x,D^0)$	$m_C^{0,h}(x)$	$m_D^{0,h}(x)$	$\Delta r^{0,h}(x)$
0	$\rho(y)$	$\rho(y)$	μ_1	μ_1	$-c$
≥ 1	$\rho(y)$	$\rho(y)$	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.

A.2 Stochastic Dynamics in 2-action Symmetric Games

Recall the update to the parameter is

$$\Delta\psi_C = \alpha\sum_{h=0}^{H-1} U^h,$$

where $U^h=(R^h_{a_h}-\beta)(1\{a_h=C\}-x)$. Define the mean vector

$$\mu(\boldsymbol{\psi},t) := \mathbb{E}[\Delta\boldsymbol{\psi}(t)],$$

and the covariance matrix by

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

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

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

where $\mu_C=\mathbb{E}[\Delta\psi_C]=\alpha x(1-x)\Delta G[\rho]$, and $\Delta G[\rho]=G_C-G_D$. To compute the covariance, note that

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

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

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

We then approximate the discrete stochastic updates with the continuous time diffusion process as defined in Equation (6). To simplify the equations, consider the logit difference given by $z = \psi_C - \psi_D$. Then $\Delta z = \Delta\psi_C - \Delta\psi_D = 2\Delta\psi_C$. The corresponding one-dimensional parameter dynamics are

$$dz = \mu_z dt + \sigma_z dW_t \tag{10}$$

where

$$\mu_z = 2\alpha x(1-x)\Delta G[\rho], \quad \sigma_z^2 = 4\Sigma_{CC}.$$

The cooperation probability of an agent is then computed with the softmax function

$$x(z) = \frac{1}{1+e^{-z}}.$$

Applying Ito's lemma with

$$x'(z) = x(1-x), \qquad x''(z) = x(1-x)(1-2x)$$

we obtain

$$dx = [x(1-x)\mu_z + \frac{1}{2}x(1-x)(1-2x)\sigma_z^2]\, dt + x(1-x)\sigma_z\, dW_t.$$

Substituting in the formulas of μ_z and σ_z gives the provided equation.

A.3 Conditional Moments

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 type at timestep l be given by Y_l. For the focal agent, the action at the conditioned round is fixed:

$$z_h = 1\{a_h = C\},$$

and for all round $l > h$, the policy is given by x. Then the actions sampled in each round $l > h$ are

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

for the opponent and focal agent, respectively. Note that the focal agent's actions are i.i.d. across the episode. The reward at round l is given by

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

At the initial 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

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

Likewise, the conditional variance when $l = h$ is

$$\begin{aligned}\text{Var}(r_h | a_h = a) &= b^2 \text{Var}(\zeta_h | a_h = a) \\ &= b^2 m^h(x)(1 - m^h(x)),\end{aligned}$$

and for $l > h$ it is

$$\begin{aligned}\text{Var}(r_l | a_h = a) &= b^2 \text{Var}(\zeta_l | a_h = a) + c^2 \text{Var}(1 - z_l) \\ &= b^2 m_a^{h,l}(x)(1 - m_a^{h,l}(x)) + c^2 x(1 - x). \end{aligned} \tag{11}$$

The conditional means can be computed using repeated substitution of the iterative partner distribution update. For the covariance terms, let $h \leq l < k \leq H - 1$. Then

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

where the final term in the second line is zero due to the independence of action selection by the focal agent. Note that for $H = 2$, the final two terms are zero as $l = h$ is the only relevant term and the focal action is fixed. In this case, this covariance is given by

$$\text{Cov}(Y_l, Y_k | a_h = a) = \mathbb{E}[Y_l Y_k | a_h = a] - m_a^{h,l}(x) m_a^{h,k}(x).$$

Computing the Covariance for $H = 2$

Beginning with the variance terms, we have for $h = 0$:

$$\begin{aligned}\mathbb{E}[r_0 | Y_0, a_0 = C] &= b\mathbb{E}[\zeta_0 | Y_0] = bY_0 \\ \text{Var}(r_0 | Y_0, a_0 = C) &= b^2 \text{Var}(\zeta_0 | Y_0) = b^2 Y_0(1 - Y_0) \\ \Rightarrow \text{Var}(r_0 | a_0 = C) &= \mathbb{E}[b^2 Y_0(1 - Y_0)] + \text{Var}(bY_0) \\ &= b^2(\mu_1 - \mu_1^2).\end{aligned}$$

Likewise for defection:

$$\begin{aligned}\mathbb{E}[r_0 | Y_0, a_0 = D] &= b\mathbb{E}[\zeta_0 | Y_0] + c = bY_0 + c \\ \text{Var}(r_0 | Y_0, a_0 = D) &= b^2 \text{Var}(\zeta_0 | Y_0) = b^2 Y_0(1 - Y_0) \\ \Rightarrow \text{Var}(r_0 | a_0 = D) &= \mathbb{E}[b^2 Y_0(1 - Y_0)] + \text{Var}(bY_0 + c) \\ &= b^2(\mu_1 - \mu_1^2).\end{aligned}$$

For the following step: $h = 1$,

$$\begin{aligned}\mathbb{E}[r_1 | Y_1, a_0 = C] &= b\mathbb{E}[\zeta_1 | Y_1] + c\mathbb{E}[1 - z_1] = bY_1 + c(1 - x) \\ \text{Var}(r_1 | Y_1, a_0 = C) &= b^2 \text{Var}(\zeta_1 | Y_1) + c^2 \text{Var}(1 - z_1) = b^2 Y_1(1 - Y_1) + c^2 x(1 - x) \\ \Rightarrow \text{Var}(r_1 | a_0 = C) &= \mathbb{E}[b^2 Y_1(1 - Y_1) + c^2 x(1 - x) | a_0 = C] + \text{Var}(bY_1 + c(1 - x) | a_0 = C) \\ &= b^2 \Big(\mathbb{E}[Y_1 | a_0 = C] - \mathbb{E}[Y_1 | a_0 = C]^2 \Big) + c^2 x(1 - x).\end{aligned}$$

Note this is the same form as Equation (11). Using the one-step Markov operator on the opponent distribution, we can derive the reward variances for each of the partner selection rules.

The updated covariances can be computed using the conditional moment

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

OFT The expected cooperation rate of the opponent in the next step given the current opponent and the focal agent's action is

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

Then, computing the conditional mean we have

$$\begin{aligned}
\mathbb{E}[Y_1|a_0 = C] &= \mathbb{E}[\mathbb{E}[Y_1|Y_0, a_0 = C]] \\
&= \mathbb{E}[Y_0^2 + (1 - Y_0)\mu_1] \\
&= \mu_2 + \mu_1 - \mu_1^2 \\
\mathbb{E}[Y_1|a_0 = D] &= \mu_1
\end{aligned}$$

For the covariance terms, we have

$$\begin{aligned}
\mathbb{E}[Y_0Y_1|a_0 = C] &= \mathbb{E}[Y_0\mathbb{E}[Y_1|Y_0, a_0 = C]] \\
&= \mathbb{E}[Y_0(Y_0^2 + (1 - Y_0)\mu_1)] \\
&= \mu_3 + \mu_1^2 - \mu_2\mu_1 \\
\mathbb{E}[Y_0Y_1|a_0 = D] &= \mathbb{E}[Y_0\mathbb{E}[Y_1|Y_0, a_0 = D]] \\
&= \mathbb{E}[Y_0\mu_1] \\
&= \mu_1^2
\end{aligned}$$

The variance terms are

$$\text{Var}(r_1|a_0 = C) = b^2(\mu_2 + \mu_1 - \mu_1^2)(1 - \mu_2 - \mu_1 + \mu_1^2) + c^2x(1 - x), \quad \text{Var}(r_1|a_0 = D) = b^2\mu_1(1 - \mu_1) + c^2x(1 - x).$$

The covariance terms are

$$\begin{aligned}
\text{Cov}(r_0, r_1|a_0 = C) &= b^2\text{Cov}(Y_0, Y_1|a_0 = C) \\
&= b^2(\mathbb{E}[Y_0Y_1|a_0 = C] - \mathbb{E}[Y_0|a_0 = C]\mathbb{E}[Y_1|a_0 = C]) \\
&= b^2(\mu_3 + \mu_1^2 - \mu_2\mu_1 - \mu_1(\mu_2 + \mu_1 - \mu_1^2)) \\
&= b^2(\mu_3 - 2\mu_2\mu_1 + \mu_1^3) \\
\text{Cov}(r_0, r_1|a_0 = D) &= b^2\text{Cov}(Y_0, Y_1|a_0 = D) \\
&= b^2(\mathbb{E}[Y_0Y_1|a_0 = D] - \mathbb{E}[Y_0|a_0 = D]\mathbb{E}[Y_1|a_0 = D]) \\
&= b^2(\mu_1^2 - \mu_1^2) = 0
\end{aligned}$$

It remains to condition on the second step, $h = 1$. Explicitly, the terms are given by

$$\begin{aligned}
\mathbb{E}[r_1|a_1 = C] &= bm^1(x) \\
&= b\int_0^1 y[xy\rho(y) + \rho(y)(1 - x\mu_1)]dy \\
&= b(\mu_1 + x(\mu_2 - \mu_1^2)) \\
\mathbb{E}[r_1|a_1 = D] &= bm^1(x) + c \\
&= b(\mu_1 + x(\mu_2 - \mu_1^2)) + c \\
\text{Var}(r_1|a_1 = a) &= b^2m^1(x)(1 - m^1(x)) \\
&= b^2(\mu_1 + x(\mu_2 - \mu_1^2))(1 - (\mu_1 + x(\mu_2 - \mu_1^2)))
\end{aligned}$$

Collating these terms, we can summarise the elements of the second moment in Table 6.

Table 6: 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	$b^2[(\mu_2 + \mu_1 - \mu_1^2)(1 - \mu_2 - \mu_1 + \mu_1^2) + (\mu_1 - \mu_1^2) + 2(\mu_3 - 2\mu_2\mu_1 + \mu_1^3)] + c^2x(1 - x)$	$[b(\mu_2 + 2\mu_1 - \mu_1^2) + c(1 - x) - \beta]^2$
	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 rate of the opponent in the next step given the current opponent and the focal agent's action is

$$\begin{aligned}\mathbb{E}[Y_1|Y_0, a_0 = C] &= \mu_1\\ \mathbb{E}[Y_1|Y_0, a_0 = D] &= Y_0 \cdot (1 - Y_0) + Y_0\mu_1\end{aligned}$$

Then computing the conditional mean,

$$\begin{aligned}\mathbb{E}[Y_1|a_0 = C] &= \mu_1\\ \mathbb{E}[Y_1|a_0 = D] &= \mathbb{E}[\mathbb{E}[Y_1|Y_0, a_0 = D]]\\ &= \mathbb{E}[Y_0(1 - Y_0) + Y_0\mu_1]\\ &= \mu_1 + \mu_1^2 - \mu_2\end{aligned}$$

For the covariance terms, we have

$$\begin{aligned}\mathbb{E}[Y_0Y_1|a_0 = C] &= \mathbb{E}[Y_0\mathbb{E}[Y_1|Y_0, a_0 = C]]\\ &= \mathbb{E}[Y_0\mu_1]\\ &= \mu_1^2\\ \mathbb{E}[Y_0Y_1|a_0 = D] &= \mathbb{E}[Y_0\mathbb{E}[Y_1|Y_0, a_0 = D]]\\ &= \mathbb{E}[Y_0(Y_0(1 - Y_0) + Y_0\mu_1)]\\ &= \mu_2 - \mu_3 + \mu_1\mu_2\end{aligned}$$

The variance terms are then:

$$\text{Var}(r_1|a_0 = C) = b^2\mu_1(1 - \mu_1) + c^2x(1 - x), \quad \text{Var}(r_1|a_0 = D) = b^2(\mu_1 + \mu_1^2 - \mu_2)(1 - \mu_1 - \mu_1^2 + \mu_2) + c^2x(1 - x)$$

The covariance terms are

$$\begin{aligned}\text{Cov}(r_0, r_1|a_0 = C) &= b^2\text{Cov}(Y_0, Y_1|a_0 = C)\\ &= b^2(\mathbb{E}[Y_0Y_1|a_0 = C] - \mathbb{E}[Y_0|a_0 = C]\mathbb{E}[Y_1|a_0 = C])\\ &= b^2(\mu_1^2 - \mu_1^2) = 0\\ \text{Cov}(r_0, r_1|a_0 = D) &= b^2\text{Cov}(Y_0, Y_1|a_0 = D)\\ &= b^2(\mathbb{E}[Y_0Y_1|a_0 = D] - \mathbb{E}[Y_0|a_0 = D]\mathbb{E}[Y_1|D])\\ &= b^2(\mu_2 - \mu_3 + \mu_1\mu_2 - \mu_1(\mu_1 + \mu_1^2 - \mu_2))\\ &= b^2(\mu_2 - \mu_1^2 + 2\mu_1\mu_2 - \mu_1^3 - \mu_3)\end{aligned}$$

It remains to condition on the second step, $h = 1$. Explicitly, the terms are given by

$$\begin{aligned}\mathbb{E}[r_1|a_1 = C] &= bm^1(x)\\ &= b\int_0^1 y[(1 - x)(1 - y)\rho(y) + \rho(y)(1 - (1 - x)(1 - \mu_1))]dy\\ &= b(\mu_1 - (1 - x)(\mu_2 - \mu_1^2))\\ \mathbb{E}[r_1|a_1 = D] &= bm^1(x) + c\\ &= b(\mu_1 - (1 - x)(\mu_2 - \mu_1^2)) + c\\ \text{Var}(r_1|a_1 = a) &= b^2m^1(x)(1 - m^1(x))\\ &= b^2(\mu_1 - (1 - x)(\mu_2 - \mu_1^2))(1 - \mu_1 + (1 - x)(\mu_2 - \mu_1^2)))\end{aligned}$$

Collating these terms, we can summarise the elements of the second moment in Table 7.

Table 7: 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	$b^2[(\mu_1 + \mu_1^2 - \mu_2)(1 - \mu_1 - \mu_1^2 + \mu_2) + \mu_1(1 - \mu_1)$ $+ 2(\mu_2 - \mu_1^2 + 2\mu_1\mu_2 - \mu_1^3 - \mu_3)] + c^2x(1 - x)$	$[b(2\mu_1 + \mu_1^2 - \mu_2) + c(2 - x) - \beta]^2$
1	C	$b^2(\mu_1 - (1 - x)(\mu_2 - \mu_1^2))(1 - \mu_1 + (1 - x)(\mu_2 - \mu_1^2))$	$[b(\mu_1 - (1 - x)(\mu_2 - \mu_1^2)) - \beta]^2$
	D	$b^2(\mu_1 - (1 - x)(\mu_2 - \mu_1^2))(1 - \mu_1 + (1 - x)(\mu_2 - \mu_1^2))$	$[b(\mu_1 - (1 - x)(\mu_2 - \mu_1^2)) + c - \beta]^2$

Always Stay Since the transition is independent of the action, for both actions $a \in \{C, D\}$ the expected cooperation rate of the opponent in the next step given the current opponent and action is

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

Then computing the conditional mean,

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

For the covariance terms, we have

$$\begin{aligned}\mathbb{E}[Y_0Y_1|a_0 = a] &= \mathbb{E}[Y_0\mathbb{E}[Y_1|Y_0, a_0 = a]] \\ &= \mathbb{E}[Y_0Y_0] \\ &= \mu_2\end{aligned}$$

The variance terms are then:

$$\text{Var}(r_1|a_0 = a) = b^2\mu_1(1 - \mu_1) + c^2x(1 - x)$$

The covariance terms are

$$\begin{aligned}\text{Cov}(r_0, r_1|a_0 = a) &= b^2\text{Cov}(Y_0, Y_1|a_0 = a) \\ &= b^2(\mathbb{E}[Y_0Y_1|a_0 = a] - \mathbb{E}[Y_0|a]\mathbb{E}[Y_1|a_0 = a]) \\ &= b^2(\mu_2 - \mu_1^2) \\ &= b^2\text{Var}(\rho)\end{aligned}$$

It remains to condition on the second step, $h = 1$. Explicitly, the terms are given by

$$\begin{aligned}\mathbb{E}[r_1|a_1 = C] &= bm^1(x) \\ &= b\int_0^1 y\rho(y)dy \\ &= b\mu_1 \\ \mathbb{E}[r_1|a_1 = D] &= bm^1(x) + c \\ &= b\mu_1 + c \\ \text{Var}(r_1|a_1 = a) &= b^2m^1(x)(1 - m^1(x)) \\ &= b^2\mu_1(1 - \mu_1)\end{aligned}$$

Collating these terms, we can summarise the elements of the second moment in Table 8.

Table 8: Second moments (S_a^h) of episodic reward when $H = 2$ under Always 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	$b^2\mu_1(1 - \mu_1)$	$[b\mu_1 + c - \beta]^2$

Always Switch Since the transition is independent of the action, for both actions $a \in \{C, D\}$ the expected cooperation rate of the opponent in the next step given the current opponent and action is

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

Then computing the conditional mean,

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

For the covariance terms, we have

$$\begin{aligned}\mathbb{E}[Y_0Y_1|a_0 = a] &= \mathbb{E}[Y_0\mathbb{E}[Y_1|Y_0, a_0 = a]] \\ &= \mathbb{E}[Y_0\mu_1] \\ &= \mu_1^2\end{aligned}$$

The variance terms are then:

$$\text{Var}(r_1|a_0 = a) = b^2\mu_1(1-\mu_1) + c^2x(1-x)$$

The covariance terms are

$$\begin{aligned}\text{Cov}(r_0, r_1|a_0 = a) &= b^2\text{Cov}(Y_0, Y_1|a_0 = a)\\ &= b^2(\mathbb{E}[Y_0Y_1|a_0 = a] - \mathbb{E}[Y_0|a_0 = a]\mathbb{E}[Y_1|a_0 = a])\\ &= b^2(\mu_1^2 - \mu_1^2) = 0\end{aligned}$$

Note that for conditioning at step $h = 1$, the same derivation as Always Stay holds. Collating these terms, we can summarise the elements of the second moment in Table 9.

Table 9: Second moments (S_a^h) of episodic reward when $H = 2$ under Always 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	$b^2\mu_1(1-\mu_1)$	$[b\mu_1 + c - \beta]^2$

B Missing Proofs

B.1 Conditional Rewards

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

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

$$\Delta m^{0,h} = \int_0^1 y\Delta\rho^{0,h}(y)dy$$

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 $h > 1$

$$g^h(y) = yg^{h-1}(y) - \mathbb{E}[Yg^{h-1}(Y)], \quad g^1(y) = y - \mu_1. \tag{12}$$

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

Lemma B.1. *Let $\rho(y)$ be the initial distribution, and g^h be defined by Equation* (12)*. Then for all $h \geq 1$, the difference in distribution and mean of the opponent is given by*

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

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

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

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

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

□

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

$$\langle f, g \rangle := \mathbb{E}[Yf(Y)g(Y)]. \tag{13}$$

Lemma B.2. *The operator $\mathcal{T}$ is self-adjoint with respect to the inner product in Equation* (13)*.*

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 B.3. *Let g^h be defined by Equation* (12)*. For any $m, n \in \mathbb{N}$,*

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

Proof. By Lemma B.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 B.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 B.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 B.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 B.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,

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

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

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

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 B.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

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

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

Lemma B.6. *For all h, the distribution $q^h(y|x)$ is given by*

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

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

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

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

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

□

Proof of Proposition 3.1

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

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

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

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

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 B.4. Note that the proof has an identical structure for the ROFT rule, invoking Lemma B.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 $\Delta m^{0,1} = \text{Var}(\rho)$, and for all $h > k+1$ the reward difference $\Delta r^{k,h}(x) \geq 0$, then

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

which is increasing in H when $\Delta G[\rho] > 0$ for $H = 2$. □

B.2 Mean Dynamics

Proof of Theorem 3.1

Proof. The characteristic flow is generated by

$$\frac{d}{dt}X_t(x) = -4\alpha c X_t^2(x)(1 - X_t(x))^2, \quad X_0(x) = x_0,$$

This is a separable ODE, which integrating gives

$$\int \frac{1}{X^2(1-X)^2} dX = \int -4\alpha c \; dt$$
$$\frac{1}{1-X_t} - \frac{1}{X_t} + 2\ln|\frac{X_t}{1-X_t}| = -4\alpha c t + C$$

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

$$F(X_t) = F(x_0) - 4\alpha c t.$$

For $c > 0$ and $\alpha > 0$, then as $t \to \infty$ we have $F(X_t) \to -\infty$ for every $x_0 \in (0,1)$. Since $F(x) \to -\infty$ as $x \to 0$, we have that $X_t \to 0$. Since the velocity field $u(x) = -4\alpha c x^2(1-x)^2 \in C^1([0,1])$, the solution to the continuity equation is given by the pushforward of the initial point under the characteristic flow:

$$\rho(t, \cdot) = (X_t)_{\#}\rho_0.$$

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

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

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

$$|\varphi(X(t,x))\rho_0(x)| \leq \|\varphi\|_\infty |\rho_0(x)|$$

By the dominated convergence theorem,

$$\int_0^1 \varphi(X(t,x))\rho_0(x)dx \to \varphi(0)\int_0^1 \rho_0(x)dx = \varphi(0)$$

which means

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

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

Lemma B.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

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

which is uniformly bounded by $\Delta G[\rho]$ on $[0,1]$. Therefore by the mean value theorem,

$$\begin{aligned}|u[\rho](x) - u[\rho](y)| &\leq \sup_z|\partial_x u[\rho](z)|\ |x-y| \\ &\leq 2\alpha|\Delta G[\rho]||x-y|\end{aligned}$$

Since the episode length and per-game reward are bounded, $|\Delta G[\rho]|$ is uniformly bounded. Therefore exists an L_1 such that $u[\rho](y)$ is spatially Lipschitz with constant L_1. The condition for sub-linearity follows from

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

which is uniformly bounded. □

Lemma B.8. *For every $\rho_1, \rho_2 \in \mathcal{P}$, there exists a positive constant $L_2 \in \mathbb{R}$ such that*

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

Proof.

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

Therefore,

$$\|u[\rho_1](\cdot) - u[\rho_2](\cdot)\|_{C^0([0,1])} \leq \frac{\alpha}{8}|\Delta G[\rho_1] - \Delta G[\rho_2]|$$

It suffices to show that

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

for some constant $C \in \mathbb{R}$. Using the Kantorovich-Rubinstein duality [45], 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 i, the ith moment is

$$m_i = \int_0^1 x^i d\rho(x)$$

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

$$\frac{1}{i}|m_i(\rho_1) - m_i(\rho_2)| \leq W_1(\rho_1, \rho_2).$$

Therefore,

$$\begin{aligned}|\Delta G[\rho_1] - \Delta G[\rho_2]| &= b|\text{Var}(\rho_1) - \text{Var}(\rho_2)| \\ &\leq b|m_2(\rho_1) - m_2(\rho_2)| + b|m_1(\rho_1)^2 - m_1(\rho_2)^2| \\ &\leq 2bW_1(\rho_1, \rho_2) + b|m_1(\rho_1) + m_1(\rho_2)||m_1(\rho_1) - m_1(\rho_2)| \\ &\leq 2bW_1(\rho_1, \rho_2) + b \cdot 2 \cdot W_1(\rho_1, \rho_2) \\ &= 4bW_1(\rho_1, \rho_2)\end{aligned}$$

Therefore the result holds with $L_2 = \frac{b}{4}$. □

B.2.1 Proof of Proposition 3.2

Proof. By Theorem 2 in [6], it suffices to show the conditions of Lemma B.7 and Lemma B.8 are satisfied. These require that $\Delta G[\rho]$ is bounded, which is always satisfied for $\rho_0 \in \mathcal{P}([0,1)]$. □

The characteristic ODE admits a solution X_t

$$\begin{aligned}\frac{d}{dt}X_t &= 2\alpha\Delta G[\rho]X_t^2(1-X_t)^2\\ \int \frac{1}{X^2(1-X)^2}dX &= 2\alpha\int \Delta G[\rho]dt\\ F(X_t) &= F(x_0) + 2\alpha K(t)\end{aligned}$$

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

$$X_K(x) = F^{-1}(F(x_0) + 2\alpha K).$$

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

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

The initial condition means

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

Note that for every $x > 0$

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

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

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

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

$$\begin{aligned}|m_1(K_1) - m_1(K_2)| &= |\mathbb{E}[X_{K_1}(X)] - \mathbb{E}[X_{K_2}(X)]|\\ &\le \mathbb{E}[|X_{K_1}(X) - X_{K_2}(X)]|\\ &\le \frac{\alpha}{8}|K_1 - K_2|.\end{aligned}$$

It follows that the second moment is Lipschitz:

$$\begin{aligned}|m_2(K_1) - m_2(K_2)| &= |\mathbb{E}[X_{K_1}^2(X)] - \mathbb{E}[X_{K_2}^2(X)]|\\ &= |\mathbb{E}[X_{K_1}^2(X) - X_{K_2}^2(X)]|\\ &\le \mathbb{E}[|X_{K_1}^2(X) - X_{K_2}^2(X)|]\\ &\le \mathbb{E}[2|X_{K_1}(X) - X_{K_2}(X)|]\\ &\le \frac{\alpha}{4}|K_1 - K_2|\end{aligned}$$

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 $K \to \infty$, $\mathbb{E}[X_K^2(X)] \to 1$, so $\text{Var}(K) = \mathbb{E}[X_K^2(X)] - \mathbb{E}[X_K(X)]^2 \to 0$. 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 all $K < K^*,\ h(K) > 0$ 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

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

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

$$\frac{d}{dt}m_1(t) = \int_0^1 A_\rho(t,x)\rho(t,x)dx = 2\alpha\int_0^1 x^2(1-x)^2\Delta G[\rho]\rho(t,x)dx + O(\alpha^2).$$

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

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

and the covariance is bounded by

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

Therefore, we have

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

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

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

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 $[0,T]$,

$$\begin{aligned}
\frac{d}{dt}m_1(t) &= 2\alpha I(t) + O(\alpha^2) \\
&\ge \alpha I(0) - \frac{1}{4}\alpha^2H^2(H(b+c)+|\beta|)^2 \\
&> 0
\end{aligned}$$

for all $\alpha < \alpha^*$. □

To show the regularity conditions required for a well-posed solution to the steady-state equation, we provide the following Lipschitz continuity result on moments

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

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

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

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

and

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

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

Proof. First note that $\Delta G[\rho]$ is a polynomial in m_1 and m_2, and Σ is a polynomial in x, m_1, m_2, m_3. Since the product of polynomials is a polynomial, we have $A^\varepsilon_\eta = f(x, m_1, m_2, m_3)$ is a polynomial such that $x, m_1, m_2, m_3 \in [0,1]$. Therefore, f is Lipschitz in each of its arguments such that

$$L_f = \sup_{[0,1]^4} \|\nabla f(x, m_1, m_2, m_3)\|_\infty < \infty.$$

Hence,

$$\begin{aligned}|f(x, m_1(\eta), m_2(\eta), m_3(\eta)) - f(x, m_1(\nu), m_2(\nu), m_3(\nu))| &\le L_f(|m_1(\eta) - m_1(\nu)| + |m_2(\eta) - m_2(\nu)| \\ &\quad + |m_3(\eta) - m_3(\nu)|) \\ &\le L_f(W_1(\eta,\nu) + 2W_1(\eta,\nu) + 3W_1(\eta,\nu)) \\ &= 6L_f W_1(\eta,\nu)\end{aligned}$$

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

$$\sup_{[0,1]^4} |\partial_x f| < \infty$$

and therefore for any η,

$$|A_\eta(x) - A_\eta(y)| \le \sup_{[0,1]^4} |\partial_x f||x-y|$$

so A_η is Lipschitz in x. Taking L_A as the maximum of each Lipschitz bound gives the result. The proof for $B^{2,\varepsilon}_\eta$ is identical, since this is also a polynomial. □

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 $A^\varepsilon_\eta(x) \le M_A$ and $B^{2,\varepsilon}_\eta(x) \le M_B$. The next result shows the iterative operator is well-defined.

Lemma B.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 $B_\eta^{2,\varepsilon}(x) \leq M_B + \varepsilon$, we also have the bound

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

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

Lemma B.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 [35], it suffices to show that the set $\mathcal{F}^\varepsilon(S^\varepsilon)$ is equibounded and equicontinuous. Firstly, by Lemma B.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,

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

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 B.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 relatively compact in $C([0,1])$. □

Proposition B.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

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

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

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

□

Proof of Proposition 4.2

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 B.10, this operator is well-defined and an invariant mapping. By Proposition B.1 it is continuous and by Lemma B.11 its relatively compact in $C([0,1])$. Applying Schauder's Fixed Point Theorem [40], 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. □

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. □

C Empirical Study

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.

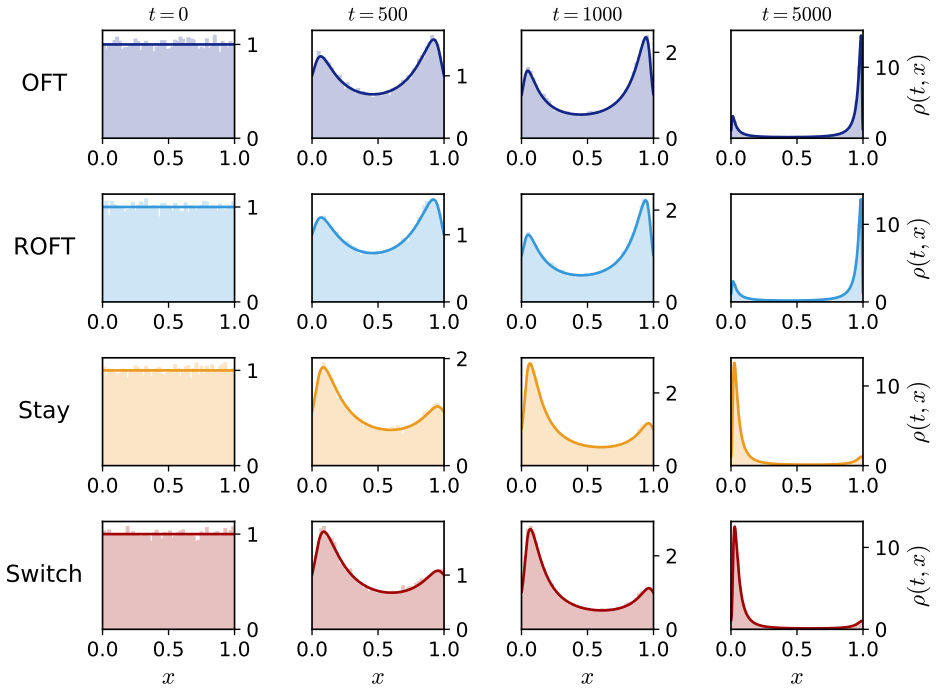

Figure 3: Evolution of the strategy distribution where the population is initialised with $X \sim \text{Uni}(0, 1)$ for the 4 rules. The theoretical solution (solid line) matches the simulations (histogram).

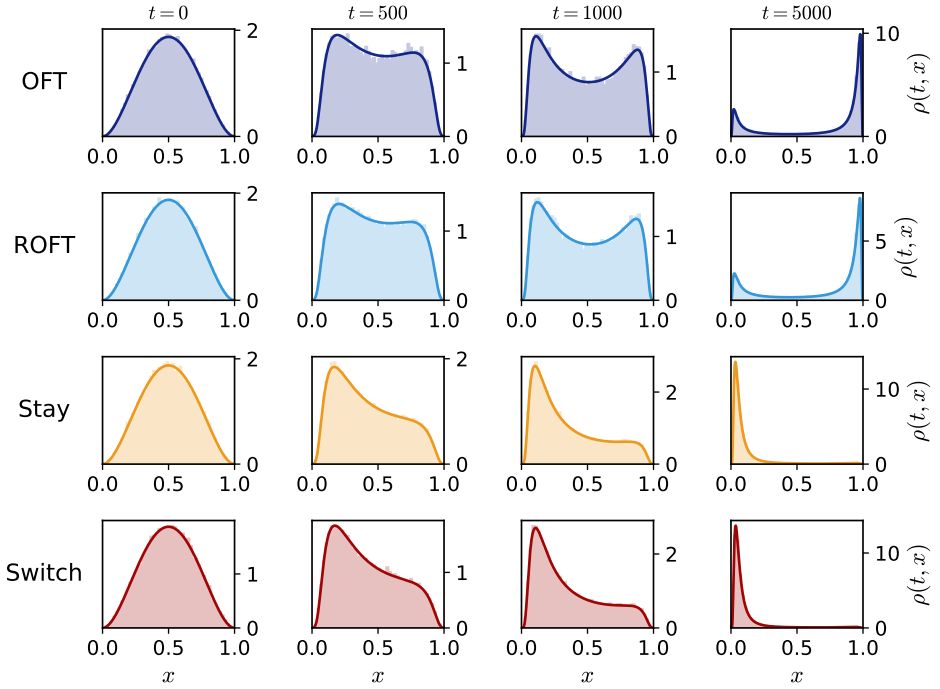

Figure 4: Evolution of the strategy distribution where the population is initialised with $X \sim$ Beta(3,3) for the 4 rules. The theoretical solution (solid line) matches the simulations (histogram).