Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

Highlights

Selfish optimization and collective learning in populations

Physica D xxx (xxxx) xxx

Alex McAvoy*, Yoichiro Mori, Joshua B. Plotkin

- Selfish learning among pairs of agents can lead to undesirable outcomes.
- Individuals often exist in populations and encounter many others while learning.
- Random encounters can improve the outcome of learning processes.
- Sociable, selfish learners can overcome conflicts of interest in populations.

Please cite this article as: A. McAvoy, Y. Mori and J.B. Plotkin, Selfish optimization and collective learning in populations, Physica D (2022) 133426, https://doi.org/10.1016/j.physd.2022.133426.

PHYSD: 133426

ARTICLE IN PRESS

Physica D xxx (xxxx) xxx

Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

Selfish optimization and collective learning in populations

Alex McAvoy [a,b,*,1], Yoichiro Mori [a,b,c,1], Joshua B. Plotkin [a,b,c,1]

[a] Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104, USA
[b] Center for Mathematical Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
[c] Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 29 November 2021
Received in revised form 17 June 2022
Accepted 18 June 2022
Available online xxxx
Communicated by Jordi Garcia-Ojalvo

Keywords:
Collective learning
Multi-agent learning
Population
Repeated game

ABSTRACT

A selfish learner seeks to maximize their own success, disregarding others. When success is measured as payoff in a game played against another learner, mutual selfishness typically fails to produce the optimal outcome for a pair of individuals. However, learners often operate in populations, and each learner may have a limited duration of interaction with any other individual. Here, we compare selfish learning in stable pairs to selfish learning with stochastic encounters in a population. We study gradient-based optimization in repeated games like the prisoner's dilemma, which feature multiple Nash equilibria, many of which are suboptimal. We find that myopic, selfish learning, when distributed in a population via ephemeral encounters, can reverse the dynamics that occur in stable pairs. In particular, when there is flexibility in partner choice, selfish learning in large populations can produce optimal payoffs in repeated social dilemmas. This result holds for the entire population, not just for a small subset of individuals. Furthermore, as the population size grows, the timescale to reach the optimal population payoff remains finite in the number of learning steps per individual. While it is not universally true that interacting with many partners in a population improves outcomes, this form of collective learning achieves optimality for several important classes of social dilemmas. We conclude that naïve learning can be surprisingly effective in populations of individuals navigating conflicts of interest.

1. Introduction

Population-based methods of optimization and learning have received considerable attention over the years. These methods are often inspired by natural selection and attribute "fitness" to candidate solutions, with higher fitness corresponding to better solutions, for a given problem at hand. Genetic algorithms are classical examples, which encode solutions using an alphabet (e.g. as binary strings) and are amenable to crossover, mutation, and fitness-based reproduction [1,2]. Applications are broad, although these algorithms are especially common in problems with an overwhelming number of candidate solutions, such as the traveling salesman problem [3] or the search for appropriate weights and topologies in artificial neural networks ("neuroevolution") [4].

In some classes of population-based optimization, the population itself is structured according to the search space, with individuals corresponding to locations in parameter space. Particle swarm optimization, for example, involves solutions moving through parameter space as "particles" with some velocity [5]. The velocity is influenced by the best solutions found to date, both by the individual particle and by the population as a whole (or at least locally within a neighborhood of the particle). One reason why this approach is successful is that it promotes exploration, as particles tend to overshoot local optima due to momentum [6].

In such metaheuristics, populations have proven remarkably effective in finding good solutions to computationally hard problems. But these problems are frequently characterized by static fitness landscapes, meaning the presence of other agents does not affect the viability of a solution found by a particular agent. In contrast, the goal of multi-agent reinforcement learning is to optimize an agent's success or fitness in the presence of other agents, who themselves might be carrying out a similar kind of optimization [7,8]. A standard model for this kind of problem is a Markov game, which involves repeated interactions among a group of individuals in some (possibly changing) environment [9,10].

Even in a simple setting without a dynamic environment, multi-agent learning is far from completely understood. A simple example is a repeated game between two individuals. Folk theorems [11,12] imply that such games often have a rich variety of equilibria, which is most striking when the strategic interaction is "mixed", meaning it is neither purely cooperative

* Corresponding author.
E-mail address: alexmcavoy@gmail.com (A. McAvoy).
[1] All authors designed research, performed research, and wrote the paper.

https://doi.org/10.1016/j.physd.2022.133426
0167-2789/

(aligned incentives) nor purely competitive (zero-sum) [13]. The traditional example is the repeated prisoner's dilemma, which has only detrimental equilibria when the game horizon is short, but many different (partly) cooperative equilibria when the horizon is longer [14]. The ability of long interaction horizons to reinforce prosocial behaviors has been referred to as the "shadow of the future" [15], and it is one of the central tenets of the theory of direct reciprocity [16,17].

Given the rich set of equilibria in repeated games, there remains the question of whether two self-concerned learners will jointly arrive at good outcomes. The answer, perhaps unsurprisingly, is no–at least not reliably. Mutual selfishness means that agents often cannot overcome the temptation to exploit the opponent, eroding prosocial actions and leading to detrimental outcomes for both individuals, despite the existence of cooperative equilibria. A recent study [18] proposes a clever way to overcome this drawback of selfish ("naïve") optimization by attempting to shape the co-player's future optimization problem. This learning rule, dubbed "learning with opponent-learning awareness", can better align incentives of two players by utilizing a short look-ahead into the co-player's learning, provided the co-player's learning rule is known. However, as in much of the work on multi-agent learning, this learning rule is implemented under the assumption that learners interact in stable pairs, which means that an agent spends a significant amount of time optimizing against a fixed opponent (who is also learning).

In a population of learners, however, there is no guarantee that any given encounter will be long enough to resemble stable learning. An agent might enter the world with little training, equipped only with a learning rule and a rough idea of how to engage with others. Encounters might be ephemeral, requiring a learner to efficiently use limited information to update its style of play, which is then brought forth into future engagements with different partners. Moreover, the goal in designing a learning algorithm may not be to produce a single agent who performs well, but rather an entire population that plays well with one another. Automated vehicles are agents like this, which are extensively trained prior to ever interacting with another car or person in the real world but must nonetheless sense and adjust to previously unseen encounters with other vehicles and pedestrians [19]. The goal of training for automated vehicles is assuredly not to produce a single vehicle capable of optimal performance, disregarding the performance of other vehicles — but, rather, to produce an entire population of agents who all perform well. Likewise human beings, whose capabilities are often targeted as goals of and inspiration for artificial intelligence, also learn from many different partners throughout life, and they may benefit from a population that collectively performs well.

In this study, we consider a blended population-based framework, with learning in place of strategy transmission and transient encounters in place of stable pairs. We disregard "vertical" transmission and consider how social encounters influence the behaviors of individuals within a fixed population of learners. The population has no central planner or controller to enforce sophisticated social preferences; there are only random encounters and selfish (greedy) strategy updates. Unlike population-based optimization techniques on static landscapes, in which an outcome might be considered successful even if only a small portion of the population finds an optimal solution (including genetic algorithms and some swarm intelligence methods [20]), our goal here is not to produce a small subset of fit individuals in a population. Rather, we seek to understand how social learning in dynamic optimization landscapes affects members of a population collectively.

Even when selfish learning in stable pairs leads to poor outcomes, as in the repeated prisoner's dilemma, we will show that selfish learning in a population with ephemeral partnerships can collectively attain the maximum possible average payoff. Moreover, the timescale to attain such an optimum does not necessarily slow down for any particular individual as the population size (and thus the number of potential partners) grows. We conclude that sociable, selfish optimizers, who learn little from any one encounter, can nonetheless be quite effective and efficient promoters of prosperity and collective learning within a population. In particular, the ability of agents to have a variety of partners within a population can drastically improve the efficacy of selfish learning in the presence of conflicts of interest.

2. Model

The players in our model learn how to behave in repeated games. Each learning step consists of several rounds of a repeated game, followed by a strategy update step. When player X interacts with player Y, they play a repeated game with a stochastic number of rounds determined by a continuation probability, $\lambda < 1$. Each round is of a one-shot (stage) game with actions A and B and payoffs

$$\begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} a_{AA} & a_{AB} \\ a_{BA} & a_{BB} \end{pmatrix} \end{array}. \tag{1}$$

In Eq. (1), the values indicate what payoff the row player receives when facing the column player. After each round, the game continues with probability λ and ends with probability $1-\lambda$. Here, we care about repeated games with sufficiently long time horizons, meaning λ is close to 1. We take $\lambda = 0.9999 < 1$ for an average of 10^4 rounds in the repeated game, both because all interactions are finite (even if long) and because the limit $\lambda \to 1$ can introduce technical issues (see Section SM1).

A memory-one strategy for X in this repeated game consists of an initial probability of playing A, $p_0 \in [0,1]$, together with a four-tuple of conditional probabilities, $(p_{AA}, p_{AB}, p_{BA}, p_{BB}) \in [0,1]^4$, where p_{xy} indicates X's probability of playing A when, in the previous round, X played x and Y played y. We denote a player's memory-one strategy, $(p_0, (p_{AA}, p_{AB}, p_{BA}, p_{BB}))$, by $\mathbf{p} \in [0,1]^5$ to simplify notation. We denote by $\pi(\mathbf{p}, \mathbf{q})$ the payoff to an individual using $\mathbf{p}$ against an opponent using $\mathbf{q}$. How this payoff is calculated, together with its functional form, is detailed in Section SM1.

A learning rule consists of *(i)* an initial distribution over strategies, representing a prior policy before any learning takes place, and *(ii)* an update rule for how to adjust this policy after an encounter. Since we consider memory-one strategies here, the initial policy is chosen from a distribution on $\mathbf{p} \in [0,1]^5$. In practice, we consider two kinds of initial distributions. The primary distribution we analyze chooses each of the five coordinates independently from an arcsine (Beta $(1/2, 1/2)$) distribution. For each coordinate, this distribution explores the corners of $[0,1]$ better than a uniform distribution does, and it is well-established in models of repeated interactions [21]. However, a uniform distribution is also a natural choice, and so we also consider examples in which each coordinate is sampled uniformly (and independently) in $[0,1]$.

Suppose that X and Y are selfish learners paired to interact. After the repeated game is completed, the players perform a gradient-based strategy update step, with both players updating simultaneously. If X has strategy $\mathbf{p}$ and Y has strategy $\mathbf{q}$ prior to this update, then their respective strategies after one gradient step are

$$\mathbf{p}' = \mathrm{proj}\left(\mathbf{p} + \eta \nabla_{\mathbf{p}} \pi(\mathbf{p}, \mathbf{q})\right); \tag{2a}$$

$$\mathbf{q}' = \mathrm{proj}\left(\mathbf{q} + \eta \nabla_{\mathbf{q}} \pi(\mathbf{q}, \mathbf{p})\right), \tag{2b}$$

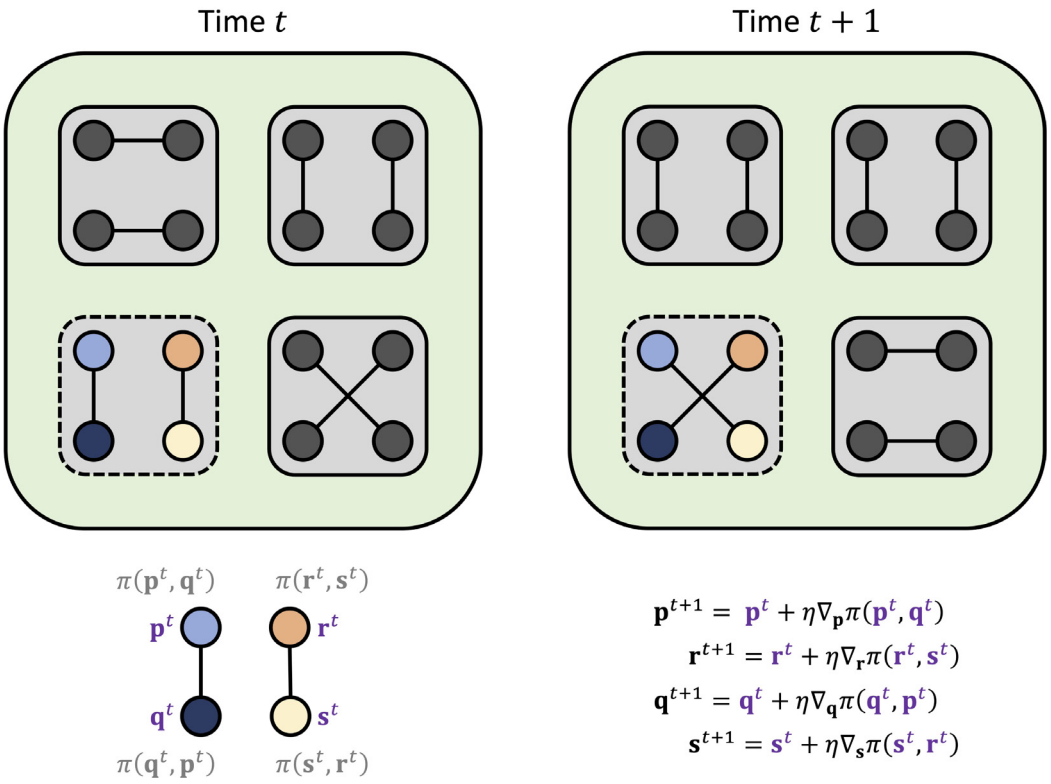

Fig. 1. Learning in isolated sub-populations. The figure depicts a population of N total individuals arranged in d sub-populations, each of size n ($d = n = 4$ and $N = dn = 16$ shown here). Individuals are paired randomly within each sub-population, receive scores from their interactions (based on π), and perform gradient-based strategy updates. The updated strategies are then brought forth into subsequent encounters, which may involve new partnerships (depicted here, for example, in the dotted sub-population). In two extreme cases, individuals interact with the same partner in every time step (stable pairs, $n = 2$) or with potentially any other member of the entire population (well-mixed population, $d = 1$). To simplify notation, the projection operator used in the gradient update is omitted.

where η is a small learning rate and proj denotes the projection operator onto the cube $[0, 1]^5$, which ensures that strategies remain in $[0, 1]^5$. Note that even if two learners both update strategies using Eq. (2), we consider their learning rules to be different if they have different initial distributions.

To describe how learners are paired with partners, we consider a population of d disjoint sub-populations, each of size n, where n is even. The total population size is thus $N = dn$. When $n = 2$, all pairings are between the same two individuals, and the model reduces to a population whose individuals learn in stable pairs. In each time step, individuals are randomly paired within each sub-population to interact and update their strategies via Eq. (2). In the next time step, players are again paired randomly within each sub-population, bringing their new strategies from the previous time step, and the process repeats. Thus, there are two timescales in our model: one timescale for repeated actions in a one-shot game, and one timescale for the strategy updates ("learning") following the repeated game. To make the distinction clear, we refer to the repeated one-shot games as "interactions" and the combination of pairing, a repeated game, and a strategy update as an "encounter". This population-based learning process is depicted schematically in Fig. 1.

For fixed population size $N = dn$, we are interested in comparing $n = 2$ to $n > 2$ to understand the effects of distributed learning in a population compared to the standard case of stable pairings. When $n > 2$, an individual's partner can change across encounters. An alternative approach would be to set $d = 1$ and study the effects of varying N ($= n$). Apart from one example, we have avoided this approach because it involves the comparison of populations of different size. We care about the mean payoff of the population, and changing N while keeping d fixed would involve averaging over a different number of individuals. In contrast, with N fixed and n varying, the population always starts off with the same distribution of mean payoff prior to learning, regardless of n, which permits a more straightforward understanding of the effects of randomness in partner choice on learning.

At any time, everyone has a strategy for the repeated game. When they are paired to play repeated games based on these strategies, the average payoff in each pair satisfies

$$\min_{x,y\in\{A,B\}}\left\{\frac{a_{xy}+a_{yx}}{2}\right\} \leqslant \frac{\pi(\mathbf{p},\mathbf{q})+\pi(\mathbf{q},\mathbf{p})}{2} \leqslant \max_{x,y\in\{A,B\}}\left\{\frac{a_{xy}+a_{yx}}{2}\right\}. \tag{3}$$

As a result, the mean payoff in the population also lies in this interval. The main question we study is how increasing n (with N fixed) affects this mean population payoff in the long run.

3. Results

We analyze our model of collective learning using two complementary approaches, "Lagrangian" and "Eulerian", borrowing these terms from fluid dynamics [22]. In the Lagrangian approach, we track the behaviors of individual learners in finite populations. Here, we are not concerned with quantitative effects of learning as a function of population size; instead, we wish to explore what the learning process looks like when the population is reasonably large. In particular, we will use the Lagrangian perspective to show that collective learning can produce optimal or near-optimal solutions (e.g. in prisoner's dilemma games) in populations of even moderate size. In the Eulerian approach, we consider the density of regions within the strategy space in an infinite population. This approach is used to take

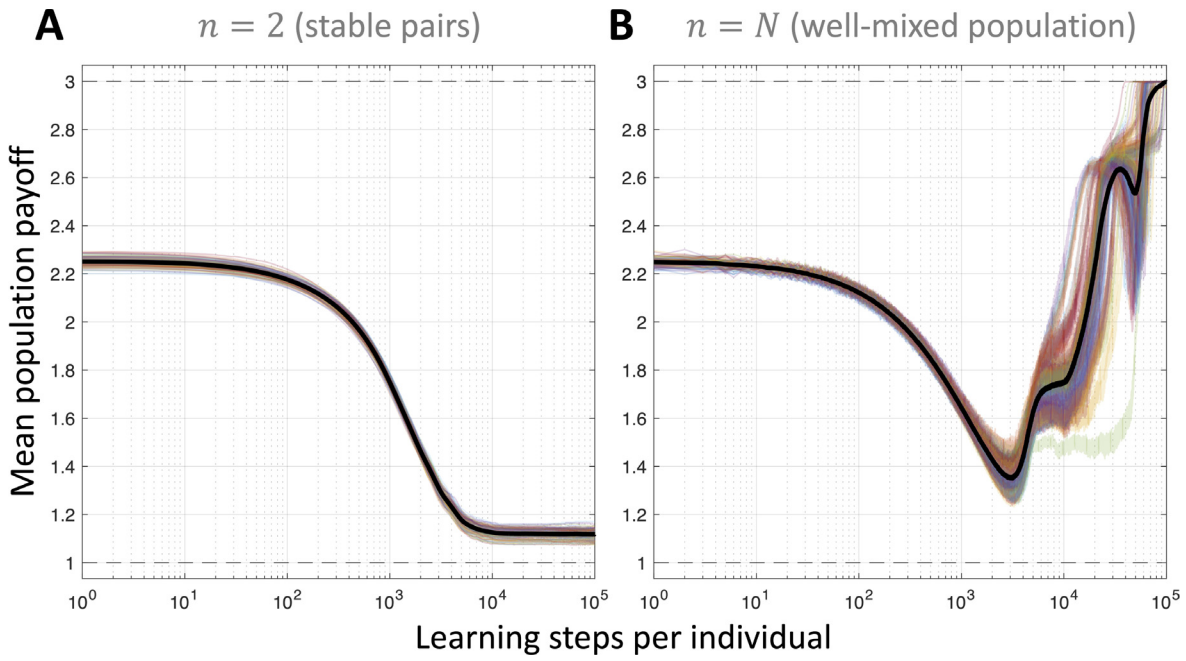

Fig. 2. Pairwise versus collective learning in a standard prisoner's dilemma. In both panels, each trajectory (colored) represents the mean payoff of a population of $N = 1000$ players as learning unfolds. In **A**, the population is sorted once into $d = N/2$ sub-populations, each of size two, and learning takes place only within these sub-populations. Initial strategies are chosen randomly from an arcsine distribution. In **B**, there is only one sub-population, and random pairing within this sub-population occurs prior to every learning step. The curves are generated by calculating the payoffs of paired partners at each step using their current strategies, and then taking the average payoff of the population. Unlike the case $d = N/2$, in which all individuals interact with a fixed learning partner throughout the process, in **B** the outcome for the population is much better. Following an equilibration period in which the mean population payoff declines, the mean payoff begins to increase (non-monotonically) until it hits the maximal social value of a_{AA} $(= 3)$. The mean over all 100 runs is shown in black, on top of each individual run (colored). In both panels, the game parameters are $(a_{AA}, a_{AB}, a_{BA}, a_{BB}) = (3, 0, 5, 1)$, the continuation probability is $\lambda = 0.9999$, and the learning rate is $\eta = 10^{-3}$. The dashed lines indicate the minimum and maximum possible average payoffs, respectively (see Eq. (3)).

the large-population limit, and it produces results that reaffirm the qualitative findings obtained using the Lagrangian approach. The Eulerian approach will also reveal an important result on the timescale of learning in prisoner's dilemma games: convergence to the optimal population outcome requires only finitely many learning steps per individual, even when the population is arbitrarily large.

3.1. Lagrangian perspective: tracking individual agents

Within the space of all payoff matrices in the form of Eq. (1), there are a several distinguished classes of games. The most common example is a prisoner's dilemma, which satisfies $a_{BA} > a_{AA} > a_{BB} > a_{AB}$. This payoff ranking implies that if the game is played once, B (called "defection") is the unique Nash equilibrium because an individual always improves its payoff by switching from A to B. Both players would prefer the payoff for A (called "cooperation") against A to the payoff for B against B. When the game is repeated, however, multiple equilibria exist and may bring different long-term payoffs, ranging from the small (defecting-type equilibria) to the large (cooperative equilibria).

Selfish learners are not able to reliably find cooperative equilibria when they interact in stable pairs. Fig. 2**A** shows learning trajectories for a population of stable pairs $(n = 2)$ when each pair starts with random memory-one strategies prior to optimization. These trajectories consistently lead the population to a low payoff, close to the minimum of a_{BB}. In contrast, when learning is distributed throughout a population via ephemeral, random encounters, the mean population payoff eventually converges to the maximum possible, a_{AA} (Fig. 2**B**). At this point, all individuals in the population have learned to cooperate in the repeated game, even though each one is seeking only to maximize its own payoff. **Fig. SM1** gives another depiction of Fig. 2**A**, showing the mean payoff trajectory for each pair. A small number of pairs do converge to optimal payoffs, which is reflected by the solid black curve in Fig. 2**A** converging to a value slightly above 1.

Although Fig. 2 depicts two extremes, we note that the sub-population size, n, should be sufficiently large to achieve optimal payoffs reliably. Small values of n greater than 2 need not lead to optimal payoffs for the population (see also Fig. 3). However, there does not appear to be a simple threshold for n and N that guarantees convergence of mean population payoffs to optimal or near-optimal values. Such values are evidently sensitive to the game parameters.

We also consider the prisoner's dilemma when each coordinate of the initial strategy is chosen uniformly at random from $[0, 1]$. **Fig. SM2** illustrates qualitatively similar results for this initial condition: the population still benefits from flexibility in partner choice. Furthermore, although we are concerned mainly with agents who have access to the exact gradients, we also consider an alternative optimization procedure based on random search [23]. Instead of calculating gradients, an agent simply samples a nearby strategy and tests this against the partner. If it improves the agent's payoff, then it is adopted; otherwise, the current strategy is retained. **Fig. SM3** shows, once again, that even for this update rule, collective learning with ephemeral partners still produces excellent outcomes for a population in this repeated prisoner's dilemma.

While it is remarkable that such an outcome can be achieved via selfishness alone, simply by swapping partners sufficiently often, a natural question is whether optimal payoffs will also be attained in different forms of prisoner's dilemmas, when mutual cooperation is inefficient. In particular, when $a_{AA} < (a_{AB} + a_{BA})/2$, players fare better with a policy of alternation rather than mutual cooperation—that is, for one player to cooperate in even rounds and defect in odd rounds, and for the other

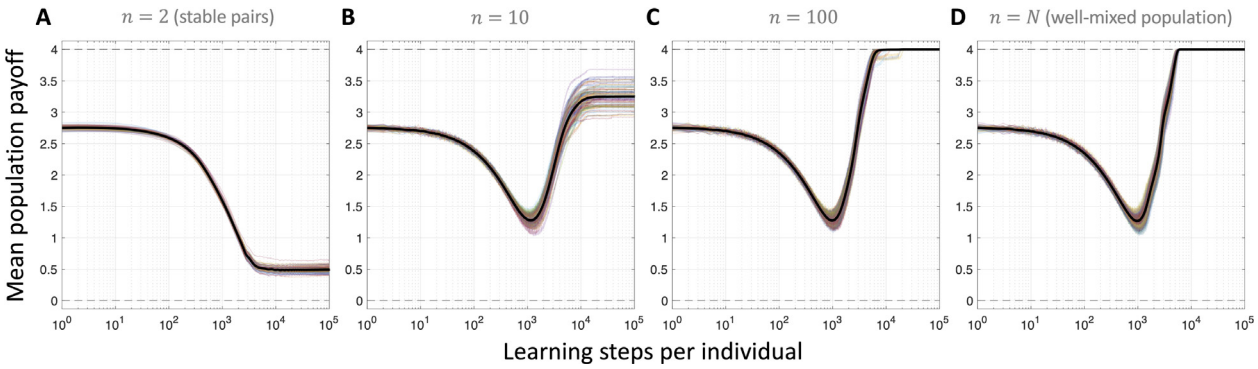

Fig. 3. Collective learning in an alternating prisoner's dilemma. All four panels depict a repeated prisoner's dilemma with $(a_{AB} + a_{BA})/2 > a_{AA}$, which means that mutual cooperation is no longer optimal. Rather, a policy of strictly alternating cooperation and defection, with X cooperating in even rounds only and Y cooperating in odd rounds only, is the best social outcome. Although sorting individuals into $d = N/2$ sub-populations of stable pairs still results in bad outcomes (primarily defection), as the population becomes more well-mixed, the optimal outcome of $(a_{AB} + a_{BA})/2$ $(= 4)$ is achieved. Thus, this collective learning process is not simply discovering how to cooperate mutually—it can also find superior outcomes in interactions for which more complex coordination is required. In all panels, the population size is $N = 1000$, the number of runs is 100, the game parameters are $(a_{AA}, a_{AB}, a_{BA}, a_{BB}) = (3, -1, 9, 0)$, the continuation probability is $\lambda = 0.9999$, and the learning rate is $\eta = 10^{-3}$.

player to cooperate in odd rounds and defect in even rounds. Fig. 3 shows how increasing the size of the sub-populations (while keeping N fixed) in this kind of prisoner's dilemma influences the mean population payoff. When $n = 2$ (stable pairs), most runs converge to defection. As n grows, the population finds cooperative strategies, and for yet larger n the population discovers policies of alternation, which produce the maximum possible mean payoff. As in Fig. 2, throughout the learning process the population initially experiences a decline in payoffs before reaching a trough, beyond which the mean payoff rises until it reaches $(a_{AB} + a_{BA})/2$.

It is also notable in this example that, for fixed N, increasing group size n does not necessarily lead to a longer timescale for achieving the optimal population payoff. Increasing n means more randomness in partner selection, and so one might expect slower convergence to a final outcome. And yet, for this repeated prisoner's dilemma, a larger group actually hastens the ascent to the maximal mean payoff.

Both Figs. 2 and 3 depict prisoner's dilemmas and suggest that *(i)* populations of randomly-interacting selfish learners can attain qualitatively better outcomes than stable selfish learning pairs and *(ii)* for fixed N, the mean population payoff is a non-decreasing function of n. Given this finding in strong social dilemmas, it is natural to ask whether different behavior is observed in other classes of interactions. A weaker social dilemma [24] called the snowdrift game, for example, is similar to the prisoner's dilemma except that the payoff ranking is $a_{BA} > a_{AA} > a_{AB} > a_{BB}$ instead of $a_{BA} > a_{AA} > a_{BB} > a_{AB}$. As a result, B is no longer a dominant strategy. There are two primary variants of the snowdrift game, one with $(a_{AB} + a_{BA})/2 < a_{AA}$ and one with $(a_{AB} + a_{BA})/2 > a_{AA}$. In both cases, collective learning via random encounters within a sub-population results in optimal payoffs. (In the latter case, this is also true for learning within stable pairs.) We also observe optimal outcomes in stag hunt games $(a_{AA} > a_{BA} > a_{BB} > a_{AB})$ provided $(a_{AB} + a_{BA})/2 > a_{BB}$. See **Fig. SM4** for details.

A population of selfish, ephemeral learners is not optimal for all types of games, however. The battle of the sexes game [25,26] provides a contrast to the preceding results. This game is characterized by two players with competing preferences for what to do (e.g. what kind of event to attend). We consider a symmetric version of this game (also known as the "hero" game [27]), in which the action A may be thought of as "go with own preference" while B is "go with partner's preference". But they would prefer to be together than apart, which leads to a coordination game with payoff ranking $a_{AB} > a_{BA} > a_{AA} > a_{BB}$. Fig. 4**A** shows how learning in stable pairs quickly brings a population close to its optimal average payoff. But this is not the case for a population of randomly-interacting learners (Fig. 4**B**), which may be attributed to the difficulty of achieving consistent anti-coordination when learning partners are not fixed. We note that this kind of anti-coordination is different from that of the alternating prisoner's dilemma (Fig. 3). Due to the nature of the underlying one-shot game, a pair of selfish learners in the battle of the sexes games tends to arrive at asymmetric outcomes, with one player receiving a_{AB} and the other getting a_{BA}. In contrast, the policies of anti-coordination in the alternating prisoner's dilemma give both players $(a_{AB} + a_{BA})/2$. The average pair payoff is $(a_{AB} + a_{BA})/2$ in both cases, but in the former it is more difficult to achieve a policy of anti-coordination in a population.

Finally, there are both prisoner's dilemma and stag hunt games satisfying $(a_{AB} + a_{BA}) < a_{BB}$. For these games, neither learning in stable pairs nor collective learning efficiently achieves a maximum average payoff (see **Fig. SM5**). As a result, across a variety of different kinds of games, learning in stable pairs is not universally better than collective learning or vice versa. But they do generally result in very different outcomes — and collectively learning can often produce optimal outcomes where learning in stable pairs does not.

3.2. Eulerian perspective: tracking distributions of agents

Our goal has been to compare two extremes: learning in stable pairs and collective learning in large populations. So far, we have focused on large but finite populations and an infinite strategy space. We now turn to a dual approach, using an infinite population and a discretized, finite strategy space. Specifically, we focus on a single sub-population $(n = N \to \infty)$, motivated by a desire to understand the extreme case in which two learners never interact more than once.

Let ρ be the density of memory-one strategies in the infinite population. Although the space of memory-one strategies may be identified with $[0, 1]^5$, it is convenient to think of ρ as a density on $\mathbb{R}^5$ that is supported on $[0, 1]^5$. At time t, the gradient direction of an individual using strategy $\mathbf{p}$ is a random variable, $\boldsymbol{V}[\rho](\mathbf{p})$, with $\rho(\mathbf{q}, t)$ the density of direction $\nabla_{\mathbf{p}}\pi(\mathbf{p}, \mathbf{q})$. This gradient field requires defining the payoff, π, outside of $[0, 1]^5$, but the extension of π to $\mathbb{R}^5$ is irrelevant due to the support of ρ.

To ensure that the support of the density is always contained in $[0, 1]^5$ for all t whenever it is at time $t = 0$, we need a modified

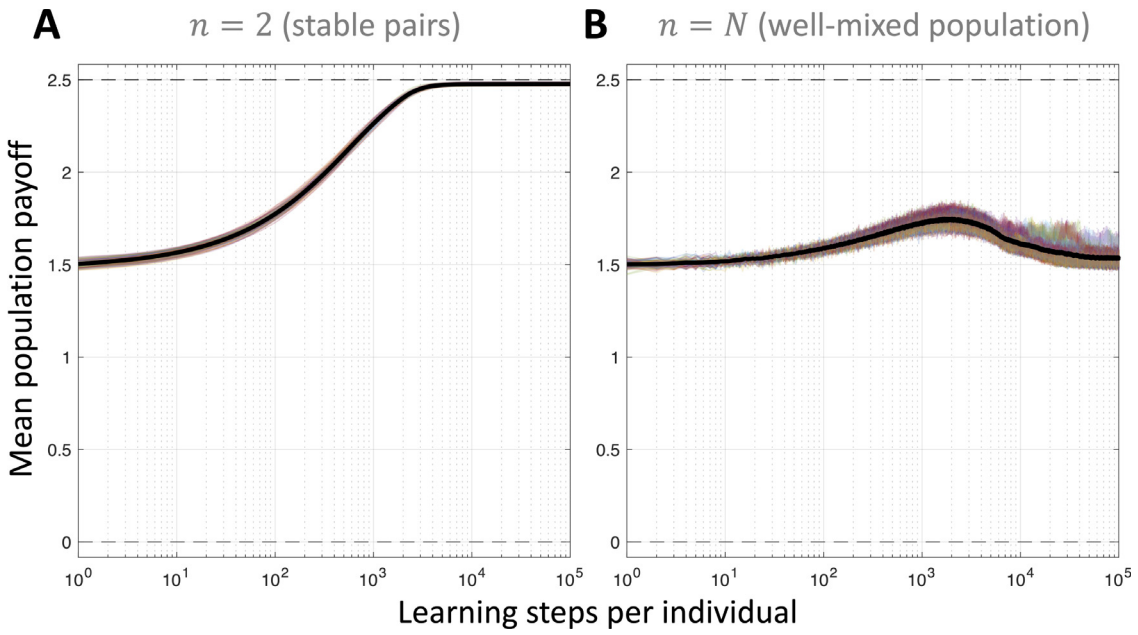

Fig. 4. Pairwise versus collective learning in the battle of the sexes game. In contrast to repeated prisoner's dilemmas, learning in stable pairs leads to near-optimal outcomes for a population in the repeated battle of the sexes game, whereas collective learning has difficulty achieving and maintaining a high mean payoff. In both panels, the population size is $N = 1000$, the number of runs is 100, the game parameters are $(a_{AA}, a_{AB}, a_{BA}, a_{BB}) = (1, 3, 2, 0)$, the continuation probability is $\lambda = 0.9999$, and the learning rate is $\eta = 10^{-3}$.

gradient field of the form $\widetilde{\boldsymbol{V}} = W\boldsymbol{V}$ for some matrix W, which represents the projection operator in Eq. (2). We wish for W to be the identity matrix away from the boundary, but for it to ensure the gradient is always inward pointing on $[0, 1]^5$ to retain the salient features of the search process. We take W to be a diagonal matrix depending on both $\mathbf{p}$ and $\mathbf{q}$ (the partner strategy of $\mathbf{p}$) on the boundary, with $W_{ii} = 1$ unless either *(i)* $p_i = 0$ and $\left(\nabla_{\mathbf{p}}\pi\left(\mathbf{p}, \mathbf{q}\right)\right)_i < 0$ or *(ii)* $p_i = 1$ and $\left(\nabla_{\mathbf{p}}\pi\left(\mathbf{p}, \mathbf{q}\right)\right)_i > 0$, in which case $W_{ii} = 0$.

With this modified gradient field, we consider a continuous-time process in which a learner randomly chooses a partner, computes the gradient direction $\widetilde{\boldsymbol{V}}\left[\rho\right]\left(\mathbf{p}^t\right)$, and then lets its strategy at time $t + \Delta t$ be $\mathbf{p}^{t+\Delta t} = \mathbf{p}^t + (\Delta t)\,\widetilde{\boldsymbol{V}}\left[\rho\right]\left(\mathbf{p}^t\right)$. Thus, a time step of $\Delta t = \eta$ corresponds to the discrete-time model presented previously. In the limit as $\Delta t \to 0$, the strategy density evolves according to the advection equation

$$\frac{\partial \rho\left(\mathbf{p}, t\right)}{\partial t} = -\nabla_{\mathbf{p}} \cdot \left(\rho\left(\mathbf{p}, t\right) \int_{\mathbf{q}\in\mathbb{R}^5} W\left(\mathbf{p}, \mathbf{q}\right) \nabla_{\mathbf{p}}\pi\left(\mathbf{p}, \mathbf{q}\right) \rho\left(\mathbf{q}, t\right)\, d\mathbf{q} \right). \tag{4}$$

Notably, this partial differential equation depends on only the mean (projected) gradient direction, which is important because in the mechanistic description of the model we do not require an individual to interact with a large subset of the population. There is no notion of a "mean" player in the definition of the model, and only one encounter takes place per individual prior to each learning step. We include a more detailed description of the Eulerian approach in Section SM2. In addition to the PDE, we also derive the large-N limit of the system describing the dynamics in discrete time (Eq. SM16).

The trade-off of this approach is that, to work with this equation numerically, we must discretize the strategy space. To simplify this numerical problem, we study lower-dimensional strategy spaces. If $\mathbf{p}$ is a "reactive" strategy, then p_{xy} depends on y only and not on x [28]. Thus, $p_{AA} = p_{BA} \coloneqq p_A$ and $p_{AB} = p_{BB} \coloneqq p_B$, so $\mathbf{p}$ is specified by a triplet, $(p_0, (p_A, p_B)) \in [0, 1]^3$. If we also assume that the game length is infinite (taking the limit $\lambda \to 1$), then the initial probability of playing A is irrelevant and payoffs depend on $\mathbf{p} = (p_A, p_B) \in [0, 1]^2$ (see Eq. SM3).

Reducing the sophistication of strategies for repeated games also limits both transient dynamics and long-term outcomes. For example, the rigidity of reactive strategies makes it more difficult to efficiently achieve policies of alternating cooperation in prisoner's dilemma games with $a_{AA} < (a_{AB} + a_{BA})/2$ (c.f. Fig. 3). But reactive strategies do preserve the qualitative behavior seen in the standard prisoner's dilemma of Fig. 2, and so we can use this simpler strategy space to study collective learning from a Eulerian perspective.

Fig. SM6 illustrates the numerical scheme to compute how ρ, the density of strategies, changes over time. In Fig. 5, we see results from the PDE that are in close qualitative agreement with individual-based (that is, Lagrangian) simulations. In addition to providing a different perspective on the learning problem (involving a different kind of calculation), the Eulerian approach also illustrates that the learning timescale does not diverge as $N \to \infty$ in repeated prisoner's dilemmas. In the Lagrangian (individual-based) approach, we observed relatively quick convergence to the mean payoff maximum in prisoner's dilemma games when $n = N \gg 0$, and indeed this is also reflected in the solution of the Eulerian PDE (Fig. 5).

The evolution of the density leading to Fig. 5 is depicted in **Vid. SM1**, as well as in snapshots in Fig. 6. This density reveals that the initial dip in mean payoff is due to a general trend toward unconditional defection ($\mathbf{p} = (0, 0)$); see Fig. 6**A**. The resulting decline in payoffs is consistent with what is seen for pairs (Fig. 2**A** and **Fig. SM1**) following a random initial state. However, some of the density remains concentrated near $(1, 0)$. Since these strategies reciprocate cooperation and punish defection, they are capable of incentivizing cooperation in selfish agents, even in those who tend to defect with high probability. As a result, we observe a wave develop outwards from unconditional defection toward more cooperative strategies (Fig. 6**B**), which increases the mean population payoff. And even though some of these incentivizing strategies are present in the early stages of the process, at that time most individuals are likely to interact with strategies

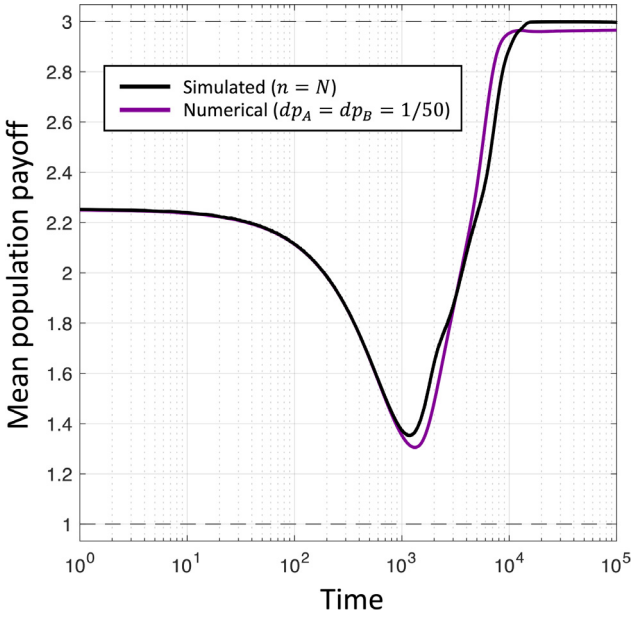

Fig. 5. Learning dynamics of reactive strategies in an infinite population. Solving Eq. (4) numerically (**Fig. SM6**), we plot the mean population payoff over time when players use a simple class of reactive strategies. The mean payoff (purple) predicted by the PDE closely tracks the mean payoff observed in agent-based simulations (black) when $n = N = 1000$. The payoffs in the numerical scheme end up slightly below the maximum payoff of 3, due to the use of a finite grid for the strategy space, obtained by dividing $[0, 1]$, the domain of each coordinate p_A and p_B, into 50 equally spaced subintervals. The time scale is chosen such that one unit of time in the continuous model corresponds to one time step in the discrete model.

that cannot elicit cooperation; as a result, learners, on average, tend to decrease their cooperation levels more frequently than they increase them, leading to a net degradation of payoff. Once defection is predominant, individuals near $(1, 0)$ benefit from being both more cooperative (increasing p_A) and more forgiving (increasing p_B). This, in turn, incentivizes cooperation in the rest of the population, provided there is not too much forgiveness for defection (Fig. 6**C**).

4. Discussion

In this study, we have compared selfish learning with stable pairs to selfish learning with stochastic encounters in a population to determine *(i)* whether there are differences in outcomes between these two learning processes and *(ii)* whether observed differences are robust across all classes of interactions. The result of our investigation can be broadly summarized as saying "yes" to question *(i)* and "no" to question *(ii)*.

Our study is related to several existing lines of inquiry in the literature on learning and evolutionary dynamics. The first is evolutionary dynamics in a single sub-population ($n = N$), where similar strategic trajectories have been noticed in populations of replicators who "learn" by copying more successful individuals. In that setting, Nowak and Sigmund [29] observed a transition from random reactive strategies, to unconditional defection, and then to tit-for-tat (copy the opponent), and finally to generous tit-for-tat (copy the opponent, but forgive errors). Indeed, the population payoff trajectories found in that context [see29, Fig. 1] resemble the qualitative behavior seen in Figs. 2 and 5. However, in stark contrast to the replicators of Nowak and Sigmund [29], our model is based on individual learning through steps of gradient ascent (locally rational, selfish updates), as opposed biased imitation of fitter individuals. In our setting, when two players meet, neither one replicates the strategy of the other, regardless of how successful the two strategies are; in fact, the learning process may bring a player's strategy further away from their opponent's. This is a fundamental difference between the learning process we consider here and models based on biased replicators. Although biased replication of successful strategies can be seen as a simplistic form of learning, the process that we study is completely different in that it requires individuals to have cognitive ability, to be aware of the payoffs in the game they are playing, and to compute how they can improve their payoff against their current opponent by a small change in strategy.

There is another important distinction between our study and the literature on traditional replicator models of strategy dynamics: whereas populations are frequently used to study the evolution of cooperation [30–35], our focus is decidedly not on cooperation per se. Cooperation is an action in several of the games we consider, but we are concerned with achieving optimal payoffs for a population, regardless of whether those are achieved by cooperation or not. Indeed, this is one reason for studying the version of the prisoner's dilemma shown in Fig. 3, where mutual cooperation is actually suboptimal and collective learning nonetheless achieves the optimal outcome.

Furthermore, we note the distinction between the deterministic formulation described in Eq. (4) and evolutionary invasion analysis [36], which also uses gradient-based dynamics. The latter, sometimes referred to as "adaptive dynamics", concerns the invasion of initially-rare mutant traits in a resident population. Our approach, by contrast, allows for multiple types coexisting in the population, and so it is more akin to the model of population games of Friedman and Ostrov [37] than to invasion analysis, except we use a higher-dimensional strategy space due to the complexity of repeated games. **Vid. SM1** and Fig. 6 demonstrate how the mechanics of strategy evolution in our learning model differ significantly from those of both evolutionary [29, Fig. 1] and adaptive [38, Fig. 6] dynamics.

Our study is also related to the literature on multi-agent learning. For many problems, the goal might be to understand the effects of learning rules when encounters are decidedly stochastic and ephemeral. Player i might never meet player j again, but the next partner will have faced someone *like* player i in the past. An agent learns from past experiences and adjusts, bringing new behavior into future encounters. In our example of the repeated prisoner's dilemma, we have seen that moving from pairwise stable learning to population-based collective learning allows selfish learning to be quite efficient at attaining optimal

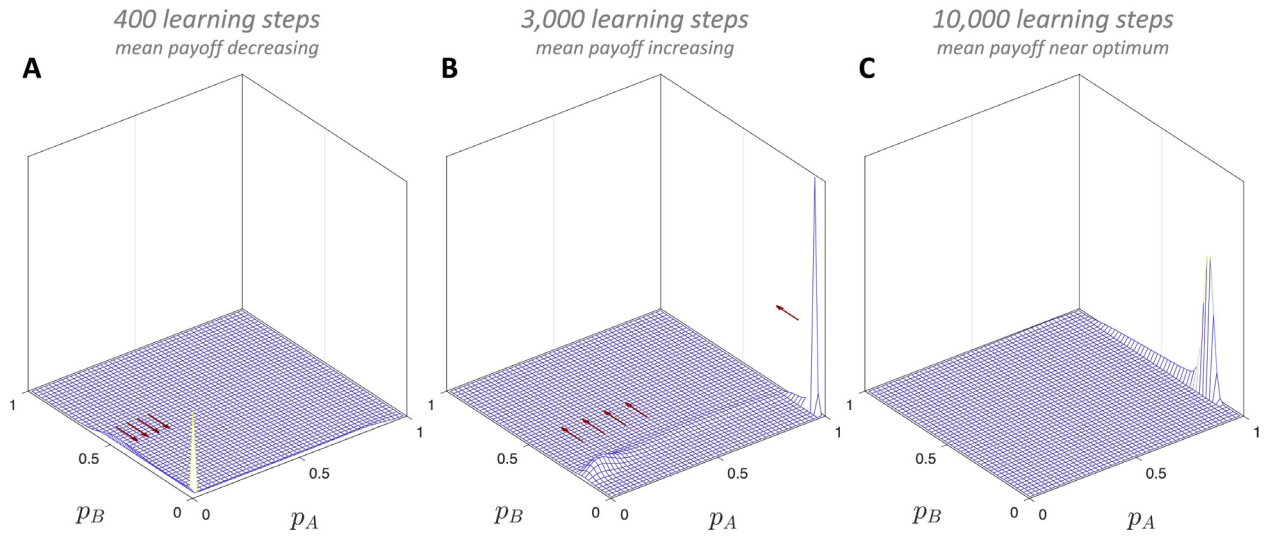

Fig. 6. Snapshots of the strategy density. Panels **A**, **B**, and **C** depict the strategy density in **Vid. SM1** after 400, 3000, and 10,000 learning steps, respectively. Following an initial attraction to defection (**A**), individuals near $\mathbf{p} = (1, 0)$ who reciprocate cooperation and punish defection can incentivize their interaction partners to cooperate. As a result, we observe an outflux from defection in **B**, which increases the mean population payoff. Finally, as the mean population payoff settles near its optimal value (3, in this example), strategies tend to reciprocate cooperation but forgive some amount of defection (p_B positive but not too large). Note that strategies develop a limited amount of forgiveness in order to avoid being exploited.

outcomes for all. But these findings are somewhat sensitive to the interaction type (c.f. Fig. 4) as well as to the exact implementation of the learning rules themselves, and so we make no claim that collective learning is universally more efficient than learning in stable pairs.

Instead, we emphasize that the population-based model is highly relevant to many real-world situations of multi-agent learning, and the outcomes of this learning process depend dramatically on how likely one is to re-encounter the same learning partner. Of course, models of multi-agent learning can be more complex, involving dynamic environments [39–41]. Even social dilemmas can be extended spatially and temporally [42]. Nevertheless, the simplicity of our model, based on repeated matrix games with purely myopic, selfish strategy updates, has some advantages. To understand multi-agent learning in more realistic application cases, we must also understand the effects of a population in simpler (even if idealized) settings. The complexity introduced by stochastic encounters in a population makes this problem difficult even for repeated matrix games, but here it is still tractable. An important area of future research will be to study this problem in more complicated settings, including spatial and temporal extensions.

From our study of repeated matrix games, it is clear that the outcomes of learning in a population will depend on the initial distribution to a significant degree. If all individuals were to start with deterministic policies of playing B (defecting) in every round of a prisoner's dilemma, then no amount of flexibility in partner choice could bring the population out of this state under selfish learning. The distributions we consider (arcsine and uniform) both introduce a large diversity of strategies in the initial population prior to any learning, which we view as important determinants of long-term prosperity. However, we note that even for stable pairs in the iterated prisoner's dilemma, the long-run outcomes cannot be predicted from the initial level of cooperation alone. For example, if two players both initially use unconditional cooperation ("ALLC"), then their optimization process will lead them to mutual defection; but if they both use the strategy tit-for-tat ("TFT"), then they remain at mutual cooperation. This property holds even though ALLC against ALLC results in the same initial level of cooperation as TFT against TFT.

We have focused on repeated games that are sufficiently long because these games are known to have rich spaces of equilibria [11]. While our concern has been payoffs (in particular, mean population payoffs) rather than Nash or subgame-perfect equilibria, long time horizons ensure that current behaviors can be punished or rewarded in the future with high probability, and thus, in principle, there are interesting strategies available for players to learn in the first place. In many classes of interactions, interesting strategies exist even when the length of the game is slightly shorter, and an open question is how the length of the game influences the learning process.

Finally, the population size ties into both the initial distribution and the length of interactions, and it may have effects on long-term learning outcomes. When individuals choose initial strategies at random, larger population sizes lead to a broader variety of behaviors present prior to learning. The effect of this diversity is almost trivial if individuals interact in stable pairs since the mean population payoff then just converges to the expected payoff for a single pair as the population grows (c.f. Fig. 2**A** and **Fig. SM1**). But with larger sub-populations, especially when any two individuals in a population can interact, this strategy variety matters a lot. In such situations, individuals are also less likely to interact in two subsequent time steps if N is large, which leaves open the possibility that somewhat large mean payoffs are possible even for shorter games.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful to Daniel Cooney, Christian Hilbe, Julian Kates-Harbeck, Eleni Katifori, Andrea Liu, Truong-Son Van, and the soft matter group at the University of Pennsylvania for comments on this manuscript and for helpful conversations related to collective learning. This work was supported by the Simons Foundation (Math+X grant to Y.M.), the National Science Foundation (grant

DMS-2042144 to Y.M.), the David & Lucile Packard Foundation (J.B.P.), and the John Templeton Foundation (J.B.P., grant 62281).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.physd.2022.133426.

References

[1] D.E. Goldberg, Artificial intelligence, in: Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Publishing Company, ISBN: 9780201157673, 1989.
[2] L.M. Schmitt, Theory of genetic algorithms, Theoret. Comput. Sci. 259 (1) (2001) 1–61, http://dx.doi.org/10.1016/S0304-3975(00)00406-0.
[3] P. Larrañaga, C.M.H. Kuijpers, R.H. Murga, I. Inza, S. Dizdarevic, Genetic algorithms for the travelling salesman problem: A review of representations and operators, Artif. Intell. Rev. 13 (1999) 129–170, http://dx.doi.org/10.1023/A:1006529012972.
[4] K.O. Stanley, R. Miikkulainen, Evolving neural networks through augmenting topologies, Evol. Comput. 10 (2) (2002) 99–127, http://dx.doi.org/10.1162/106365602320169811.
[5] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of ICNN '95 - International Conference on Neural Networks, IEEE, 1995, http://dx.doi.org/10.1109/icnn.1995.488968.
[6] R. Eberhart, J. Kennedy, A new optimizer using particle swarm theory, in: MHS '95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, IEEE, 1995, http://dx.doi.org/10.1109/mhs.1995.494215.
[7] L. Panait, S. Luke, Cooperative multi-agent learning: The state of the art, Auton. Agents Multi-Agent Syst. 11 (3) (2005) 387–434, http://dx.doi.org/10.1007/s10458-005-2631-2.
[8] P.J. Hoen, K. Tuyls, L. Panait, S. Luke, J.A. La Poutré, An overview of cooperative and competitive multiagent learning, in: Learning and Adaption in Multi-Agent Systems, Springer Berlin Heidelberg, 2006, pp. 1–46, http://dx.doi.org/10.1007/11691839_1.
[9] L.S. Shapley, Stochastic games, Proc. Natl. Acad. Sci. 39 (10) (1953) 1095–1100, http://dx.doi.org/10.1073/pnas.39.10.1953.
[10] M.L. Littman, Markov games as a framework for multi-agent reinforcement learning, in: Machine Learning Proceedings, 1994, Elsevier, 1994 pp. 157–163, http://dx.doi.org/10.1016/b978-1-55860-335-6.50027-1.
[11] D. Fudenberg, E. Maskin, The folk theorem in repeated games with discounting or with incomplete information, Econometrica 54 (3) (1986) 533, http://dx.doi.org/10.2307/1911307.
[12] M. Barlo, G. Carmona, H. Sabourian, Repeated games with one-memory, J. Econom. Theory 144 (1) (2009) 312–336, http://dx.doi.org/10.1016/j.jet.2008.04.003.
[13] K. Zhang, Z. Yang, T. Başar, Multi-agent reinforcement learning: A selective overview of theories and algorithms, 2019, arXiv preprint arXiv:1911.10635.
[14] M.A. Nowak, Science 314 (5805) (2006) 1560–1563, http://dx.doi.org/10.1126/science.1133755.
[15] R. Axelrod, The Evolution of Cooperation, Basic Books, 1984.
[16] R.L. Trivers, The evolution of reciprocal altruism, Q. Rev. Biol. 46 (1) (1971) 35–57, http://dx.doi.org/10.1086/406755.
[17] C. Hilbe, K. Chatterjee, M.A. Nowak, Partners and rivals in direct reciprocity, Nat. Hum. Behav. (2018) http://dx.doi.org/10.1038/s41562-018-0320-9.
[18] J. Foerster, R.Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, I. Mordatch, Learning with opponent-learning awareness, in: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, in: AAMAS '18, International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 122–130.
[19] P.A. Hancock, I. Nourbakhsh, J. Stewart, On the future of transportation in an era of automated and autonomous vehicles, Proc. Natl. Acad. Sci. 116 (16) (2019) 7684–7691, http://dx.doi.org/10.1073/pnas.1805770115.
[20] K. Zielinski, R. Laur, Stopping criteria for a constrained single-objective particle swarm optimization algorithm, Informatica 31 (2007) 51–59.
[21] M. Nowak, K. Sigmund, A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game, Nature 364 (6432) (1993) 56–58, http://dx.doi.org/10.1038/364056a0.
[22] G.K. Batchelor, An Introduction To Fluid Dynamics, Cambridge Mathematical Library. Cambridge University Press, 2000, http://dx.doi.org/10.1017/CBO9780511800955.
[23] F.J. Solis, R.J.-B. Wets, Minimization by random search techniques, Math. Oper. Res. 6 (1) (1981) 19–30, http://dx.doi.org/10.1287/moor.6.1.19.
[24] C. Hauert, Effects of space in 2 ×2 games, Int. J. Bifurcation Chaos 12 (07) (2002) 1531–1548, http://dx.doi.org/10.1142/s0218127402005273.
[25] J. Maynard Smith, Evolution and the Theory of Games, Cambridge University Press, 1982, http://dx.doi.org/10.1017/cbo9780511806292.
[26] R.D. Luce, H. Raiffa, Games and Decisions: Introduction and Critical Survey, Dover Books on Mathematics. Dover Publications, ISBN: 9780486659435, 1989.
[27] J. Tanimoto, Fundamentals of Evolutionary Game Theory and Its Applications, Springer, Japan, 2015, http://dx.doi.org/10.1007/978-4-431-54962-8.
[28] E. Kalai, D. Samet, W. Stanford, A note on reactive equilibria in the discounted prisoner's dilemma and associated games, Internat. J. Game Theory 17 (3) (1988) 177–186.
[29] M.A. Nowak, K. Sigmund, Tit for tat in heterogeneous populations, Nature 355 (6357) (1992) 250–253, http://dx.doi.org/10.1038/355250a0.
[30] R. Axelrod, The emergence of cooperation among egoists, Am. Political Sci. Rev. 75 (2) (1981) 306–318, http://dx.doi.org/10.2307/1961366.
[31] M.A. Nowak, A. Sasaki, C. Taylor, D. Fudenberg, Emergence of cooperation and evolutionary stability in finite populations, Nature 428 (6983) (2004) 646–650, http://dx.doi.org/10.1038/nature02414.
[32] F.C. Santos, J.M. Pacheco, Scale-free networks provide a unifying framework for the emergence of cooperation, Phys. Rev. Lett. 95 (9) (2005) http://dx.doi.org/10.1103/physrevlett.95.098104.
[33] M.A. Nowak, Five rules for the evolution of cooperation, Science 314 (5805) (2006) 1560–1563, http://dx.doi.org/10.1126/science.1133755.
[34] G. Szabó, G. Fáth, Evolutionary games on graphs, Phys. Rep. 446 (4) (2007) 97–216, http://dx.doi.org/10.1016/j.physrep.2007.04.004.
[35] L.G. Moyano, A. Sánchez, Evolving learning rules and emergence of cooperation in spatial prisoner's dilemma, J. Theoret. Biol. 259 (1) (2009) 84–95, http://dx.doi.org/10.1016/j.jtbi.2009.03.002.
[36] S.A.H. Geritz, É. Kisdi, G. Meszéna, J.A.J. Metz, Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree, Evol. Ecol. 12 (1) (1998) 35–57, http://dx.doi.org/10.1023/a:1006554906681.
[37] D. Friedman, D.N. Ostrov, Evolutionary dynamics over continuous action spaces for population games that arise from symmetric two-player games, J. Econom. Theory 148 (2) (2013) 743–777, http://dx.doi.org/10.1016/j.jet.2012.07.004.
[38] M.A. Nowak, K. Sigmund, Evolutionary dynamics of biological games, Science 303 (5659) (2004) 793–799, http://dx.doi.org/10.1126/science.1093411.
[39] E. Akiyama, K. Kaneko, Dynamical systems game theory and dynamics of games, Physica D 147 (3-4) (2000) 221–258, http://dx.doi.org/10.1016/s0167-2789(00)00157-3.
[40] Y. Sato, E. Akiyama, J.P. Crutchfield, Stability and diversity in collective adaptation, Physica D 210 (1) (2005) 21–57, http://dx.doi.org/10.1016/j.physd.2005.06.031.
[41] W. Barfuss, J.F. Donges, J. Kurths, Deterministic limit of temporal difference reinforcement learning for stochastic games, Phys. Rev. E 99 (4) (2019) http://dx.doi.org/10.1103/physreve.99.043305.
[42] J.Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, T. Graepel, Multi-agent reinforcement learning in sequential social dilemmas, in: Proceedings of the 16th International Conference on Autonomous Agents and MultiAgent Systems, in: AAMAS '17, International Foundation for Autonomous Agents and Multiagent Systems, 2017, pp. 464–473.