PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research

Cite this article: Bauer J, West S, Alonso E, Broom M. 2026 Mutation-bias learning: an evolutionary game dynamics approach to convergence analysis in multi-agent reinforcement learning. *Proc. R. Soc. A* **482**: 20250449.
https://doi.org/10.1098/rspa.2025.0449

Received: 23 May 2025
Accepted: 7 November 2025

Subject Areas:
applied mathematics, differential equations, artificial intelligence

Keywords:
replicator dynamics, evolutionary games, mutation, multi-agent reinforcement learning

Author for correspondence:
Johann Bauer
e-mail: johann.bauer@tu-darmstadt.de

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.8174357.

THE ROYAL SOCIETY PUBLISHING

Mutation-bias learning: an evolutionary game dynamics approach to convergence analysis in multi-agent reinforcement learning

Johann Bauer[1,3], Sheldon West[2], Eduardo Alonso[2] and Mark Broom[1]

[1]Department of Mathematics, and [2]Department of Computer Science, University of London, City St. George's, UK
[3]Centre for Cognitive Science, TU Darmstadt, Germany

JB, 0000-0003-3635-1235

We pursue a mathematically rigorous approach connecting stochastic multi-agent reinforcement learning (MARL) processes to deterministic dynamical systems of replicator–mutator dynamics (RMD) type from evolutionary game theory (EGT). This dynamical systems perspective makes the rich literature on evolutionary game dynamics directly available for establishing theoretical guarantees for algorithm convergence in complex multi-agent environments, addressing a fundamental challenge in the field. We demonstrate this approach by presenting and analysing mutation-bias learning with direct policy updates (MBL-DPU), a MARL algorithm that provably approximates RMD, and show convergence in stable games. Through experiments across games of increasing complexity and dimensionality, we demonstrate our dynamical systems analysis and the convergence of MBL-DPU while win-or-learn-fast policy-hill-climbing (WoLF-PHC) and frequency-adjusted Q-learning (FAQ), algorithms with fewer theoretical guarantees, deteriorate unexpectedly in higher dimensions. Beyond specific algorithms, the approach demonstrates a principled route to transferring results from EGT to multi-agent learning, allowing a systematic comparison of evolutionary game dynamics with MARL algorithms and enabling the derivation of further MARL algorithms from

evolutionary dynamics. This approach further questions the assumption that algorithm complexity generally improves performance and underscores the necessity of mathematical rigour in analysing MARL algorithms. For a systematic comparison, we also introduce and experimentally analyse mutationbias learning with logistic choice (MBL-LC), a variant closer to Q-learning but lacking the theoretical guarantees of MBL-DPU.

1. Introduction

Reinforcement learning algorithms have been employed in a wide range of problem settings with great success (e.g. [1]), and for the single-agent case the conditions for convergence of, e.g. Q-learning have been clarified [2]. However, for multi-agent reinforcement learning (MARL), questions of convergence are still very much open. Even simple two-player settings, e.g. the rock–paper–scissors (RPS) game, can exhibit chaotic behaviour under simple dynamics [3], and make a rigorous *a priori* analysis challenging. For more complicated algorithms, an analysis beyond experimental evaluation is often hardly possible. However, more general analyses are highly informative of why algorithms behave in a certain way and theoretical guarantees for at least the simplest of settings are highly desirable in order to assess how reliably MARL algorithms will generalize to similar settings.

In particular, as MARL algorithms often lead to stochastic discrete-time dynamic systems, insights from the fields of learning dynamics in games and of evolutionary game theory (EGT) have been particularly relevant. EGT approaches and specifically the established replicator dynamics (RD) have informed a number of constructions or analyses of learning algorithms in multi-agent settings (e.g. [4,5]). The potential of EGT to inform learning algorithms is illustrated, as a particularly prominent example, by the fact that the win-or-learn-fast policy-hill-climbing (WoLF-PHC) learning algorithm [6], keeps track of the past average policy. In light of RD, this is particularly useful, as the time-average policy in RD converges to a Nash equilibrium under self-play in zero-sum games (e.g. [7, proposition 3.6, p. 92]), providing an intuition for how WoLF-PHC can learn Nash equilibria in self-play in a number of settings.

In the spirit of further contributing to understanding the relation between MARL systems and the rich results on evolutionary game dynamics, and building on the relation between RD and a simple form of reinforcement learning, called Cross learning [8,9], we present and analyse two variants of a reinforcement learning algorithm: mutation-bias learning (MBL) with direct policy updates (MBL-DPU)—a least complexity modification of Cross learning—and mutation-bias learning with logistic choice (MBL-LC), which more closely aligns with the softmax policy in reinforcement learning. Our analysis explicitly takes into account the full stochasticity of the problem, and proves rigorously that MBL-DPU can be approximated by a mutation-perturbed RD (equation (RMD)) which we had specified and analysed previously in [10], a nonlinear *continuous-time deterministic* dynamics whose stability properties can still be studied analytically to a certain degree. Although in general, Lyapunov stability and other properties of a continuous-time dynamics do not always transfer to a corresponding discretized dynamics—a prominent example is the RPS game [7]—we show that asymptotic stability in the continuous dynamics does imply the convergence of the MARL algorithm, in spite of non-vanishing discrete step size and stochasticity. Our focus on RMD allows us to avoid a well-known fundamental limitation of the regular multi-population RD, which cannot have asymptotically stable interior equilibria (e.g. [11, lemma 1]). Hence, simple Cross learning is fundamentally unable to learn interior equilibria and will quickly deviate from RD in cases of merely neutral stability, such as in RPS games. In contrast to RD and Cross learning, we have proved that RMD allows interior equilibria to be asymptotically stable [10], enabling the proposed MBL algorithm to overcome this fundamental limitation and approach interior Nash equilibria arbitrarily closely. Hence, with RMD admitting asymptotically stable interior equilibria, we can show that the MBL processes revisit arbitrary neighbourhoods of such equilibria infinitely often almost surely even in the case of finite step

size, particularly in zero-sum games. In contrast to more complicated algorithms, the rigorous link we prove between the stochastic MBL trajectories and the deterministic RMD allows a general analytic approach to the question of transient dynamics as well as to the question of asymptotic convergence of MBL to an ε-equilibrium in a given class of games or any particular given game *a priori*, be it zero-sum or not. In particular, this directly addresses questions of last-iterate convergence, in contrast to convergence of the time-average of iterates. Importantly, our results allow us to understand when convergence should or should not be expected, irrespective of parameter choices, by studying the properties of RMD in the setting of interest, since the behaviour of MBL follows directly due to our analysis. We demonstrate this by proving that MBL-DPU converges in zero-sum games and more generally in stable games as a direct result of the convergence of RMD in such games. To our knowledge, MBL is among the simplest uncoupled algorithms—in the sense of [6,12]—that can learn or approximate interior equilibria and among the few such for which a more general rigorous dynamic system analysis—beyond very restricted game classes—is available.

Furthermore, our analysis of MBL fully takes into account that individual game outcomes and payoffs are results of stochastic choices and does not rely on assuming the knowledge of expected outcomes. That the transition from expected outcomes to actual sampled outcomes is not trivial is demonstrated by the treatment in [13] and by the comparisons we present in the experimental settings where intuitively well-behaved algorithms clearly demonstrate the very limited validity of such intuitions. In light of the current centrality of algorithms that rely on samples or batches of samples in real-life applications, this underscores the necessity of comprehensive rigorous analyses not hinging on intuition in order to establish their reliability.

The rest of this paper proceeds as follows: After relating our results to the literature, we state the necessary evolutionary game theoretic preliminaries. We then introduce the two MBL variants, MBL-DPU and MBL-LC, which demonstrates an alternative approach to include the mutation perturbation term closer to Q-learning inspired approaches, and state the propositions on the relation of MBL-DPU to RMD and apply these to prove the convergence properties of MBL-DPU in stable games resulting from RMD. Although intuitively appealing, MBL-LC does not allow a similar treatment and its main purpose lies in charting the relation to more standard Q-learning related algorithms. We then illustrate the theoretical results with numerical experiments in a range of two-player games, as well as a three-player game, and compare the behaviours of the two MBL variants to those of frequency-adjusted Q-learning (FAQ) [14], and WoLF-PHC [6], highlighting the points where the behaviours of all but MBL-DPU start deteriorating and underscoring the utility of a mathematically rigorous link to dynamic system analysis in the study of MARL algorithms.[1]

(a) Related results

An overview over a larger class of stochastic reinforcement learning rules is provided in [16], with a focus on their relations to systems of RD type, which forms the base dynamics, incurring the difficulties mentioned earlier. As an extension, systems of RD type with additional perturbations have been related to learning rules, including such with entropy-related perturbation terms [17], and exponential learning based on a logit model [18]. Some analyses focus specifically on Q-learning based algorithms. For instance [19] considers the stability and convergence properties of Q-learning in the two-player setting; however, the Q-values enter as expectations, not as random variables, and therefore the effects of stochasticity are not fully considered—a crucial factor in a rigorous analysis, which requires attention, as demonstrated in [13]. A similar approach is pursued by FAQ in [20] with a correction given in [14]. However, beyond an intuitive elaboration no proofs on the algorithm's relation to evolutionary dynamics were provided, and therefore a rigorous analysis is missing. Nonetheless, we choose FAQ as a comparison, as [14] claims it to be linked to an ODE system similar to RMD and as it is a

[1]Portions of this manuscript are derived from one of the authors' unpublished PhD dissertation [15].

sufficiently simple uncoupled algorithm very close to Q-learning, making it a natural candidate for comparison. As a second candidate for comparison, we choose WoLF-PHC [6], since its variant WoLF-IGA is linked to a dynamic systems perspective and WoLF-PHC, too, is an uncoupled and relatively simple algorithm, close to Q-learning. Although its theoretical analysis is more thorough than that for FAQ, only the two-player two-action analysis of its WoLF-IGA variant is available. Both algorithms have demonstrated that they are able to learn Nash equilibria in simple settings under self-play, where simpler algorithms such as pure policy-hill-climbing would fail.

A separate and quite rigorous approach to MARL convergence analysis is pursued via mulitple-time-scales algorithms, where Q-value estimates are learned from payoff samples more quickly than policy changes occur [13]. Here, the convergence analysis relates to smoothed best-response dynamics. However, the time-scale separation results in a fundamentally more complicated approach and more complicated algorithms, including the additional requirement to keep track of time scales and ensure a sufficient separation. For the case of ε-greedy multi-agent Q-learning under stochastic payoffs, convergence conditions are given in [21]. However, this algorithm operates on joint actions, which requires agents to be able to observe the actions chosen by all agents, and is therefore not uncoupled in the sense of [6].

Probably closest to our approach [22–24] take RMD and our analysis in [10] as their departure point to formulate various MARL algorithms, albeit with a considerably different focus. There, the analysis is tightly bound to zero-sum games and stable games, respectively. In particular, this restricts the analysis to settings where the Nash equilibrium set is convex (e.g. [25, theorem 2.3.5]), and therefore only a single connected Nash equilibrium component exists. It is clear that this precludes any settings with multiple isolated equilibria, clearly a large class of games with generally high relevance. Importantly, our results on the relation between MBL and RMD are not limited to a comparably specific game class and our results allow a convenient transfer of results on evolutionary games for different classes of games. Furthermore, the authors assume expected payoffs with at most some very limited noise, instead of a fully stochastic approach. As mentioned earlier and as demonstrated in the experiments, this generally imposes significant limitations on the validity in sample-based settings. Furthermore, we are careful to provide a rigorously proven relation between our discrete dynamics and its continuous-time counterpart. That this transition indeed requires special attention is aptly demonstrated for the RD and its various discrete-time counterparts by [26]. Similar considerations apply to approaches pursued in [27–29], where analysis relies on either zero-sum or stable settings and expected payoffs, not taking into account the stochasticity fully. It is overall clear that approaches operating on different levels of generality will yield differing perspectives even with similar reference dynamics in mind.

We do not take into account proximal policy optimization (PPO) algorithms [30], for our comparison, since they require an agent to construct an approximation of the actual target function and solve a constrained optimization problem at each learning step with a suitable sampling strategy in-between learning and to keep track of a potentially large number of estimates. This results in a much more complicated algorithm than analysed here and convergence analysis even in the single-agent setting is challenging (e.g. [28]). We are not aware of a rigorous MARL convergence analysis in non-cooperative games, although *experimental* results in this direction exist (e.g. [31]) for n-player RPS games with convergence only in very limited cases, or [32] extending PPO to WoLF-PPO in experimental studies of matching pennies (MP) and two-player RPS.

2. Preliminaries

As our analysis of multi-agent learning is formulated in the setting of (evolutionary) game theory, we give short definitions of the main concepts employed and refer the reader to the standard literature for further details (e.g. [7,33]).

(a) Finite normal-form games

A normal-form game is a tuple (P, A, r), where $P = \{1, \ldots, N\}$ represents the set of players, $A = \times_{i \in P} A_i$ where $A_i = \{1, \ldots, n_i\}$ is the set of pure strategies of each player i,[2] and $r = (r_i)_{i \in P}$ is a family of functions with $r_i : A \to \mathbb{R}$ mapping the pure strategy profiles in A to the payoffs of player i. For each player $i \in P$, we assume that the player chooses a pure strategy from A_i according to some probability distribution x_i over A_i, i.e. according to some tuple $(x_{ih})_{h \in A_i} \in \mathcal{D}_i := \{\xi \in \mathbb{R}^{A_i}_{\geq 0} : \sum_{h \in A_i} \xi_h = 1\}$. We call such an x_i the mixed strategy of player i.[3] We call mixed strategies simply *strategies*, where there is no danger of confusion.

(b) Nash equilibrium

We call a strategy profile $x^* := (x_i^*)_{i \in P} \in \mathcal{D} := \times_{i \in P} \mathcal{D}_i$ a Nash equilibrium if for all players $i \in P$ and all mixed strategies $x_i \in \mathcal{D}_i \setminus \{x_i^*\}$, we have

$$\mathbb{E}[r_i(a)|x^*] \geq \mathbb{E}[r_i(a)|(x_i, x^*_{-i})], \tag{2.1}$$

where $(x_i, x^*_{-i}) \in \mathcal{D}$ denotes the mixed strategy profile for which $(x_i, x^*_{-i})_{ih} = x_{ih}$ $(\forall\, h \in A_i)$ and $(x_i, x^*_{-i})_{jh} = x^*_{jh}$ $(\forall\, j \in P \setminus \{i\}, h \in A_j)$. The equilibrium is called a *strict* Nash equilibrium if the inequality is strict for all $i \in P$. The well-known intuition of this concept is that no player has an incentive to deviate from the Nash equilibrium strategy given that all other players play the Nash equilibrium strategy profile, since for each player $i \in P$, x_i^* is a *best-response* to x^*. Equivalently, no pure strategy has a higher payoff than the Nash equilibrium strategy:

$$\forall\, i \in P, \quad h \in A_i : \mathbb{E}[r_i(a)|x^*] \geq \mathbb{E}[r_i(a)|x^*, a_i = h]. \tag{2.2}$$

As a useful relaxation of this concept, we call a strategy profile $(\tilde{x}_i)_{i \in P} \in \mathcal{D}$ an *ε-equilibrium* if

$$\exists \varepsilon > 0 \quad \forall\, i \in P, h \in A_i : \mathbb{E}[r_i(a)|\tilde{x}] \geq \mathbb{E}[r_i(a)|\tilde{x}, a_i = h] - \varepsilon, \tag{2.3}$$

i.e. every pure strategy is by at most ε better than $(\tilde{x}_i)_{i \in P}$, and for all players $i \in P$, $(\tilde{x}_i)_{i \in P}$ is an *ε-best-response* to $\tilde{x}$.

(c) Repeated games, learning and rationality

Given a finite normal-form game, we consider an infinitely repeated game to be a repetition of the normal-form game for each round $t \in \mathbb{N}$. In particular, assuming that in each round t the players choose a pure strategy profile $a(t) \in A$ according to the mixed strategy profile $x(t) = (x_i(t))_{i \in P}$, the pure strategy profiles constitute a stochastic process $\{a(t)\}_{t \in \mathbb{N}}$. In turn, an algorithm which adapts the mixed strategy profile in each round t, defines a potentially stochastic process $\{x(t)\}_{t \in \mathbb{N}}$. It is this resulting process and its properties which are the focus of our convergence analysis. Following the definition given by [6], we call such a process *rational*, if a player i's mixed strategy $\{x_i(t)\}_{t \in \mathbb{N}}$ converges to a best-response whenever all other players' strategies converge to a stationary policy. We call a process *ε-rational* if it converges to an ε-best-response. It is clear that in the case of stationary policies for all other players, the focal player faces a Markov decision process and the best-response strategy maximizes the player's average expected payoff. In the simplest case, where players cannot observe other players' actions and have no memory, as considered here, the usual state space and the state-dependency of policies disappear.

(d) Replicator-mutator dynamics

We consider the multi-population replicator–mutator dynamics (RMD) we formulated in [10], which is a special case of general RMD (e.g. [34]): For all $i \in P$, let $M_i > 0$ be a mutation parameter,

[2] A is usually denoted S in the game theory literature, and players are conceived as populations of pure strategies in the EGT literature. In the simplest case, pure strategies correspond to actions in the reinforcement learning literature. We use the terms 'player' and 'agent' synonymously.

[3] This would be referred to as a policy in the reinforcement learning literature.

$c_i \in \mathcal{D}_i^\circ$ (denoting the interior of $\mathcal{D}_i$) some fixed parameter and $f_i : \mathcal{D} \to \mathbb{R}^{A_i}$ a continuously differentiable fitness function. Then the RMD is given for $i \in P$, $h \in A_i$ by

$$\dot{x}_{ih}(t) = x_{ih}(t)\left(f_{ih}(x(t)) - \sum_{k\in A_i} x_{ik}(t) f_{ik}(x(t))\right) + M_i(c_{ih} - x_{ih}(t)). \tag{RMD}$$

In case that $M_i = 0$ for all $i \in P$, RMD reduces to the standard multi-population RD. One possible (and usual) conceptualization of the fitness f_{ih} of a pure strategy $h \in A_i$ is to assume that it is the expected payoff of playing h, given all other players' strategies, or more concretely, given a strategy profile $x \in \mathcal{D}$ let the fitness f_{ih} satisfy $f_{ih}(x) = \mathbb{E}[r_i(a)|x, a_i = h]$. It is clear that all fitness functions are continuously differentiable in this case.

Remark. The equilibria of RMD, also called *mutation equilibria*, in general are not Nash equilibria of the underlying game. Instead, they are ε-equilibria, where ε depends on $(M_i)_{i\in P}$ [10].

3. Mutation-bias learning

We can now introduce the stochastic learning rules and specify their relation to RMD. We provide two variants of MBL: one, based on direct policy updates (MBL-DPU, algorithm 1)—where the policy update corresponds to Cross learning [9], with a mutation-bias as a perturbation term; the other, based on logistic choice (MBL-LC, algorithm 2)—where the policy corresponds to logistic choice based on action-value estimates which are updated with a mutation-bias perturbation.

MBL with direct policy update (MBL-DPU). MBL-DPU, algorithm 1, is the simpler of the two variants with a direct policy update and no estimation of Q-values. It is an additive linear perturbation of Cross learning with perturbation term $\theta M_i(c_{ih} - x_{ih})$, line 5, and becomes identical to Cross learning [8,9], for $M_i = 0$ ($\forall\, i \in P$). In this sense it can be said to be a least complexity modification of Cross learning, since only few elementary computations are required in addition to simple Cross learning. We note that the assumption in Cross learning, that the payoffs r_i be restricted to $[0, 1]$ is not necessary. It suffices that payoffs are non-negative and bounded. In this case, θ has to be chosen small enough to ensure well-definition of MBL-DPU. Note that this assumption is not restrictive for finite games, as boundedness is trivially satisfied for finite games and non-negativity can be ensured by adding a constant C_i to all payoffs r_i, affecting neither the Nash equilibria nor the dynamics in the deterministic limit—a straightforward property of RD and RMD.

MBL with logistic choice (MBL-LC). Clearly, the simple perturbation in MBL-DPU can be combined with a wide class of transformations on the payoffs without affecting the additive character of the perturbation. A somewhat more involved possibility to combine the mutation-like perturbation with a policy update is based on a Boltzmann distribution or multinomial logistic choice, as frequently encountered in Q-learning. In MBL-LC, algorithm 2, the perturbation affects the action-value updates instead of the policy. Hence, this version resembles the perturbation term of FAQ [14,19], and allows for a closer comparison. In particular, restricting the adjustment in line 6 by applying a minimum is parallel to FAQ. Rewriting the perturbation along the lines presented as a heuristic rather than as a rigorous proof in [14,19] would suggest that MBL-LC results in RMD in the deterministic limit, which is far from clear as will become clear in the experimental section. One can see that the logistic-choice policy can still be expressed as a policy update with modified payoffs:

$$x_{ih} \leftarrow \begin{cases} x_{ih} + (1 - x_{ih})\tilde{r}_i & \text{if } h = a_i, \\ x_{ih} - x_{ih}\tilde{r}_i & \text{otherwise,} \end{cases} \qquad \text{with } \tilde{r}_i = \frac{x_{ia_i}(e^{\tau \Delta Q_{ia_i}} - 1)}{x_{ia_i}(e^{\tau \Delta Q_{ia_i}} - 1) + 1}, \tag{3.1}$$

where Q denotes an action-value function and ΔQ_{ia_i} denotes the update of the action-value of the chosen action a_i. From this it is clear that an intermediate approach could be using the simpler MBL-DPU combined with unperturbed Q-learning, which is equivalent to transforming payoffs accordingly.

Algorithm 1. (MBL-DPU) MBL with direct policy update for generic player $i \in P$.

Initialize: Choose learning rate θ, mutation parameters $M_i > 0$ and $c_i \in \mathcal{D}_i^\circ$, initial $x_i \in \mathcal{D}_i$.
for all times t **do**
  Select strategy $a_i \in A_i$ with probabilities $\Pr(a_i = h) = x_{ih}$ $(\forall\, h \in A_i)$.
  Observe payoff r_i resulting from strategy profile $(a_j)_{j\in P}$.
  For all $h \in A_i$, set: $x_{ih} \leftarrow \begin{cases} x_{ih} + \theta(1 - x_{ih})r_i + \theta M_i(c_{ih} - x_{ih}) & \text{if } h = a_i, \\ x_{ih} - \theta x_{ih} r_i + \theta M_i(c_{ih} - x_{ih}) & \text{otherwise.} \end{cases}$
end for

Algorithm 2 . (MBL-LC) MBL with logistic choice for generic player $i \in P$.

Initialize: Choose learning rate θ, $M_i > 0$ and $c_i \in \mathcal{D}_i^\circ$, $Q_i \in \mathbb{R}^{A_i}$. Choose $\beta > 0$, $\tau > 0$.
for all times t **do**
  For all $h \in A_i$, set: $x_{ih} \leftarrow \frac{e^{\tau Q_{ih}}}{\sum_{k\in A_i} e^{\tau Q_{ik}}}$.
  Select strategy $a_i \in A_i$ with probabilities $\Pr(a_i = h) = x_{ih}$ $(\forall\, h \in A_i)$.
  Observe payoff r_i resulting from strategy profile $(a_j)_{j\in P}$.
  For $h = a_i$, set: $Q_{ih} \leftarrow Q_{ih} + \min\left\{\frac{\beta}{x_{ih}}, 1\right\}\theta\left(r_i + M_i\frac{c_{ih}}{x_{ih}}\right)$.
end for

(a) Convergence of MBL-DPU

We address the question of convergence in two steps. First, we determine whether the stochastic process induced by the learning algorithm can be approximated by a deterministic dynamics. Second, we transfer the convergence properties of the deterministic dynamics to the stochastic process. For MBL-DPU we have the following convergence result (proved in appendix A):

Proposition 3.1. *For every time $T < \infty$, the family of stochastic processes $\{(X^\theta_{ih}(t))_{i,h}\}_{t\in\mathbb{N}_0}$ induced by MBL-DPU converges to RMD in the sense that for all $\varepsilon > 0$:*

$$\sup_{x_0\in\mathcal{D}} \Pr(||X^\theta(n_\theta) - \Phi(x_0, T)|| > \varepsilon) \to 0 \quad \text{as} \quad \theta \to 0, \tag{3.2}$$

where $n_\theta\theta \to T$ for $\theta \to 0$, x_0 is a.s. the initial state of the stochastic processes, i.e. $X^\theta(0) = x_0$ a.s., and $\Phi(x_0, \cdot)$ is the unique solution of RMD with $\Phi(x_0, 0) = x_0$.

Remark. As discussed in [8,35], proposition 3.1 on its own does not yield an analysis of the asymptotic behaviour of the stochastic process. However, if a mutation equilibrium x^M of RMD is asymptotically stable and x_0 lies in the basin of attraction of x^M, then we have $\Phi(x_0, T) \to x^M$ as $T \to \infty$. Hence, with the asymptotic stability of x^M, we have that for T large enough, $\Phi(x_0, T)$ is arbitrarily close to x^M and together with proposition 3.1, any neighbourhood of x^M will be reached by the learning process $\{X^\theta(t)\}_{t\geq 0}$ with an arbitrary degree of certainty after finitely many steps for suitable choice of θ. Although this does not imply that the process must remain in this neighbourhood afterwards, it will revisit the neighbourhood with arbitrary probability depending on θ.

Attracting mutation limits. In [10] we showed that every game has at least one connected Nash equilibrium component that is approximated by mutation equilibria irrespective of the choice of the mutation parameter c, as $M \to 0$, called a *mutation limit*. Furthermore, it was shown that for the game of MP the Nash equilibrium is approximated by asymptotically stable mutation equilibria, warranting the name *attracting mutation limit* for such Nash equilibria. This implies the following consequence (proved in appendix A):

Proposition 3.2. *If a unique Nash equilibrium $x^* \in \mathcal{D}^\circ$ is an attracting mutation limit and U a neighbourhood of x^*, then for every mutation parameter $c \in \mathcal{D}^\circ$ there are $M > 0$, $\theta > 0$ such that the*

stochastic process $\{(X^\theta(t))\}_{t\in\mathbb{N}_0}$ *induced by MBL-DPU visits* U *at a finite time a.s., i.e. with probability* 1 *there is* $S \in \mathbb{N}_0$ *with* $X^\theta(S) \in U$. *In fact,* $\{(X^\theta(t))\}_{t\in\mathbb{N}_0}$ *a.s. visits* U *infinitely often.*

With the relation between MBL-DPU and RMD spelled out clearly, analysing the behaviour of MBL-DPU in various game classes becomes a matter of inspecting RMD. This convenience manifests itself in the clarity with which the proof of the following proposition on the convergence of MBL-DPU in stable games can be formulated (see appendix A):

Proposition 3.3. *Let* $f \in C^1(\mathcal{D}, \mathbb{R}^{A_1\times\ldots\times A_n})$ *be a continuously differentiable fitness function, such that* f *is a stable game in the sense of [36, definition 3.3.1], i.e.:*

$$\forall\, x, y \in \mathcal{D} : (y - x)^T(f(y) - f(x)) \leq 0. \tag{3.3}$$

Then the Nash equilibrium for f *is an attracting mutation limit and, for every open neighbourhood* U *of the Nash equilibrium, there are* $c \in \mathcal{D}^\circ$, $M > 0$ *and* $\theta > 0$ *such that the stochastic process* $\{X^\theta(n)\}_{n\geq 0}$ *induced by MBL-DPU visits* U *infinitely often almost surely.*

With zero-sum games being a subclass of stable games, this implies that MBL-DPU approximates Nash equilibria in zero-sum games to any desired precision. In contrast to MBL-DPU, we do not have a proof of an analogous result for MBL-LC, yet. In [14,19] it is assumed that FAQ, a similar logistic-choice learning rule based on Q-learning, converges to a perturbation of the RD, although no proof is given. Although it seems plausible for MBL-LC to behave similarly to MBL-DPU, the experimental results indicate that MBL-LC is probably more sensitive to the choice of learning rate than MBL-DPU, since the logistic choice can cause a stronger variance of the strategy at each learning step, as indicated in the more detailed results for MBL-LC in the electronic supplementary material, S. The larger variance in the learning step is also the reason why our proof strategy for MBL-DPU cannot be translated to MBL-LC in a trivial manner.

Perturbation creates a trade-off between accuracy and speed. We note that neither MBL-DPU nor MBL-LC converge to a Nash equilibrium but only to an ε-equilibrium and in particular, that both stay away from the boundary of $\mathcal{D}$. For MBL-DPU this is clear from the fact that the equilibria of RMD are not Nash equilibria and that the boundary of $\mathcal{D}$ is repelling. For MBL-LC this is also due to the exploration parameter τ. For the latter, it is further the case that τ cannot be allowed to approach ∞ as this collides with the $\theta \to 0$ limit and makes the time derivative of the policy unbounded. This results in a highly increased variance in the stochastic process, preventing effective learning of equilibria. This particular aspect applies also to other logistic-choice based algorithms, particularly FAQ. However, if MBL-LC and FAQ indeed converge to the corresponding ODE systems, then these include τ as a simple scaling parameter. Since constant positive rescalings do not change the trajectories, the systems can be rescaled by $1/\tau$ in such a way that τ effectively regulates the perturbation's strength relative to the RD. In the case of RMD, $1/\tau$ can be absorbed by the mutation strength M. Thus an increase of τ has the same effect as a decrease of M which results in all mutation equilibria moving closer to a Nash equilibrium, as desired. A reduction in the perturbation strength also results in a longer time to approach equilibria and this creates a trade-off between accuracy and speed for both MBL-LC and MBL-DPU.

4. Experimental results

We illustrate the theoretical results and in particular some cases where the importance of rigorous analysis becomes clear and the intuition about apparently well-behaved algorithms starts to become unreliable in a number of experimental settings: The prisoner's dilemma (PD), one of the most studied games in game theory, is the clearest and most straightforward case in terms of learning dynamics, with a unique and strict Nash equilibrium. The second game, MP, has been used to model antagonistic interactions (e.g. host-parasite and pursuit-evasion settings [37,38]), and is an example for an asymmetric game. In terms of complexity, similar to PD, it has a

unique equilibrium, which lies in the interior of the joint strategy space and is not asymptotically stable under the RD without mutation, presenting a more challenging case for learning dynamics than PD. For both, PD and MP, the learning dynamics reduce to two-dimensional systems and with the Poincaré–Bendixson theorem, the complexity of the limit sets and therefore the asymptotic behaviour of any continuous learning dynamics is quite restricted (e.g. [39, theorem 7.16]). However, it is known that chaotic behaviour can arise already in two-player RPS games, effectively a four-dimensional system (e.g. [3,17]). Therefore, we compare the algorithms on RPS with 3, 5, and 9 strategies, as cases where the Nash equilibrium is still unique and Lyapunov stable under the RD without mutation, such that the main difference to MP is the higher dimensionality, permitting for potentially chaotic behaviours to arise, yet not presenting a structurally more challenging case than MP. As an example setting for understanding when and why convergence might not be achieved, we also consider the three-player MP (3MP) game. We compare MBL-DPU and MBL-LC to FAQ [14], and WoLF-PHC [6]. For details on the games' payoffs and further experiments, cf. electronic supplementary material, S.

(a) Prisoner's dilemma

PD is an example of a game with a strict Nash equilibrium at a vertex of the joint strategy space $\mathcal{D}$. It is known that strict Nash equilibria are asymptotically stable under RD (e.g. [7]). In this case, plain Cross learning would also converge to the Nash equilibrium. We had shown previously that RMD does not destabilize asymptotically stable equilibria of RD [10, lemma 4.8]. Hence, the mutation equilibrium resulting from the mutation perturbation remains asymptotically stable and, with our result, MBL-DPU also learns an approximation of the Nash equilibrium. In this sense, PD is the least challenging setting in terms of the ease with which the Nash equilibrium can be learned. The setting serves mainly to illustrate the fact that the learned equilibria of MBL-DPU and MBL-LC in fact lie away from the boundary Nash equilibrium, in particular since mutation pushes the trajectories away from the boundary of $\mathcal{D}$, in contrast to the other two algorithms. With decreasing mutation strength M, both algorithms are able to better approach the Nash equilibrium, as would be expected from RMD. This case also illustrates that the more elementary MBL-DPU converges more slowly than either of MBL-LC, FAQ, or WoLF-PHC. For more details and figures on this benign case, we refer the reader to electronic supplementary material, appendix S(a).

(b) Zero-sum games—matching pennies (MP)

As a second, structurally different case, we consider zero-sum or equivalent games which have interior Nash equilibria. As an example, we consider MP with player 1's payoffs given by $[R_1]_{h,k}$ and player 2's payoffs given by $[R_2]_{h,k}$ where h is player 1's strategy and k is player 2's strategy, specifically: $[R_1]_{1,1} = 1$, $[R_1]_{1,2} = -2.3$, $[R_1]_{2,1} = -0.4$, $[R_1]_{2,2} = 1$; $[R_2]_{1,1} = -2.3$, $[R_2]_{1,2} = 1$, $[R_2]_{2,1} = 1$, $[R_2]_{2,2} = -0.4$; the Nash equilibrium x^* lying at $x_1^* = (14/47, 33/47)$, $x_2^* = (33/47, 14/47)$. For this game it is straightforward to check that the eigenvalues of the Jacobian of RMD in the neighbourhood of the Nash equilibrium have only negative real parts. Equivalently, one can check that the eigenvalues of the Jacobian of RD are purely imaginary in the neighbourhood of the Nash equilibrium and consider that RMD shifts the eigenvalues towards the negative half-plane, rendering the Nash equilibrium an attracting mutation limit. With propositions 3.1, 3.2 and 3.3, MBL-DPU is of course already guaranteed to converge in these cases. In fact, we observe convergence in the MP setting for MBL-DPU, MBL-LC, as well as our comparisons, FAQ and WoLF-PHC, figure 1. This setting illustrates that MBL-DPU overcomes the limitations of Cross learning at a minimal cost in increased complexity. Similar to the PD setting, MBL-DPU converges more slowly than the more complicated algorithms: In our simulations, MBL-LC, FAQ and WoLF-PHC were factors of approximately 2.0, 1.45 and 1.2, respectively, faster than MBL-DPU, which takes approximately 5×10^5 steps on MP, when measuring the number of

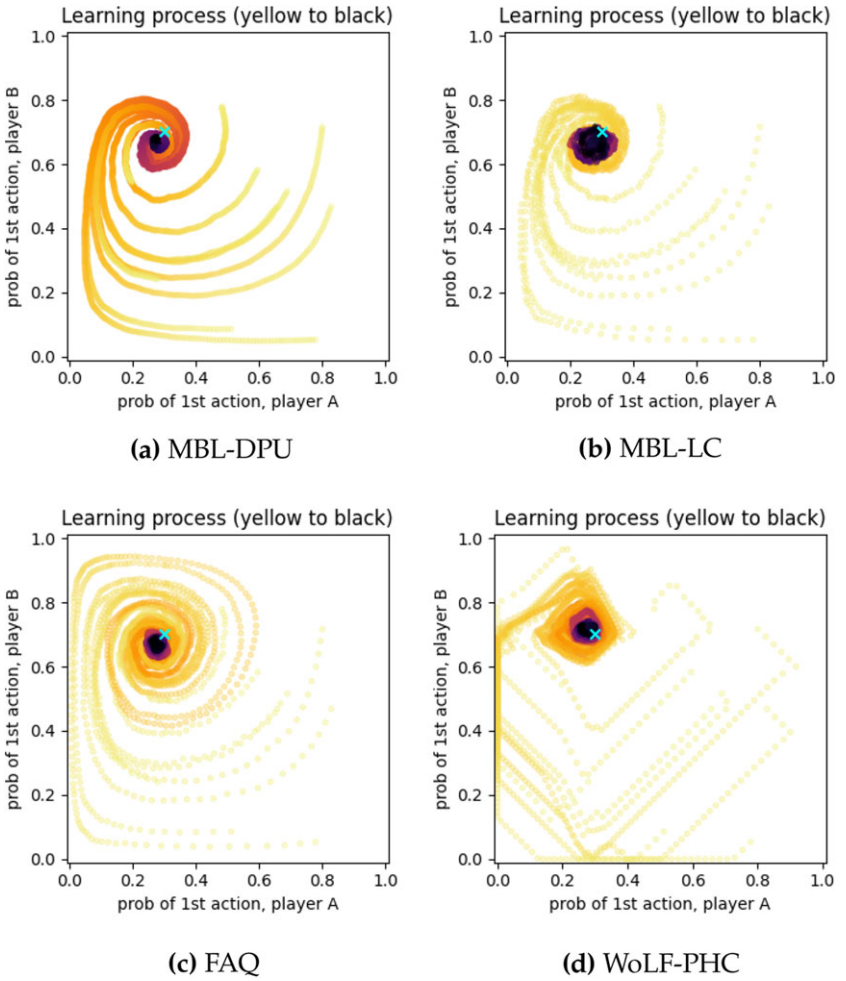
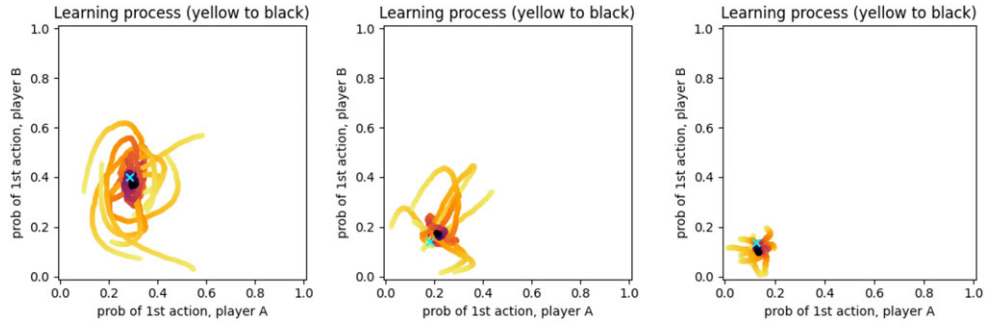


Figure 1. Self-play on the MP game; for 10 different initial conditions. Each subfigure shows the ten trajectories in the projection onto the first components of the players' strategies, in this case the 'defect' strategy, with the first player on the horizontal axis and the second on the vertical axis. Points coloured yellow correspond to earlier points in time, changing over orange and violet to black for later points in time. The position of the game's Nash equilibrium is marked with a blue cross in the projection plane. (Parameter values for MBL-DPU, MBL-LC and FAQ: $M^{-1} = \tau = 20$; for WoLF-PHC with initial learning rate 10^{-1} for Q, win learning rate $1/2 \times 10^{-4}$.)

steps required such that the system state is at most at twice the distance of the residual distance, which remains non-zero owing to the system's stochasticity.

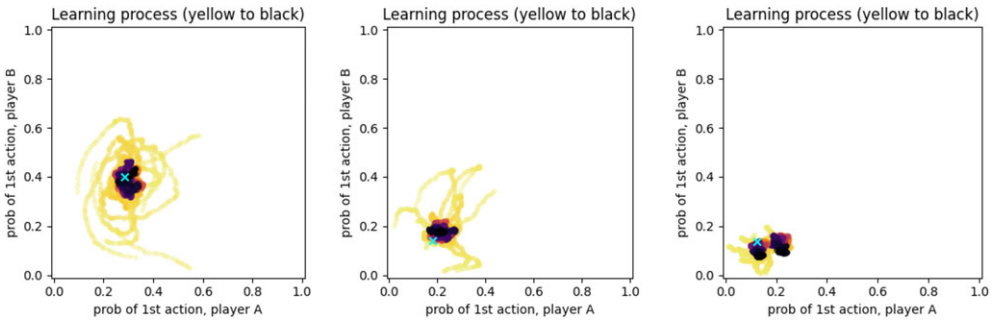
(c) Zero-sum games—RPS

For the higher-dimensional settings, i.e. RPS with 3, 5 and 9 strategies, we still observe convergence for MBL-DPU, figure 2, as guaranteed by the Nash equilibrium being an attracting mutation limit. Naturally, the trajectories of the resulting 4, 8 and 16 dimensional systems appear less intuitive in the two-dimensional projection. For MBL-LC, figure 3, and FAQ, figure 4, we observe convergence in the RPS-3 case, but both algorithms deteriorate in higher dimensions, MBL-LC for RPS-9, figure 3c, and FAQ for RPS-5 and RPS-9, figure 4b,c, with both showing the convergence region splitting up such that some trajectories stop approximating the Nash equilibrium. Similarly, while WoLF-PHC seems to approach the Nash equilibrium in RPS-3 and RPS-5, figure 5, it loses the ability to learn the Nash equilibrium for RPS-9, figure 5c, with trajectories seemingly getting stuck near the boundary of $\mathcal{D}$. In RPS-3, all four algorithms require



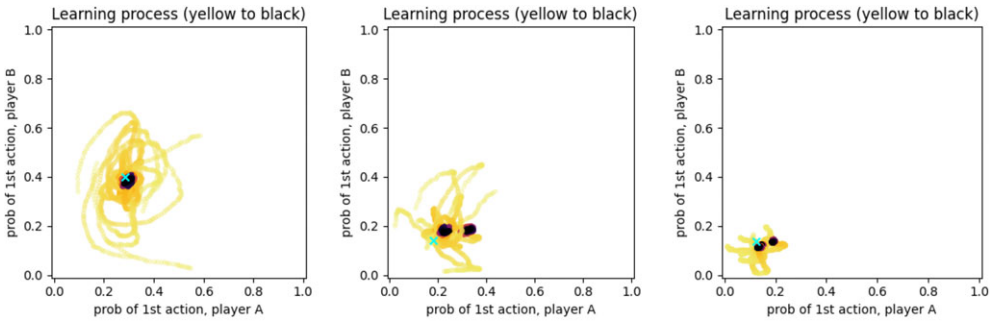
(a) MBL-DPU on RPS-3. **(b)** MBL-DPU on RPS-5. **(c)** MBL-DPU on RPS-9.

Figure 2. Self-play of MBL-DPU on RPS-3, RPS-5 and RPS-9 games, with $M^{-1} = 20$, with payoffs for RPS-3 given by: $[R_1]_{1,1} = 0$, $[R_1]_{1,2} = -2$, $[R_1]_{1,3} = 3$, $[R_1]_{2,1} = 2$, $[R_1]_{2,2} = 0$, $[R_1]_{2,3} = -2$, $[R_1]_{3,1} = -1$, $[R_1]_{3,2} = 2$, $[R_1]_{3,3} = 0$, $R_2 = -R_1$. For payoffs for RPS-5 and RPS-9, see electronic supplementary material, S(b).



(a) MBL-LC on RPS-3. **(b)** MBL-LC on RPS-5. **(c)** MBL-LC on RPS-9.

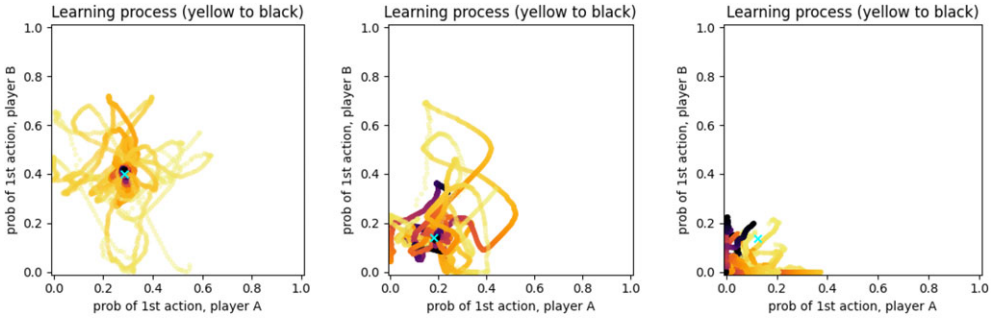
Figure 3. Self-play of MBL-LC on RPS-3, RPS-5 and RPS-9 games, with $M^{-1} = \tau = 20$.



(a) FAQ on RPS-3. **(b)** FAQ on RPS-5. **(c)** FAQ on RPS-9.

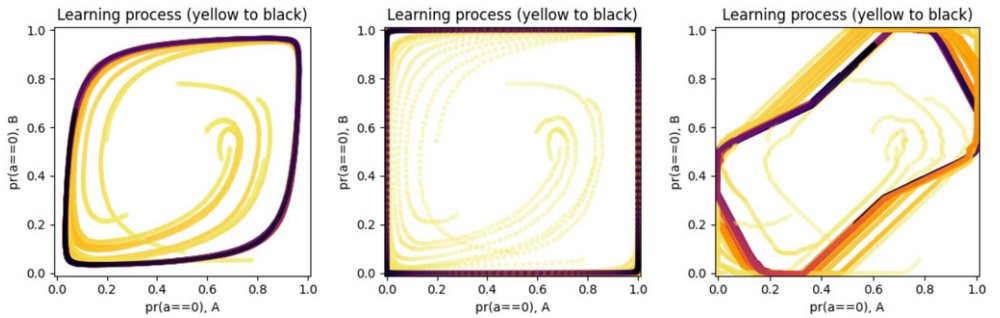
Figure 4. Self-play of FAQ on RPS-3, RPS-5 and RPS-9 games, with $\tau = 20$.

a comparable number of approximately 7×10^5 steps until they remain within at most twice the residual distance from their respective long-term means. In RPS-5 and RPS-9, comparisons become meaningless owing to the algorithms not converging to a single mean on all runs. MBL-DPU requires approximately 7.5×10^5 steps on RPS-5 and approximately 9×10^5 steps on RPS-9.



(a) WoLF-PHC on RPS-3. **(b)** WoLF-PHC on RPS-5. **(c)** WoLF-PHC on RPS-9.

Figure 5. Self-play of WoLF-PHC-learning on RPS-3, RPS-5 and RPS-9 games, with initial learning rate 10^{-1} for Q, win learning rate $1/2 \times 10^{-4}$.



(a) MBL-DPU. **(b)** FAQ. **(c)** WoLF-PHC.

Figure 6. Self-play on 3MP by (a) MBL-DPU with $M^{-1} = 20$, (b) FAQ with $\tau = 20$ and (c) WoLF-PHC with initial learning rate 10^{-1} for Q, win learning rate $1/2 \times 10^{-4}$.

(d) Three-player matching pennies (3MP)

Beyond the two-player case, we compare MBL in a 3MP setting introduced in [40]. In short, the three players have a shared pure strategy space, i.e. $A_1 = A_2 = A_3$, with two pure strategies, where player 1 wants to match player 2, player 2 wants to match player 3, and player 3 wants not to match player 1. The unique Nash equilibrium lies at the centre of $\mathcal{D}$. All four algorithms fail to learn the Nash equilibrium, figure 6 (MBL-LC not shown, cf. appendix S(c)). Instead, they seem to approach a seemingly stable periodic orbit.

5. Discussion

The experimental results illustrate the difficulties in relying on intuition and experimental results alone. WoLF-PHC, FAQ and MBL-LC all show quicker convergence in those cases where they actually do converge and they would seem the better choice than MBL-DPU. Not surprisingly, this is the case in PD, which has a strict Nash equilibrium, and in MP which is a planar system and cannot exhibit too complex behaviours. However, we see that behaviours start becoming less clear when we move to higher dimensions in the RPS variants. While all algorithms seem to approximate the Nash equilibrium in RPS-3, we see unexpected behaviour in RPS-5 for FAQ with a split up convergence region. In RPS-9 we see FAQ deteriorate further and MBL-LC now also fails to converge with a split in the convergence regions. WoLF-PHC now too fails to learn the Nash equilibrium, with trajectories stalling or getting stuck near the boundary. In RPS-9 no algorithm except for MBL-DPU—the simplest among the four, and the only one with a convergence proof

available—manages to reliably approach the Nash equilibrium. This loss of convergence for the more complex algorithms is unexpected, since RPS-9 does not fundamentally differ from RPS-3 in the game structure and the failure to learn when moving from RPS-3 to RPS-9 would be hard to anticipate *a priori*. In contrast, with the theoretical results on MBL-DPU we have an indication of how well it will generalize to a structurally comparable but higher-dimensional scenario.

The failure of FAQ, WoLF-PHC and MBL-LC in RPS-9 does not imply that there are no parameter choices that could potentially restore the convergence of the respective algorithms. For example tweaking the learning rates might restore convergence in these specific cases, without guaranteeing convergence in higher-dimensional scenarios. However, the absence of analytical tools leaves the existence of such parameter values an open question. Even where such parameter choices exist the problem remains potentially intractable without an indication of where to look for them in the parameter space—even more so for algorithms with more parameters. Together with the unpredictability of failure to converge when moving from a low- to a higher-dimensional setting, this questions the reliability of algorithms that seem to make sense intuitively and look promising in some experiments but for which we lack fundamental results—particularly for even more complicated algorithms not considered here. In this situation, the utility of the mathematical guarantees available for MBL-DPU becomes obvious. Given a payoff structure, conditions for convergence can be checked by analysing the corresponding ODE system. In specific cases, this allows a very straightforward analysis of classes of settings, such as we have provided for stable games in proposition 3.3 by showing that RMD stabilizes equilibria and therefore allows MBL-DPU to converge to neighbourhoods of the Nash equilibrium. We further understand where exactly MBL-DPU is headed and that empirical non-convergence becomes less likely with smaller learning rates. This gives an indication of where to look for a suitable learning rate. More importantly, our results allow an approach that is not fundamentally restricted to any particular game class as long as RMD can be analysed. Finally, where MBL-DPU fails to converge, as in 3MP, just as the other algorithms, the underpinning ODE makes this expectable and understandable, since an analysis of the corresponding RMD system quickly shows that the Jacobian of the system has eigenvalues with positive real parts at the Nash equilibrium, making the equilibrium unstable for sufficiently small mutation strengths. Overall, the result on the general connection between MBL-DPU and RMD allows one to further deduce—without requiring separate proofs—that MBL-DPU will converge wherever RD converges, since RMD converges wherever RD does, as clarified in [10], which e.g. includes potential games [36, theorem 7.1.6]. This demonstrates that such theoretical results enable us to understand when a given algorithm is not the best choice for a setting, instead of searching for parameter values that might or might not restore convergence, as we would be forced to do otherwise.

Note that we have left out any modifications to further improve MBL-DPU with the purpose of analysing the least complex variant with few parameters. In particular, as we had clarified in [10], the choice of the mutation parameters c and sufficiently small M does not qualitatively affect the behaviour of RMD and hence MBL-DPU and their relation to RD. This approach can be relaxed by varying either c or M, as we had also mentioned in [15]. The theoretical perspective makes it quite plausible that mutation strength can be chosen according to a reduction schedule, starting with high mutation and fast convergence and reducing mutation over time, increasing the accuracy with which the Nash equilibrium is approximated. Note further that the mutation strength is linked to a measure of the Nash condition not being satisfied, since the equilibria of RMD are ε-equilibria. Hence, every player can use the current violation of the Nash condition, i.e. its own distance from a current best-response, as a guide to adjust its mutation strength, e.g. by adjusting the mutation strength to be slightly lower than the current violation of the Nash condition. We conjecture that this would result in the system being driven towards a state that is not worse than the current state, as measured by the Nash condition, while keeping the convergence speed as high as possible. We would expect this to speed up convergence and improve the speed-accuracy trade-off, making MBL-DPU more attractive as a simple, predictable and theoretically founded MARL algorithm. Apart from such practical considerations, the current analysis still leaves open the questions of analysing MBL-DPU's behaviour in non-zero-sum games without

strict Nash equilibria and its behaviour in a wider range of n-player settings with more than two players. Note that some generalizations will be covered by the presented results, such as other sources of stochasticity which do not affect the ability to compute expectations and leave variances bounded. Other generalizations, such as to sequential games, might require potentially non-trivial extensions, which might build on the presented results as starting points. In addition, a clarification of the convergence properties of MBL-LC would allow us to determine whether a smaller learning rate would recover convergence, since the logistic-choice policy shows much larger variance than the direct policy update and might thus be more sensitive to the learning rate.

Data accessibility. The code that has been used to run the numerical simulations is made available as electronic supplementary material [41]. However, no simulations are required to verify the mathematical claims central to the paper, and all simulations are mainly providing illustration of the theoretical points made.

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. J.B.: conceptualization, formal analysis, investigation, methodology, writing—original draft, writing—review and editing; S.W.: conceptualization, formal analysis, investigation, methodology, writing—original draft, writing—review and editing; E.A.: conceptualization, formal analysis, investigation, methodology, writing—original draft, writing—review and editing; M.B.: conceptualization, formal analysis, investigation, methodology, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 690817, as part of the Research and Innovation Staff Exchange (RISE) programme.

Appendix A. Proofs

The proofs employ a result proved in [35, p. 118], which we state in the following appendix A(a), and then proceed to prove propositions 3.1 and 3.2 in appendix A(b) and proposition 3.2 in appendix A(c).

(a) A theorem on learning with small steps

The result from [35] we employ is phrased in the following terms: Let $J \subset \mathbb{R}_{>0}$ be a parameter set with $\inf J = 0$ and $N \in \mathbb{N}$, such that for every $\theta \in J$, $\{X_n^\theta\}_{n\geq 0} \subset I_\theta \subset \mathbb{R}^N$ is a Markov process with stationary probabilities. We denote by $\mathbb{E}_x[X_n^\theta]$ the expected value of X_n^θ given $X_0^\theta = x$. Let further I be the minimal closed convex set with $\bigcup_\theta I_\theta \subset I$. Define

$$H_n^\theta = \frac{\Delta X_n^\theta}{\theta},$$

and let $w(x,\theta)$, $S(x,\theta)$, $s(x,\theta)$ and $r(x,\theta)$ for $(x,\theta) \in I \times J$ be given as:

$$w(x,\theta) = \mathbb{E}[H_n^\theta | X_n^\theta = x] \in \mathbb{R}^N,$$

$$S(x,\theta) = \mathbb{E}[(H_n^\theta)^2 | X_n^\theta = x] \in \mathbb{R}^{N\times N},$$

$$s(x,\theta) = \mathbb{E}[(H_n^\theta - w(x,\theta))^2 | X_n^\theta = x] = S(x,\theta) - w(x,\theta)^2 \in \mathbb{R}^{N\times N}$$

and

$$r(x,\theta) = \mathbb{E}[||H_n^\theta||^3 | X_n^\theta = x] \in \mathbb{R},$$

where $x^2 = xx^T$ and $||x|| = \sqrt{x^T x}$ for $x \in \mathbb{R}^N$.

We can now state theorem 8.1.1 from [35, p. 118] (omitting part (C)):

Theorem A.1 (Norman). *In the above situation, let the following conditions be satisfied: The family of sets $(I_\theta)_\theta$ satisfies*

$$\forall\, x \in I : \lim_{\theta \to 0} \inf_{y \in I_\theta} ||x - y|| = 0. \tag{A 1}$$

There are functions w and s on I such that

$$\sup_{x\in I_\theta} ||w(x,\theta) - w(x)|| \in \mathcal{O}(\theta) \tag{A 2}$$

and

$$\sup_{x\in I_\theta} ||s(x,\theta) - s(x)|| \to 0 \quad \text{for } \theta \to 0, \tag{A 3}$$

where $\mathcal{O}$ refers to the Bachmann–Landau notation.

The function w is differentiable, i.e. there is a function w' such that for all $x \in I$:

$$\lim_{\substack{y\to x \\ y\in I}} \frac{||w(y) - w(x) - w'(x)(y-x)||}{||y-x||} = 0. \tag{A 4}$$

The function w' is bounded:

$$\sup_{x\in I} ||w'(x)|| < \infty. \tag{A 5}$$

The functions w' and s satisfy the Lipschitz condition:

$$\sup_{x,y\in I, x\neq y} \frac{||w'(x) - w'(y)||}{||x-y||} < \infty \tag{A 6}$$

and

$$\sup_{x,y\in I, x\neq y} \frac{||s(x) - s(y)||}{||x-y||} < \infty. \tag{A 7}$$

The function r is bounded:

$$\sup_{\theta\in J, x\in I_\theta} r(x,\theta) < \infty. \tag{A 8}$$

Let further for $\theta \in J$ and $x \in I_\theta$, $\mu_n(x,\theta) = \mathbb{E}_x[X_n^\theta]$ and $\omega_n(x,\theta) = \mathbb{E}_x[||X_n^\theta - \mu_n(x,\theta)||^2]$.

In this case, the following hold:

(i) $\omega_n(x,\theta) \in \mathcal{O}(\theta)$ uniformly in $x \in I_\theta$ and $n\theta \le T$ for any $T < \infty$;

(ii) For any $x \in I$, the differential equation

$$f'(t) = w(f(t))$$

has a unique solution $f(t) = f(x,t)$ with $f(0) = x$. For all $t \ge 0$, we have $f(t) \in I$, and

$$\mu_n(x,\theta) - f(x,n\theta) \in \mathcal{O}(\theta)$$

uniformly in $x \in I_\theta$ and $n\theta \le T$.

Remark A.2. We note that parts (i) and (ii) imply that for all $\varepsilon > 0$,

$$\sup_{x\in I_\theta} \Pr(||X_n^\theta - f(x,T)|| > \varepsilon) \to 0,$$

for $n\theta \to T$, $\theta \to 0$, and given that $X_0^\theta = x$ almost certainly for all θ.

(b) Convergence of MBL-DPU

We restate the simple reinforcement-mutation rule of MBL-DPU in the setting layed out above, denoting the mixed strategies with an upper-case X to underscore that this is a random variable and denoting the dependence on a parameter θ, denoting the whole family of stochastic processes as $\{(X_{ih}^\theta(n))_{i\in P, h\in A_i}\}_{n\ge 0}$. Let $R(x) = (R_{ih}(x))_{i\in P, h\in A_i}$ be a random variable whose probability distribution depends smoothly on $x \in I$ with a discrete, non-negative support which

is independent of x, and let $M_i < \overline{M}$ for some upper bound $\overline{M} < \infty$ and all $i \in P$. For a player $i \in P$ and a chosen pure strategy $h \in A_i$, the update rule then is given as follows:

$$\left.\begin{aligned} X^\theta_{ih}(n+1) &= X^\theta_{ih}(n) + \theta((1 - X^\theta_{ih}(n))R_{ih}(X^\theta(n))) + \theta M_i(c_{ih} - X^\theta_{ih}(n)) \\ \text{and} \qquad X^\theta_{ik}(n+1) &= X^\theta_{ik}(n) + \theta((-X^\theta_{ik}(n))R_{ih}(X^\theta(n))) + \theta M_i(c_{ik} - X^\theta_{ik}(n)), \quad \text{for } k \neq h. \end{aligned}\right\} \tag{A 9}$$

We can now show proposition 3.1, i.e. that this rule indeed approximates RMD for $\theta \to 0$ in the sense of remark A.2:

Proposition A.3. *There is J such that the family of stochastic processes $\{(X^\theta_{ih}(n))_{i\in P, h\in A_i}\}_{n\geq 0}$ given by equation (A 9) approximates the RMD for $\theta \to 0$ in the sense of remark A.2 if $X^\theta(0) \in I$ for all $\theta \in J$.*

Proof. The proof proceeds by showing that $\{(X^\theta_{ih}(n))_{i\in P, h\in A_i}\}_{n\geq 0}$ satisfies the conditions of theorem A.1. For a player $i \in P$ and a chosen strategy $h \in A_i$ we have

$$H^\theta_{ih}(n+1) = \frac{\Delta X^\theta_{ih}(n+1)}{\theta} = (1 - X^\theta_{ih}(n))R_{ih}(X^\theta(n)) + M_i(c_{ih} - X^\theta_{ih}(n))$$

and

$$H^\theta_{ik}(n+1) = \frac{\Delta X^\theta_{ik}(n+1)}{\theta} = -X^\theta_{ik}(n)R_{ih}(X^\theta(n)) + M_i(c_{ik} - X^\theta_{ik}(n)), \quad \text{for } k \neq h.$$

Note that in this case, $H^\theta_{ih}(n+1)$ is independent of θ if $X^\theta(n)$ is given, which simplifies the analysis. Let us set $f_{ih}(x) = \mathbb{E}[R_{ih}(X^\theta(n))|X^\theta(n) = x]$, where it is clear that there is no dependence on n. Note that f is smooth, being a composition of smooth functions.

Condition (A 1): In our case, I is given as the polyhedron $\times_i \mathcal{D}_i$ and $I_\theta = I$ for all θ and thus condition (A 1) is satisfied. It remains to show that indeed $\{(X^\theta_{ih}(n))_{i\in P, h\in A_i}\}_{n\geq 0} \subset I$: Note that R_{ih} is a discrete non-negative random variable and thus bounded by some $C < \infty$. For $\theta < (C + \overline{M})^{-1}$, we have $\theta M_i \leq 1$. Assume that $X^\theta_{ih}(n) = x \in I$, then for a player $i \in P$ and a chosen strategy $h \in A_i$ we have

$$\begin{aligned} X^\theta_{ih}(n+1) &= x_{ih} + \theta((1 - x_{ih})R_{ih}(n+1) + M_i(c_{ih} - x_{ih})) \\ &= x_{ih}(1 - \theta M_i) + \theta(1 - x_{ih})R_{ih}(n+1) + \theta M_i c_{ih} \geq 0 \end{aligned}$$

and for some other pure strategy $k \neq h$, we have

$$\begin{aligned} X^\theta_{ik}(n+1) &= x_{ik} + \theta((-x_{ik})R_{ih}(n+1) + M_i(c_{ik} - x_{ik})) \\ &= x_{ik}(1 - \underbrace{\theta(R_{ih}(n+1) + M_i)}_{\leq 1}) + \theta M_i c_{ik} \geq 0. \end{aligned}$$

A simple calculation shows that $\sum_k X^\theta_{ik}(n+1) = 1$ if $x \in I$. Thus $\{(X^\theta_{ih}(n))_{i\in P, h\in A_i}\}_{n\geq 0} \subset I$ if $X^\theta(0) \in I$ for all θ and we can choose $J = (0, (C + \overline{M})^{-1})$.

Conditions (A 1) and (A 3): Consider first the function w:

$$\begin{aligned} w_{ih}(x,\theta) &= \mathbb{E}[H^\theta(n)|X^\theta(n) = x] \\ &= x_{ih}(1 - x_{ih})\mathbb{E}[R_{ih}(n+1)|X^\theta(n) = x] + x_{ih}M_i(c_{ih} - x_{ih}) \\ &\quad + \sum_{k\neq h} x_{ik}(-x_{ih})\mathbb{E}[R_{ik}(n+1)|X^\theta(n) = x] + x_{ik}M_i(c_{ih} - x_{ih}) \\ &= x_{ih}\left(f_{ih}(x) - \sum_k x_{ik}f_{ik}(x)\right) + M_i(c_{ih} - x_{ih}). \end{aligned}$$

It is clear that w does not depend on θ and that condition (A 2) is trivially satisfied. Similarly, $S(x,\theta)$ and $s(x,\theta)$ do not depend on θ and condition (A 3) is trivially satisfied.

Conditions (A 4)–(A 7): Since the function f is smooth, so is w. In particular, we have that $\sup_{x\in I} ||w'(x)|| < \infty$ because I is compact and w' is continuously differentiable, from which follows that w' satisfies the Lipschitz condition (A 6) on I. Similarly, s is smooth and satisfies condition (A 7).

Condition (A 8): Again, r does not depend on θ, and is smooth on I, which is compact. Thus it is bounded on I and condition (A 8) is satisfied.

As a consequence, we can apply theorem A.1 to the family $\{X^\theta(n)\}_{n\geq 0}$ and with remark A.2 we have that for all $\varepsilon > 0$,

$$\sup_{x\in I} \Pr(||X^\theta(n) - \Phi(x,T)|| > \varepsilon) \to 0,$$

for $n\theta \to T$, $\theta \to 0$, and given that $X^\theta(0) = x$ for all θ, where for all $i \in P$ and $h \in A_i$, Φ is the unique solution of the differential equations

$$\begin{aligned}\dot{\Phi}_{ih}(x,t) &= w_{ih}(\Phi(x,t)) \\ &= \Phi_{ih}(t)\left(f_{ih}(\Phi(x,t)) - \sum_k \Phi_{ik}(x,t) f_{ik}(\Phi(x,t))\right) + M_i(c_{ih} - \Phi_{ih}(x,t)),\end{aligned}$$

with $\Phi(x,0) = x$. ■

Proposition A.4. *Let x^M be an equilibrium of RMD and U an open neighbourhood of x^M. If x^M is globally asymptotically stable, then there is $\theta > 0$ such that the stochastic process $\{(X^\theta_{ih}(n))_{i\in P, h\in A_i}\}_{n\geq 0}$ defined in condition (A 9) visits U almost surely after finitely many steps.*

Proof. Let $\Phi(x,\cdot) : \mathbb{R}_{\geq 0} \to \mathcal{D}$ satisfy RMD with $\Phi(x,0) = x$ for all $x \in \mathcal{D}$. Let further $U' \subset U$ such that $x^M \in U'$ and $\bigcup_{x\in U'} B_\delta(x) \subset U$ for some $\delta > 0$, where $B_\delta(x)$ denotes an open ball with radius δ around x. As x^M is globally asymptotically stable, there is for each $x \in \mathcal{D}$ a $t' < \infty$ such that for all $t > t'$: $\Phi(x,t) \in U'$.

This is because there is a neighbourhood $V \subset U'$ of x^M such that $\forall\, x^0 \in V, t > 0 : \Phi(x^0,t) \in U'$ due to the Lyapunov stability of x^M. Since x^M is asymptotically stable, for every x there is a $t > 0$ such that $\Phi(x,t) \in V$ and hence the solution will remain in U' afterwards. Therefore, define

$$\tau : \mathcal{D} \to \mathbb{R} \quad \text{with } \tau(x) = \inf\{T > 0 : \Phi(x,T) \in V\}.$$

Since the r.h.s. of RMD is continuously differentiable by assumption, it is also Lipschitz continuous. Thus, Φ is continuous in the first argument and so is τ as the following argument shows:

Let $x \in \mathcal{D}$ and $\varepsilon_1 > 0$. Then there is $t > \tau(x)$ such that $\Phi(x,s) \in V$ for $s \in (\tau(x), t]$. Choose $s \in (\tau(x), t]$ such that $|\tau(x) - s| < \varepsilon_1$. Then $\Phi(x,s) \in V$ and there is a neighbourhood U_x of x such that for all $y \in U_x$, $\Phi(y,s) \in V$. Hence $\tau(y) < s < \tau(x) + \varepsilon_1$. We also have $\tau(y) > \tau(x) - \varepsilon_1$ owing to the following: consider $d := \inf\{||\Phi(x,\tau(x) - \varepsilon_1) - v|| : v \in V\} > 0$. Note that the Lipschitz condition implies that

$$\exists L > 0 \quad \forall\, t > 0, \quad y \in \mathcal{D} : ||\Phi(x,t) - \Phi(y,t)|| \leq ||x - y|| e^{Lt},$$

and for all $t \in [0, \tau(x) - \varepsilon_1]$,

$$||\Phi(x,t) - \Phi(y,t)|| \leq ||x - y|| e^{L(\tau(x)-\varepsilon_1)},$$

and without loss of generality, we can assume that $\forall\, y \in U_x$, we have $||x - y|| e^{L(\tau(x)-\varepsilon_1)} < d/2$. Thus we have $\forall\, v \in V$:

$$\begin{aligned} 0 < d \leq ||\Phi(x,t) - v|| &= ||\Phi(x,t) - \Phi(y,t) + \Phi(y,t) - v|| \\ &\leq ||\Phi(x,t) - \Phi(y,t)|| + ||\Phi(y,t) - v|| \\ &\leq ||x - y|| e^{L(\tau(x)-\varepsilon_1)} + ||\Phi(y,t) - v|| < \frac{d}{2} + ||\Phi(y,t) - v||, \end{aligned}$$

and so for all $y \in U_x$, we have $\inf\{||\Phi(y,t) - v|| : v \in V, t \in [0, \tau(x) - \varepsilon_1]\} \geq d/2 > 0$ and thus $\tau(y) > \tau(x) - \varepsilon_1$. So τ is continuous on $\mathcal{D}$. Let then $T := \sup_{x\in\mathcal{D}} \tau(x) < \infty$. Note that for all $x \in \mathcal{D}$ we have that for all $t > T$, $\Phi(x,t) \in U'$ and $B_\delta(\Phi(x,t)) \subset U$.

Let further $\eta > 0$. Then with proposition A.3, there are $\theta > 0$, $n_\theta \in \mathbb{N}$ such that for all $x \in \mathcal{D}$,

$$\Pr(X^\theta(n_\theta) \in B_\delta(\Phi(x,T)) \subset U | X^\theta(0) = x) > \eta,$$

and so

$$\Pr(X^\theta(n_\theta) \in U) > \eta.$$

From here it is easy to see that the first hit time of U for $\{X^\theta(t)\}_{t\in\mathbb{N}_0}$ is almost surely finite, i.e. the earliest time t for which $X^\theta(t) \in U$: Let $Z(k) := X^\theta(kn_\theta)$ for $k \in \mathbb{N}_0$ and let S be the first hit time of U for $\{Z(k)\}_{k\in\mathbb{N}_0}$, such that S is a random variable with values in $\mathbb{N}_0 \cup \{\infty\}$. Clearly the first hit time of U for $\{X^\theta(t)\}_{t\in\mathbb{N}_0}$ is smaller than for $\{Z(k)\}_{k\in\mathbb{N}_0}$.

We have that for all $z \in \mathcal{D}$ and all $k \in \mathbb{N}$:

$$\Pr(Z_{k+1} \in B_\delta(\Phi(z,T)) \subset U | Z_k = z) > \eta$$

and hence

$$\Pr(Z_{k+1} \in U) > \eta.$$

Then we have for S,

$$\Pr(S \le k+1) = \Pr(S \le k) + (1 - \Pr(S \le k)) \Pr(Z_{k+1} \in U) > \Pr(S \le k)(1-\eta) + \eta,$$

and a quick induction argument yields

$$\Pr(S \le k+1) > 1 - (1-\eta)^k (1 - (1-\eta)\Pr(S = 0)).$$

The probability of a finite hitting time is then

$$\Pr(S \in \mathbb{N}_0) = \lim_{k\to\infty} \Pr(S \le k+1) \ge 1 - \lim_{k\to\infty} (1-\eta)^k (1 - (1-\eta)\Pr(S=0)) = 1.$$

In particular, the hitting time of U for $\{X^\theta(t)\}_{t\in\mathbb{N}_0}$ is finite almost surely. ■

The previous proposition A.4 together with the consideration that an attracting mutation limit is approximated by asymptotically stable mutation equilibria and the immediately following corollary show proposition 3.2:

Corollary A.5. *If x^M is a globally asymptotically stable equilibrium of RMD and U an open neighbourhood of x^M, then there is $\theta > 0$ such that the stochastic process $\{X^\theta(n)\}_{n\ge0}$ defined in condition (A 9) visits U infinitely often almost surely.*

Proof. Consider for any finite $t' \in \mathbb{N}_0$ the probability that $\{X^\theta(n)\}_{n\ge0}$ will not visit U afterwards. This is clearly the same as the probability that the process $\{Z^\theta(n)\}_{n\ge0}$ induced by condition (A 9) and starting in $X^\theta(t')$, i.e. $Z^\theta(0) = X^\theta(t')$ almost surely, will not visit U at all. The previous proposition A.4 shows that this probability is 0, which concludes the proof. ■

(c) MBL-DPU in stable games

The following proposition shows that, in the class of stable games, the Nash equilibrium is an attracting mutation limit in the sense of [10, definition 4.7], i.e. that it is approximated by asymptotically stable equilibria of RMD, regardless of the choice of mutation parameters for diminishing mutation. From this, the convergence of MDL-DPU follows directly with our main result on MBL-DPU and RMD as stated in the subsequent corollary.

Proposition A.6. *Let $f \in C^1(\mathcal{D}, \mathbb{R}^{A_1\times\ldots\times A_n})$ be a continuously differentiable fitness function (or equivalently a payoff function), such that f is a stable population game in the sense of [36, definition 3.3.1], i.e.*

$$\forall\, x, y \in \mathcal{D} : (y-x)^T (f(y) - f(x)) \le 0. \tag{A 10}$$

Then the Nash equilibrium for f is an attracting mutation limit.

Proof. We need to consider that the Nash equilibrium need not be unique. However, it is a convex set for stable population games, as is known from the variational inequality corresponding to the stability definition (e.g. [25, theorem 2.3.5]). In particular, the Nash equilibrium is a single connected component in this case, and with [10, proposition 4.3] it is a mutation limit, and hence approximated by equilibria of RMD. It remains to show that these equilibria are asymptotically stable. Note that we can rewrite RMD as follows:

$$\begin{aligned}\dot{x}_{ih}(t) &= x_{ih}(t)\left(f_{ih}(x(t)) - \sum_{k\in A_i} x_{ik}(t) f_{ik}(x(t))\right) + M_i(c_{ih} - x_{ih}(t)) \\ &= x_{ih}(t)\left(f_{ih}(x(t)) + M_i\frac{c_{i_h}}{x_{ih}(t)} - \sum_{k\in A_i} x_{ik}(t)\left(f_{ik}(x(t)) + M_i\frac{c_{i_k}}{x_{ik}(t)}\right)\right).\end{aligned}$$

With $m_{ih}(x) = M_i\frac{c_{ih}}{x_{ih}}$, this is the RD for the population game $f+m$. Asymptotic stability then follows from condition (A 10) holding for $f+m$ as a strict inequality, whenever $x \neq y$ for x, $y \in \mathcal{D}^\circ$:

$$\begin{aligned}&(y-x)^T(f(y)+m(y)-f(x)-m(x)) \\ &= \underbrace{(y-x)^T(f(y)-f(x))}_{\leq 0} + (y-x)^T(m(y)-m(x)) \\ &\leq (y-x)^T(m(y)-m(x)) = \sum_{i,h} M_i y_{ih}\frac{c_{ih}}{y_{ih}} - \sum_{i,h} M_i y_{ih}\frac{c_{ih}}{x_{ih}} - \sum_{i,h} M_i x_{ih}\frac{c_{ih}}{y_{ih}} + \sum_{i,h} M_i x_{ih}\frac{c_{ih}}{x_{ih}} \\ &= \sum_{i\in P} M_i \underbrace{\sum_{h\in A_i} c_{ih}}_{=1} - \sum_{i,h} M_i y_{ih}\frac{c_{ih}}{x_{ih}} - \sum_{i,h} M_i x_{ih}\frac{c_{ih}}{y_{ih}} + \sum_{i\in P} M_i \underbrace{\sum_{h\in A_i} c_{ih}}_{=1} \\ &= \sum_{i\in P} M_i\left(2 - \sum_{h\in A_i} c_{ih}\underbrace{\left(\frac{y_{ih}}{x_{ih}} + \frac{x_{ih}}{y_{ih}}\right)}_{>2 \text{ for } x\neq y}\right) < \sum_{i\in P} M_i\left(2 - 2\sum_{h\in A_i} c_{ih}\right) = 0.\end{aligned}$$

With this, $f+m$ is a strictly stable game in the sense of [36, theorem 7.2.4] which states that its equilibrium is unique and globally asymptotically stable. Hence, for every choice of $c \in \mathcal{D}^\circ$ and strictly positive $(M_i)_{i\in P}$, the equilibrium is asymptotically stable, making the Nash equilibrium for f an attracting mutation limit. ■

The following corollary then proves proposition 3.3:

Corollary A.7. *Let $f \in C^1(\mathcal{D}, \mathbb{R}^{A_1\times\ldots\times A_n})$ be a stable game in the sense of proposition A.6. Then for every open neighbourhood U of the Nash equilibrium (set), there is $\theta > 0$ such that the stochastic process $\{X^\theta(n)\}_{n\geq 0}$ defined in condition (A 9) visits U infinitely often almost surely.*

Proof. The claim directly follows from the Nash equilibrium being an attracting mutation limit according to proposition A.6 and thus being approximated by a sequence of globally asymptotically stable equilibria of RMD. Applying corollary A.5 concludes the proof. ■

References

1. Silver D *et al.* 2017 Mastering the game of Go without human knowledge. *Nature* **550**, 354–359. (doi:10.1038/nature24270)
2. Watkins CJ, Dayan P. 1992 Q-learning. *Mach. Learn.* **8**, 279–292. (doi:10.1023/A:1022676722315)
3. Sato Y, Akiyama E, Farmer JD. 2002 Chaos in learning a simple two-person game. *Proc. Natl Acad. Sci. USA* **99**, 4748–4751. (doi:10.1073/pnas.032086299)

4. Mertikopoulos P, Sandholm WH. 2016 Learning in games via reinforcement and regularization. *Math. Oper. Res.* **41**, 1297–1324. (doi:10.1287/moor.2016.0778)
5. Omidshafiei S *et al.* 2019 α-rank: multi-agent evaluation by evolution. *Sci. Rep.* **9**, 9937. (doi:10.1038/s41598-019-45619-9)
6. Bowling M, Veloso M. 2002 Multiagent learning using a variable learning rate. *Artif. Intell.* **136**, 215–250. (doi:10.1016/S0004-3702(02)00121-2)
7. Weibull JW. 1995 *Evolutionary game theory*. Cambridge, Mass: MIT Press.
8. Börgers T, Sarin R. 1997 Learning through reinforcement and replicator dynamics. *J. Econ. Theory* **77**, 1–14. (doi:10.1006/jeth.1997.2319)
9. Cross JG. 1973 A stochastic learning model of economic behavior. *Quart. J. Econ.* **87**, 239–266. (doi:10.2307/1882186)
10. Bauer J, Broom M, Alonso E. 2019 The stabilization of equilibria in evolutionary game dynamics through mutation: mutation limits in evolutionary games. *Proc. R. Soc. A* **475**, 20190355 (doi:10.1098/rspa.2019.0355)
11. Ritzberger K, Weibull JW. 1995 Evolutionary selection in normal-form games. *Econometrica* **63**, 1371–1399. (doi:10.2307/2171774)
12. Hart S, Mas-Colell A. 2003 Uncoupled dynamics do not lead to Nash equilibrium. *Am. Econ. Rev.* **93**, 1830–1836. (doi:10.1257/000282803322655581)
13. Leslie DS, Collins EJ. 2003 Convergent multiple-timescales reinforcement learning algorithms in normal form games. *Ann. Appl. Probab.* **13**, 1231–1251. (doi:10.1214/aoap/1069786497)
14. Kaisers M, Tuyls K. 2010 Frequency adjusted multi-agent Q-learning. In *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)* (eds van der Hoek, Kaminka, Lespérance, Luck and Sen), *10–14 May 2010, Toronto, Canada*, pp. 309–315. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).
15. Bauer J. 2020 *Mutation in evolutionary game dynamics and learning; towards evolving topologies of interaction networks*. PhD thesis, University of London.
16. Rustichini A. 1999 Optimal properties of stimulus–response learning models. *Games Econ. Behav.* **29**, 244–273. (doi:10.1006/game.1999.0712)
17. Sato Y, Crutchfield JP. 2003 Coupled replicator equations for the dynamics of learning in multiagent systems. *Phys. Rev. E* **67**, 015206. (doi:10.1103/PhysRevE.67.015206)
18. Marsili M, Challet D, Zecchina R. 2000 Exact solution of a modified El Farol's bar problem: efficiency and the role of market impact. *Physica A* **280**, 522–553. (doi:10.1016/S0378-4371(99)00610-X)
19. Kianercy A, Galstyan A. 2012 Dynamics of Boltzmann Q learning in two-player two-action games. *Phys. Rev. E* **85**, 041145. (doi:10.1103/PhysRevE.85.041145)
20. Tuyls K, Vanschoenwinkel B. 2006 An evolutionary dynamical analysis of multi-agent learning in iterated games. *Auton. Agents Multi-Agent Syst.* **12**, 115–153. (doi:10.1007/s10458-005-3783-9)
21. Chapman AC, Leslie DS, Rogers A, Jennings NR. 2013 Convergent learning algorithms for unknown reward games. *SIAM J. Control Optim.* **51**, 3154–3180. (doi:10.1137/120893501)
22. Abe K, Sakamoto M, Iwasaki A. 2022 Mutation-driven follow the regularized leader for last-iterate convergence in zero-sum games. In *Proc. of the Thirty-Eighth Conf. on Uncertainty in Artificial Intelligence*, pp. 1–10. PMLR.
23. Abe K, Ariu K, Sakamoto M, Toyoshima K, Iwasaki A. 2023 Last-iterate convergence with full and noisy feedback in two-player zero-sum games. In *Proc. of The 26th Int. Conf. on Artificial Intelligence and Statistics*, pp. 7999–8028. PMLR.
24. Abe K, Ariu K, Sakamoto M, Iwasaki A. 2024 Adaptively perturbed mirror descent for learning in games. In *Forty-First Int. Conf. on Machine Learning*. PMLR.
25. Facchinei F, Pang JS. 2003 *Finite-dimensional variational inequalities and complementarity problems*, vol. 1. Springer series in operations research. New York: Springer.
26. Falniowski F, Mertikopoulos P. 2024 On the discrete-time origins of the replicator dynamics: from convergence to instability and chaos. (doi:10.48550/arXiv.2402.09824)
27. Cai Y, Oikonomou A, Zheng W. 2022 Tight last-iterate convergence of the extragradient and the optimistic gradient descent-ascent algorithm for constrained Monotone variational inequalities. (https://doi.org/10.48550/arXiv.2204.09228)
28. Liu Q, Weisz G, György A, Jin C, Szepesvari C. 2023 Optimistic natural policy gradient: a simple efficient policy optimization framework for online RL. *Adv. Neural Inf. Process. Syst.* **36**, 3560–3577.

29. Sokota S, D'Orazio R, Kolter JZ, Loizou N, Lanctot M, Mitliagkas I, Brown N, Kroer C. 2023 A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *The Eleventh Int. Conf. on Learning Representations*. International Conference on Learning Representations (ICLR).
30. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. 2017 Proximal policy optimization algorithms. (https://doi.org/10.48550/arXiv.1707.06347)
31. Mali IG, Czibula G. 2023 Policy-based reinforcement learning in the generalized rock-paper-scissors game. In *ESANN 2023 proceesdings*, pp. 345–350. Bruges (Belgium) and online. Louvain-la-Neuve, Belgium: CIACO. (doi:10.14428/esann/2023.ES2023-92)
32. Ratcliffe DS, Hofmann K, Devlin S. 2019 Win or learn fast proximal policy optimisation. In *2019 IEEE Conference on Games (CoG)*, pp. 1–4 London, UK. IEEE. (doi:10.1109/CIG.2019.8848100)
33. Hofbauer J, Sigmund K. 1998 *Evolutionary games and population dynamics*. Cambridge: Cambridge University Press.
34. Page KM, Nowak MA. 2002 Unifying evolutionary dynamics. *J. Theor. Biol.* **219**, 93–98. (doi:10.1006/jtbi.2002.3112)
35. Norman MF. 1972 *Markov processes and learning models*. Number vol. 84 in Mathematics in science and engineering. New York: Academic Press.
36. Sandholm WH. 2010 *Population games and evolutionary dynamics*. Economic learning and social evolution. Cambridge, Mass: MIT Press.
37. Nee S. 1989 Antagonistic co-evolution and the evolution of genotypic randomization. *J. Theor. Biol.* **140**, 499–518. (doi:10.1016/S0022-5193(89)80111-0)
38. Song Y, Gokhale CS, Papkou A, Schulenburg H, Traulsen A. 2015 Host-parasite coevolution in populations of constant and variable size. *BMC Evol. Biol.* **15**, 212. (doi:10.1186/s12862-015-0462-6)
39. Teschl G. 2012 *Ordinary differential equations and dynamical systems*. Providence, RI: American Mathematical Society.
40. Jordan J. 1993 Three problems in learning mixed-strategy nash equilibria. *Games Econ. Behav.* **5**, 368–386. (doi:10.1006/game.1993.1022)
41. Bauer J, West S, Alonso E, Broom M. 2025 Mutation-bias learning: an evolutionary game dynamics approach to convergence analysis in multi-agent reinforcement learning. Figshare. (doi:10.6084/m9.figshare.c.8174357)