



City Research Online

City St George's, University of London

Citation: 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. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 482(2329), 20250449. doi: 10.1098/rspa.2025.0449

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://city-ref-test.eprints-hosting.org/id/eprint/36250/>

Link to published version: <https://doi.org/10.1098/rspa.2025.0449>

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).



Subject Areas:

xxxxx, xxxxx, xxxxx

Keywords:

xxxx, xxxx, xxxxx

Author for correspondence:

Johann Bauer

e-mail:

johann.bauer@tu-darmstadt.de

Mutation-Bias Learning in Games

Johann Bauer^{1,3}, Sheldon West², Eduardo Alonso², Mark Broom¹

¹ Department of Mathematics, City St. George's, University of London, UK

² Department of Computer Science, City St. George's, University of London, UK

³ Centre for Cognitive Science, TU Darmstadt, Germany

We present two variants of a multi-agent reinforcement learning algorithm based on evolutionary game theoretic considerations. The intentional simplicity of one variant enables us to rigorously prove results on its relationship to a system of ordinary differential equations of replicator-mutator dynamics type, allowing us to present proofs on the algorithm's convergence conditions in various settings via its ODE counterpart, including in stable and zero-sum games. The more complicated variant enables comparisons to Q-learning based algorithms. We further compare both variants experimentally to WoLF-PHC and frequency-adjusted Q-learning on a range of settings, illustrating cases of increasing dimensionality where our variants preserve convergence in contrast to more complicated algorithms. The availability of analytic results provides a degree of transferability of results as compared to a focus on specific game classes or a purely empirical case studies, illustrating the general utility of a dynamical systems perspective on multi-agent reinforcement learning when addressing questions of convergence and reliable generalisation.

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 generalise 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 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, prop. 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 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 replicator dynamics (RMD) which we had specified and analysed previously in [10], a non-linear *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 had 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 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 Win-or-Learn-Fast Policy-Hill-Climbing (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

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. A similar approach is pursued by the frequency-adjusted Q-learning algorithm (FAQ) in [20] with a correction given in [14]. However, both strands hinge on assumptions which have not been proved, and therefore no rigorous analyses are provided. Nonetheless, we choose FAQ-learning as a comparison, as [14] claims it to be linked to an ODE system similar to RMD and as it is a 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 for FAQ-learning, only the two-player two-action analysis of the 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 multiple timescales 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 timescale separation results in a fundamentally more complicated approach and more complicated algorithms, including the additional requirement to keep track of timescales 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. 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

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

demonstrated for the replicator dynamics 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 optimisation 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 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].

Finite normal-form games. A normal-form game is a tuple (P, A, r) , where $P = \{1, \dots, N\}$ represents the set of players, $A = \times_{i \in P} A_i$ where $A_i = \{1, \dots, n_i\}$ is the set of pure strategies of each player i ,² and $r = (r_i)_{i \in P}$ is a family of functions with $r_i : A \rightarrow \mathbb{R}$ 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 $(x_{ih})_{h \in A_i} \in \mathcal{D}_i := \{\xi \in \mathbb{R}_{\geq 0}^{A_i} : \sum_{h \in A_i} \xi_h = 1\}$. We call such an x_i the mixed strategy of player i .³ We will call mixed strategies simply *strategies*, where there is no danger of confusion.

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

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

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:

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

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

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

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 $\tilde{x}$.

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.

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 maximises 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.

Replicator-mutator dynamics. We consider the multi-population replicator-mutator dynamics we formulated in [10], which is a special case of general replicator-mutator dynamics, e.g., [34]: For all $i \in P$, let $M_i > 0$ be a mutation parameter, $c_i \in \mathcal{D}_i^\circ$ (denoting the interior of $\mathcal{D}_i$) some fixed parameter and $f_i : \mathcal{D} \rightarrow \mathbb{R}^{A_i}$ a continuously differentiable fitness function. Then the replicator-mutator dynamics is given for $i \in P$, $h \in A_i$ by

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

In case that $M_i = 0$ for all $i \in P$, **RMD** reduces to the standard multi-population replicator dynamics (RD). One possible (and usual) conceptualisation 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, alg. 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, alg. 2)–where the policy corresponds to logistic choice based on action-value estimates which are updated with a mutation bias perturbation.

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

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

MBL with direct policy update (MBL-DPU). MBL-DPU, alg. 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.

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

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

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, alg. 2, the perturbation affects the action-value updates instead of the policy. Hence, this version resembles the perturbation term of FAQ-learning [14,19], and allows for a closer comparison. In particular, restricting the adjustment in line 6 by applying a minimum is parallel to FAQ-learning. 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:

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

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.

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 $\{(X_{ih}^\theta(t))_{i,h}\}_{t \in \mathbb{N}_0}$ induced by MBL-DPU converges to RMD in the sense that for all $\varepsilon > 0$:*

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

where $n_\theta \theta \rightarrow T$ for $\theta \rightarrow 0$, 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 $\Phi(x_0, T) \rightarrow x^M$ as $T \rightarrow \infty$. 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 $M \rightarrow 0$, called a *mutation limit*. Furthermore, it was shown that for the game of Matching Pennies 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 $f \in C^1(\mathcal{D}, \mathbb{R}^{A_1 \times \dots \times A_n})$ be a continuously differentiable fitness function, such that f is a stable game in the sense of [36, definition 3.3.1], i.e.:*

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

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-learning, a similar logistic choice learning rule based on Q-learning, converges to a perturbation of the replicator dynamics, 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 likely 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 appendix B. 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 let to approach ∞ as this collides with the $\theta \rightarrow 0$ 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 replicator dynamics. 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), Matching Pennies (MP), Rock-Paper-Scissors (RPS) with 3, 5 and 9 available strategies, and the three-player Matching Pennies (3MP) games. We compare MBL-DPU and MBL-LC to FAQ, [14], and WoLF-PHC, [6]. For details on the games' payoffs and further experiments, cf. appendix B.

Prisoner's Dilemma (PD). 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 destabilise 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 appendix B(a).

Zero-sum games—Matching Pennies (MP). As a second, structurally different case, we consider zero-sum games which have interior Nash equilibria. For the games considered here it is straightforward to check that the eigenvalues of the Jacobian of **RMD** in the neighbourhood of the Nash equilibrium only have 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-learning and WoLF-PHC, fig. 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, MBL-LC, FAQ, or WoLF-PHC. With MP being a planar system and the Poincaré-Bendixson theorem, the complexity of the system is still relatively small.

Zero-sum games—Rock-Paper-Scissors (RPS). For the higher dimensional settings, i.e., RPS with 3, 5 and 9 strategies, we still observe convergence for MBL-DPU, fig. 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 2D-projection. For MBL-LC, fig. 3, and FAQ, fig. 4, we observe convergence in the RPS-3 case, but both algorithms deteriorate in higher dimensions, MBL-LC for RPS-9, fig. 3c, and FAQ for RPS-5 and RPS-9, figs. 4b and 4c, 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, fig. 5, it loses the ability to learn the Nash equilibrium for RPS-9, fig. 5c, with trajectories seemingly getting stuck near the boundary of $\mathcal{D}$.

Three-player Matching Pennies. Beyond the two-player case, we compare MBL in a three-player Matching Pennies setting introduced in [37]. 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 center of $\mathcal{D}$. All four algorithms fail to learn the Nash equilibrium, fig. 6 (MBL-LC not shown, cf. appendix B(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 generalise 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. E.g., 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 stabilises 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 ODE underpinning 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 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.

It should be noted 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. Additionally, a clarification of the convergence properties of MBL-LC would allow 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.

Acknowledgements. Insert acknowledgment text here.

References

1. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen Y, Lillicrap T, Hui F, Sifre L, Driessche Gvd, Graepel T, Hassabis D. 2017 Mastering the game of Go without human knowledge. *Nature* **550**, 354–359. ([10.1038/nature24270](https://doi.org/10.1038/nature24270))
2. Watkins CJ, Dayan P. 1992 Q-Learning. *Machine Learning* **8**, 279–292. ([10.1023/A:1022676722315](https://doi.org/10.1023/A:1022676722315))
3. Sato Y, Akiyama E, Farmer JD. 2002 Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences* **99**, 4748–4751. ([10.1073/pnas.032086299](https://doi.org/10.1073/pnas.032086299))
4. Mertikopoulos P, Sandholm WH. 2016 Learning in Games via Reinforcement and Regularization. *Mathematics of Operations Research* **41**, 1297–1324. ([10.1287/moor.2016.0778](https://doi.org/10.1287/moor.2016.0778))
5. Omidshafiei S, Papadimitriou C, Piliouras G, Tuyls K, Rowland M, Lespiau JB, Czarnecki WM, Lanctot M, Perolat J, Munos R. 2019 -Rank: Multi-Agent Evaluation by Evolution. *Scientific Reports* **9**. ([10.1038/s41598-019-45619-9](https://doi.org/10.1038/s41598-019-45619-9))
6. Bowling M, Veloso M. 2002 Multiagent learning using a variable learning rate. *Artificial*

- Intelligence **136**, 215–250. ([10.1016/S0004-3702\(02\)00121-2](https://doi.org/10.1016/S0004-3702(02)00121-2))
7. Weibull JW. 1995 *Evolutionary game theory*. Cambridge, Mass.: MIT Press. OCLC: 255505178.
 8. Börgers T, Sarin R. 1997 Learning Through Reinforcement and Replicator Dynamics. *Journal of Economic Theory* **77**, 1–14. ([10.1006/jeth.1997.2319](https://doi.org/10.1006/jeth.1997.2319))
 9. Cross JG. 1973 A Stochastic Learning Model of Economic Behavior. *The Quarterly Journal of Economics* **87**, 239–266. ([10.2307/1882186](https://doi.org/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. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **475**, 20190355. Publisher: Royal Society ([10.1098/rspa.2019.0355](https://doi.org/10.1098/rspa.2019.0355))
 11. Ritzberger K, Weibull JW. 1995 Evolutionary Selection in Normal-Form Games. *Econometrica* **63**, 1371–1399. ([10.2307/2171774](https://doi.org/10.2307/2171774))
 12. Hart S, Mas-Colell A. 2003 Uncoupled Dynamics Do Not Lead to Nash Equilibrium. *American Economic Review* **93**, 1830–1836. ([10.1257/000282803322655581](https://doi.org/10.1257/000282803322655581))
 13. Leslie DS, Collins EJ. 2003 Convergent multiple-timescales reinforcement learning algorithms in normal form games. *The Annals of Applied Probability* **13**, 1231–1251. ([10.1214/aoap/1069786497](https://doi.org/10.1214/aoap/1069786497))
 14. Kaisers M, Tuyls K. 2010 Frequency Adjusted Multi-agent Q-learning. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1* AAMAS '10 pp. 309–316 Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems. event-place: Toronto, Canada.
 15. Bauer J. 2020 *Mutation in evolutionary game dynamics and learning; towards evolving topologies of interaction networks*. PhD Thesis City, University of London. <https://openaccess.city.ac.uk/id/eprint/26476/>.
 16. Rustichini A. 1999 Optimal Properties of Stimulus—Response Learning Models. *Games and Economic Behavior* **29**, 244–273. ([10.1006/game.1999.0712](https://doi.org/10.1006/game.1999.0712))
 17. Sato Y, Crutchfield JP. 2003 Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E* **67**. ([10.1103/PhysRevE.67.015206](https://doi.org/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: Statistical Mechanics and its Applications* **280**, 522–553. ([10.1016/S0378-4371\(99\)00610-X](https://doi.org/10.1016/S0378-4371(99)00610-X))
 19. Kianercy A, Galstyan A. 2012 Dynamics of Boltzmann Q learning in two-player two-action games. *Physical Review E* **85**, 041145. ([10.1103/PhysRevE.85.041145](https://doi.org/10.1103/PhysRevE.85.041145))
 20. Tuyls K, 't Hoen PJ, Vanschoenwinkel B. 2006 An Evolutionary Dynamical Analysis of Multi-Agent Learning in Iterated Games. *Autonomous Agents and Multi-Agent Systems* **12**, 115–153. ([10.1007/s10458-005-3783-9](https://doi.org/10.1007/s10458-005-3783-9))
 21. Chapman AC, Leslie DS, Rogers A, Jennings NR. 2013 Convergent Learning Algorithms for Unknown Reward Games. *SIAM Journal on Control and Optimization* **51**, 3154–3180. ([10.1137/120893501](https://doi.org/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 *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence* pp. 1–10. PMLR. ISSN: 2640-3498.
 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 *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* pp. 7999–8028. PMLR. ISSN: 2640-3498.
 24. Abe K, Ariu K, Sakamoto M, Iwasaki A. 2024 Adaptively Perturbed Mirror Descent for Learning in Games. .
 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. arXiv:2402.09824 [cs, math] ([10.48550/arXiv.2402.09824](https://doi.org/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. arXiv:2204.09228 [cs, math] ([10.48550/arXiv.2204.09228](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. *Advances in Neural Information Processing Systems* **36**, 3560–3577.
 29. Sokota S, D'Orazio R, Kolter JZ, Loizou N, Lanctot M, Mitliagkas I, Brown N, Kroer C. 2022 A Unified Approach to Reinforcement Learning, Quantal Response Equilibria, and Two-Player Zero-Sum Games. .

30. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. 2017 Proximal Policy Optimization Algorithms. *arXiv:1707.06347*. arXiv:1707.06347 [cs] ([10.48550/arXiv.1707.06347](https://arxiv.org/abs/1707.06347))
31. Mali IG, Czibula G. 2023 Policy-Based Reinforcement Learning in the Generalized Rock-Paper-Scissors Game. In *ESANN 2023 proceedings* pp. 345–350 Bruges (Belgium) and online. Ciaco - i6doc.com. ([10.14428/esann/2023.ES2023-92](https://ciaco-i6doc.com/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, United Kingdom. IEEE. ([10.1109/CIG.2019.8848100](https://doi.org/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. *Journal of Theoretical Biology* **219**, 93–98. ([10.1006/jtbi.2002.3112](https://doi.org/10.1006/jtbi.2002.3112))
35. Norman MF. 1972 *Markov processes and learning models*. Number v. 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. OCLC: 706018134.
37. Jordan J. 1993 Three Problems in Learning Mixed-Strategy Nash Equilibria. *Games and Economic Behavior* **5**, 368–386. ([10.1006/game.1993.1022](https://doi.org/10.1006/game.1993.1022))
38. Teschl G. 2012 *Ordinary differential equations and dynamical systems*. Providence, RI: American Mathematical Society.

A. Proofs

The proofs employ a result proved in [35, p. 118], which we state in the following and then proceed to prove propositions 3.1 and 3.2.

(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

$$H_n^\theta = \Delta X_n^\theta / \theta$$

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:

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

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

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

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

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

$$\forall x \in I : \lim_{\theta \rightarrow 0} \inf_{y \in I_\theta} \|x - y\| = 0. \quad (\text{a.1})$$

There are functions w and s on I such that:

$$\sup_{x \in I_\theta} \|w(x, \theta) - w(x)\| \in \mathcal{O}(\theta), \quad (\text{a.2})$$

$$\sup_{x \in I_\theta} \|s(x, \theta) - s(x)\| \rightarrow 0 \text{ for } \theta \rightarrow 0, \quad (\text{a.3})$$

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

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

$$\lim_{\substack{y \rightarrow x \\ y \in I}} \frac{\|w(y) - w(x) - w'(x)(y - x)\|}{\|y - x\|} = 0. \quad (\text{b.1})$$

513 The function w' is bounded:

$$\sup_{x \in I} \|w'(x)\| < \infty. \quad (\text{b.2})$$

514 The functions w' and s satisfy the Lipschitz condition:

$$\sup_{x, y \in I, x \neq y} \frac{\|w'(x) - w'(y)\|}{\|x - y\|} < \infty, \quad (\text{b.3})$$

$$\sup_{x, y \in I, x \neq y} \frac{\|s(x) - s(y)\|}{\|x - y\|} < \infty. \quad (\text{b.4})$$

515 The function r is bounded:

$$\sup_{\theta \in J, x \in I_\theta} r(x, \theta) < \infty. \quad (\text{c})$$

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

517 In this case, the following hold:

518 (A) $\omega_n(x, \theta) \in \mathcal{O}(\theta)$ uniformly in $x \in I_\theta$ and $n\theta \leq T$ for any $T < \infty$.

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

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

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

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

521 uniformly in $x \in I_\theta$ and $n\theta \leq T$.

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

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

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

524 (b) Convergence of MBL-DPU

525 We restate the simple reinforcement-mutation rule of MBL-DPU in the setting layed out above,
526 denoting the mixed strategies with an upper-case X to underscore that this is a random variable
527 and denoting the dependence on a parameter θ , denoting the whole family of stochastic processes
528 as $\{(X_{ih}^\theta(n))_{i \in P, h \in A_i}\}_{n \geq 0}$. Let $R(x) = (R_{ih}(x))_{i \in P, h \in A_i}$ be a random variable whose probability
529 distribution depends smoothly on $x \in I$ with a discrete, non-negative support which is independent
530 of x , and let $M_i < \bar{M}$ for some upper bound $\bar{M} < \infty$ and all $i \in P$. For a player $i \in P$ and a chosen
531 pure strategy $h \in A_i$, the update rule then is given as follows:

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

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

Proposition A.3. *There is J such that the family of stochastic processes $\{(X_{ih}^\theta(n))_{i \in P, h \in A_i}\}_{n \geq 0}$ given by (A.1) approximates the replicator-mutator dynamics for $\theta \rightarrow 0$ 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 $\{(X_{ih}^\theta(n))_{i \in P, h \in A_i}\}_{n \geq 0}$ satisfies the conditions of theorem A.1. For a player $i \in P$ and a chosen strategy $h \in A_i$ we have:

$$\begin{aligned} H_{ih}^\theta(n+1) &= \Delta X_{ih}^\theta(n+1)/\theta = (1 - X_{ih}^\theta(n))R_{ih}(X^\theta(n)) + M_i(c_{ih} - X_{ih}^\theta(n)) \\ H_{ik}^\theta(n+1) &= \Delta X_{ik}^\theta(n+1)/\theta = -X_{ik}^\theta(n)R_{ih}(X^\theta(n)) + M_i(c_{ik} - X_{ik}^\theta(n)) \text{ for } k \neq h \end{aligned}$$

Note that in this case, $H_{ih}^\theta(n+1)$ 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 $\{(X_{ih}^\theta(n))_{i \in P, h \in A_i}\}_{n \geq 0} \subset I$: 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 $X_{ih}^\theta(n) = x \in I$, then for a player $i \in P$ and a chosen strategy $h \in A_i$ we have

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

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

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

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

Conditions (a.2) & (a.3): Consider first the function w :

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

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 (b.1)–(b.4): 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 (b.3) on I . Similarly, s is smooth and satisfies (b.4).

Condition (c): Again, r does not depend on θ , and is smooth on I , which is compact. Thus it is bounded on I and condition (c) 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 $\varepsilon > 0$,

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

for $n\theta \rightarrow T$, $\theta \rightarrow 0$, 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 $\{(X_{ih}^\theta(n))_{i \in P, h \in A_i}\}_{n \geq 0}$ defined in (A 1) visits U almost surely after finitely many steps.

Proof. Let $\Phi(x, \cdot) : \mathbb{R}_{\geq 0} \rightarrow \mathcal{D}$ 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 $\tau : \mathcal{D} \rightarrow \mathbb{R}$ such that:

$$\tau(x) = \inf\{T > 0 : \Phi(x, T) \in V\}$$

Since the RHS 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 $\tau(y) > \tau(x) - \varepsilon_1$ due to the following:

Consider $d := \inf\{\|\Phi(x, \tau(x) - \varepsilon_1) - v\| : v \in V\} > 0$. Note that the Lipschitz condition implies that there is $L > 0$ such that for all $t > 0$ and all $y \in \mathcal{D}$

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

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 w.l.o.g. we can assume that $\forall y \in U_x$, we have $\|x - y\| e^{L(\tau(x) - \varepsilon_1)} < \frac{d}{2}$. Thus we have for all $v \in V$

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

and so for all $y \in U_x$, we have $\inf\{\|\Phi(y, t) - v\| : v \in V, t \in [0, \tau(x) - \varepsilon_1]\} \geq \frac{d}{2} > 0$ 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 ,

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

and a quick induction argument yields:

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

The probability of a finite hitting time is then:

$$\Pr(S \in \mathbb{N}_0) = \lim_{k \rightarrow \infty} \Pr(S \leq k+1) \geq 1 - \lim_{k \rightarrow \infty} (1 - \eta)^k (1 - (1 - \eta) \Pr(S = 0)) = 1$$

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

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 $\{X^\theta(n)\}_{n \geq 0}$ defined in (A 1) visits U infinitely often almost surely.*

Proof. Consider for any finite $t' \in \mathbb{N}_0$ the probability that $\{X^\theta(n)\}_{n \geq 0}$ will not visit U afterwards. This is clearly the same as the probability that the process $\{Z^\theta(n)\}_{n \geq 0}$ induced by (A 1) 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. $\square$

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 $f \in C^1(\mathcal{D}, \mathbb{R}^{A_1 \times \dots \times A_n})$ 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.*

$$\forall x, y \in \mathcal{D} : (y - x)^T (f(y) - f(x)) \leq 0 \quad (\text{A } 2)$$

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, prop. 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:

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

With $m_{ih}(x) = M_i \frac{c_{ih}}{x_{ih}}$, this is the replicator dynamics for the population game $f + m$. Asymptotic stability then follows from (A 2) holding for $f + m$ as a strict inequality, whenever $x \neq y$ for $x, y \in \mathcal{D}^\circ$:

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

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. This holds independently of the specific choice of $c \in \mathcal{D}^\circ$ and strictly positive $(M_i)_{i \in P}$, which concludes the proof. $\square$

The following corollary then proves proposition 3.3:

Corollary A.7. *Let $f \in C^1(\mathcal{D}, \mathbb{R}^{A_1 \times \dots \times A_n})$ 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 (A 1) 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. $\square$

B. Specification of experiments and further results

This section provides the specification details for the experimental results of section 4 and further results for a broader range of parameter values. It is structured as follows: Each game setting is introduced with its payoff structure together with further results and a short description of the results, in the order of Prisoner's Dilemma (B(a)), Matching Pennies (i), RPS- n games (ii), and three-player Matching Pennies (B(c)). For the two-player settings, the payoff values are given as matrices R_1 and R_2 , giving the payoffs for players one and two respectively, such that if player one chooses the i -th pure strategy from A_1 and player two chooses the j -th pure strategy from A_2 , then the payoffs are given as $r_1(i, j) = [R_1]_{ij}$ and $r_2(i, j) = [R_2]_{ij}$ respectively. The experiments were run on a small cluster of multi-kernel CPUs, but we have checked that they can easily be run on personal hardware.

(a) Prisoner's Dilemma

The experimental results for the Prisoner's Dilemma are based on the following payoff structure:

$$R_1 = \begin{pmatrix} 1 & 5 \\ 0 & 3 \end{pmatrix} \quad R_2 = \begin{pmatrix} 1 & 0 \\ 5 & 3 \end{pmatrix}$$

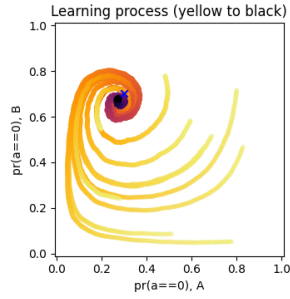
This version has a strict unique Nash equilibrium x^* at:

$$x_1^* = \begin{pmatrix} 1 & 0 \end{pmatrix}^T \quad x_2^* = \begin{pmatrix} 1 & 0 \end{pmatrix}^T$$

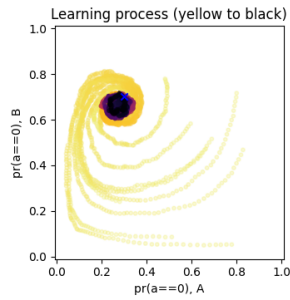
MBL-DPU and MBL-LC. The experimental results (figures 7, 8) illustrate the behaviour of MBL-DPU and its convergence for different mutation strengths M . In accordance with intuition, convergence is quick for high mutation strength at the price of the mutation equilibrium being further away from the Nash equilibrium. For lower values of M , we have that the mutation equilibrium moves closer to the Nash equilibrium while convergence becomes slower. In comparison, MBL-LC (figures 9, 10) behaves similarly while converging much more quickly. An intuition for this is provided when considering that MBL-DPU can be viewed as a linear approximation to MBL-LC for small τ .

FAQ-learning. For FAQ-learning (figures 11, 12), the role of τ corresponds to that of M^{-1} in MBL. We have that, similarly to both MBL variants, with increasing values of τ (i.e., decreasing values of M), the dynamics approaches a region that lies closer to the Nash equilibrium. The intuition here is provided by the fact that the deterministic limit of FAQ is claimed to be a replicator dynamics with a perturbative term whose effect depends on τ and which pulls the system towards the centre of $\mathcal{D}$. Furthermore, convergence is the slower the weaker the perturbative term is, much like in the two MBL variants. In contrast to the MBL variants, FAQ-learning defaults to the usual Q-learning when $x_{ih} \leq \beta$. This effectively neutralises the repelling dynamics at the boundary of $\mathcal{D}$, which would otherwise result in very large (unbounded) changes in the Q-values for very low values of x_{ih} . Note that MBL-LC has x_{ih} occurring in the denominator twice and hence retains the repelling effect at the boundary of $\mathcal{D}$.

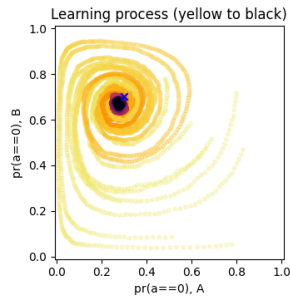
WoLF-PHC. In contrast to the other algorithms, WoLF-PHC (figure 13) follows a chosen direction for some time until it is replaced by a new direction, which results in a discrete sequence of directions and non-smooth trajectories. Convergence to the Nash equilibrium occurs much faster than for the other algorithms in the case of PD. However, strict Nash equilibria are also asymptotically stable in RD and thus PD is a base case which illustrates the different behaviours in a clear-cut situation, as opposed to more challenging and ambiguous situations without strict Nash equilibria.



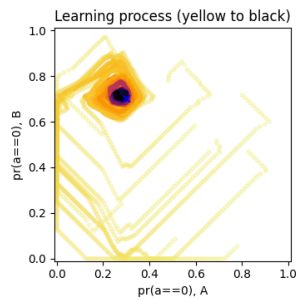
(a) MBL-DPU with $M^{-1} = 20$.



(b) MBL-LC with $M^{-1} = \tau = 20$.



(c) FAQ with $\tau = 20$.



(d) WoLF-PHC with initial learning rate 10^{-1} for Q , win learning rate $1/2 \cdot 10^{-4}$.

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.

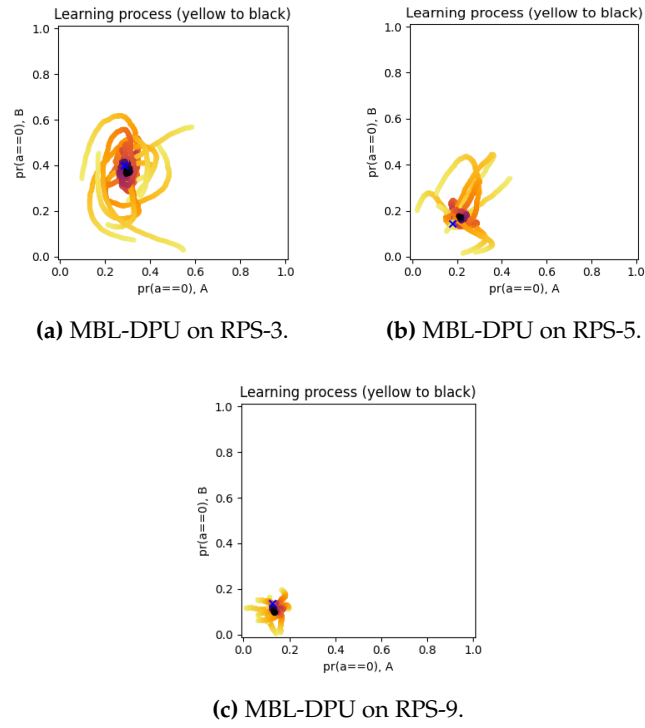


Figure 2: Self-play of MBL-DPU on RPS-3, RPS-5 and RPS-9 games, with $M^{-1} = 20$.

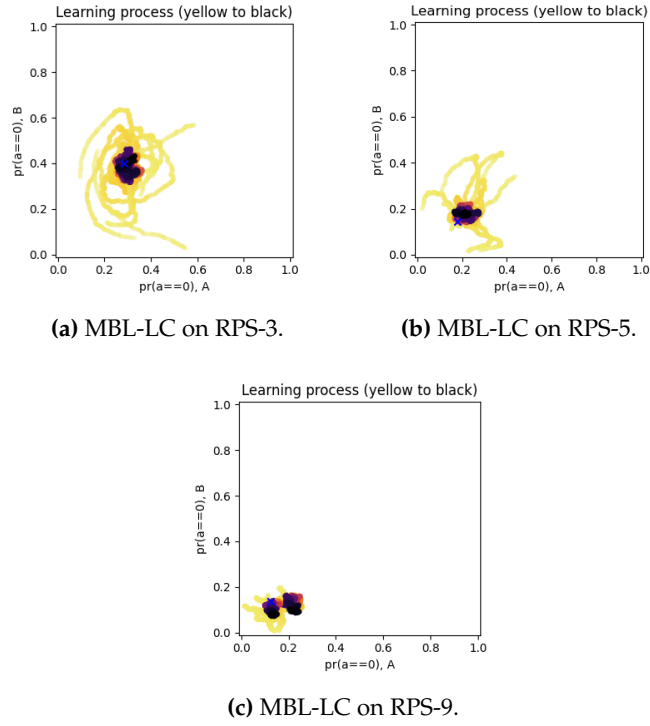


Figure 3: Self-play of MBL-LC on RPS-3, RPS-5 and RPS-9 games, with $M^{-1} = \tau = 20$.

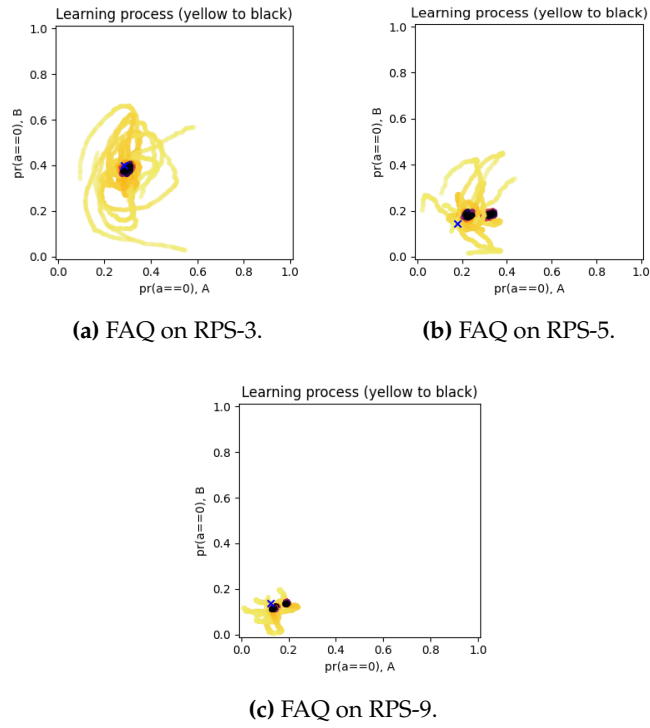


Figure 4: Self-play of FAQ-learning on RPS-3, RPS-5 and RPS-9 games, with $\tau = 20$.

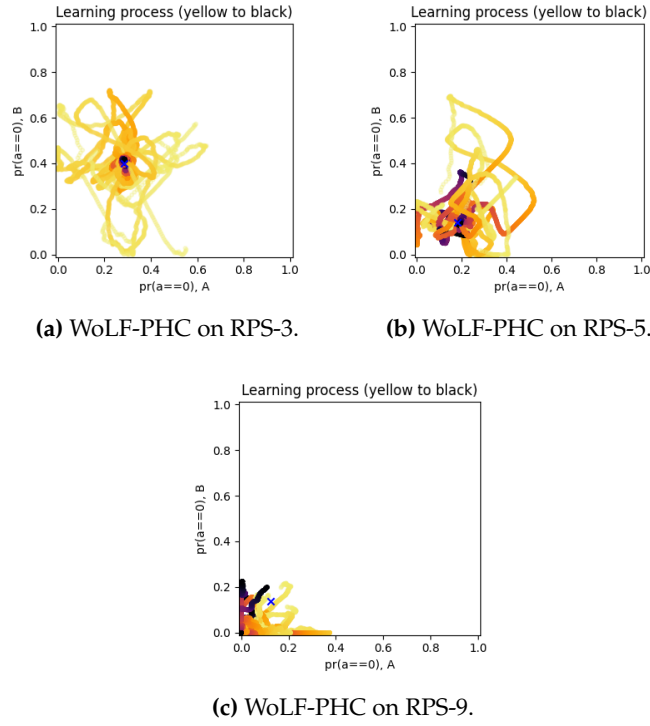


Figure 5: Self-play of WoLF-PHC-learning on RPS-3, RPS-5 and RPS-9 games, with initial learning rate 10^{-1} for Q , win learning rate $1/2 \cdot 10^{-4}$.

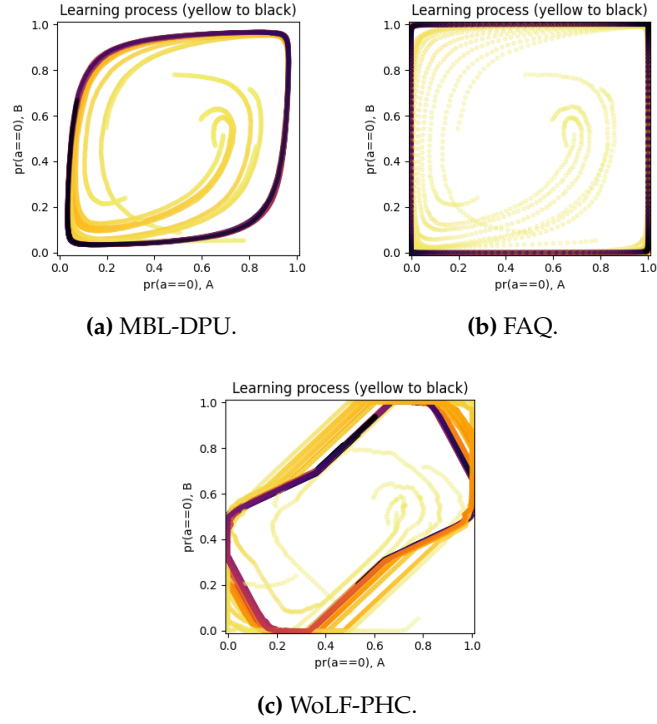


Figure 6: Self-play on 3MP by (a) MBL-DPU with $M^{-1} = 20$, (b) FAQ with $\tau = 20$, and (c) WoLF-PHC with initial learning rate 10^{-1} for Q , win learning rate $1/2 \cdot 10^{-4}$.

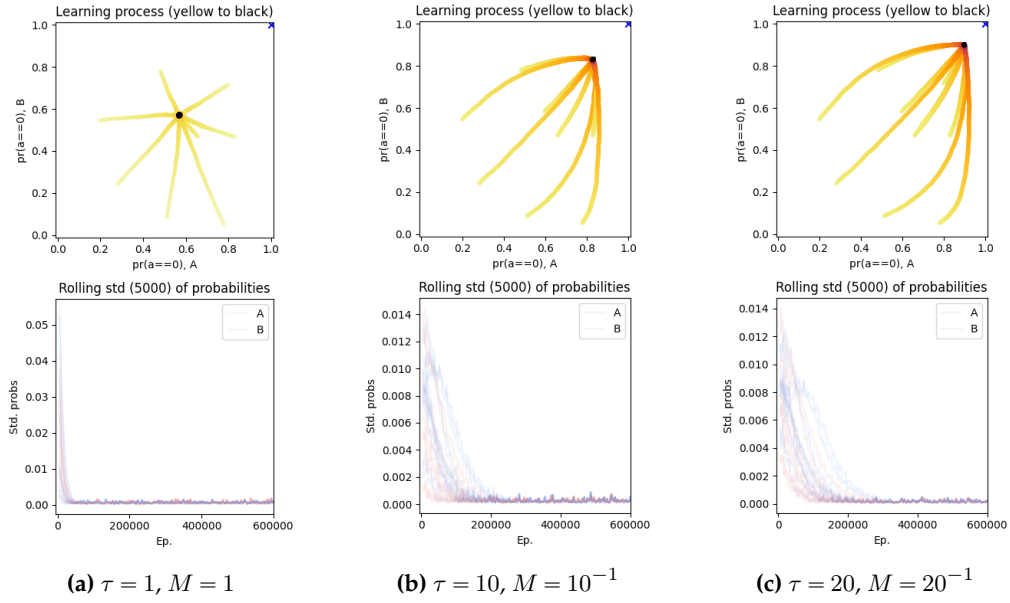


Figure 7: MBL-DPU in self-play on the PD game with different values for τ (1, 10, 20) or M ($1, 10^{-1}, 20^{-1}$) equivalently; $\theta = 10^{-4}$; for 10 different initial conditions. In each subfigure, the upper graph shows the ten trajectories in the projection on the first components of the players' strategies, in this case the 'defect' strategy, with the first player given on the horizontal axis and the second player 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. The lower graph shows the standard deviation of all components of the players' strategies for each point in time over the past 5000 time steps, for each of the ten initial conditions, coloured red and blue for the two players. Time is given on the horizontal axis. The standard deviation is computed with the usual Euclidean metric.

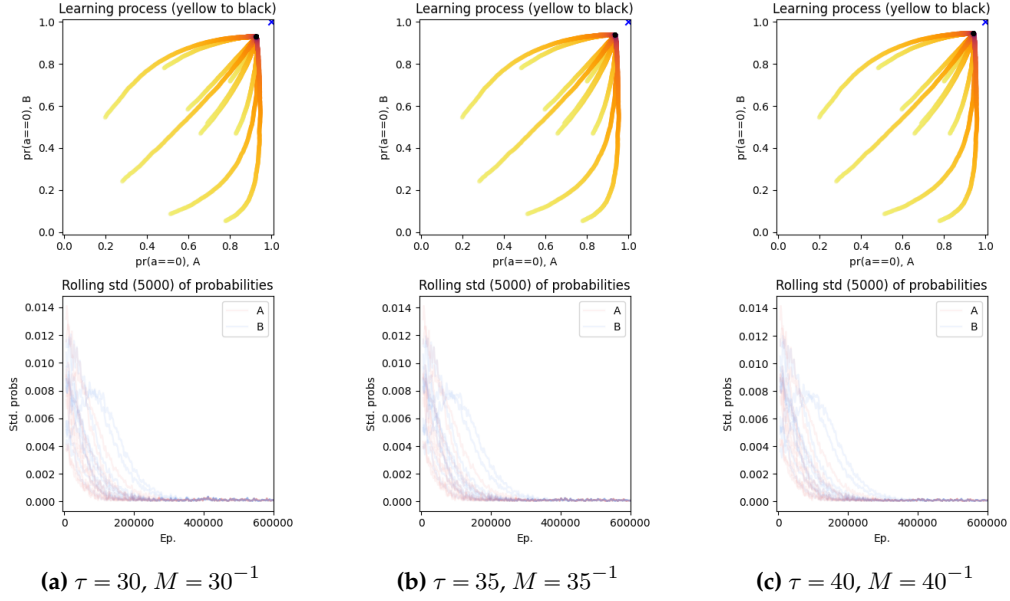


Figure 8: MBL-DPU in self-play on the PD game with different values for τ (30, 35, 40) or M (30^{-1} , 35^{-1} , 40^{-1}) equivalently; $\theta = 10^{-4}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

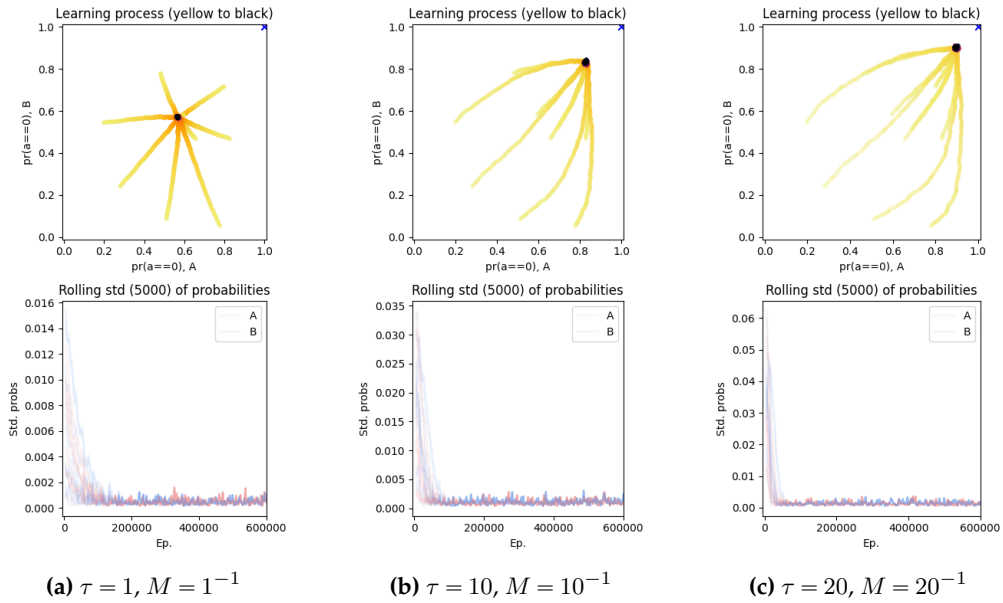


Figure 9: MBL-LC in self-play on the PD game with different values for τ (1, 10, 20) or M (1 , 10^{-1} , 20^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

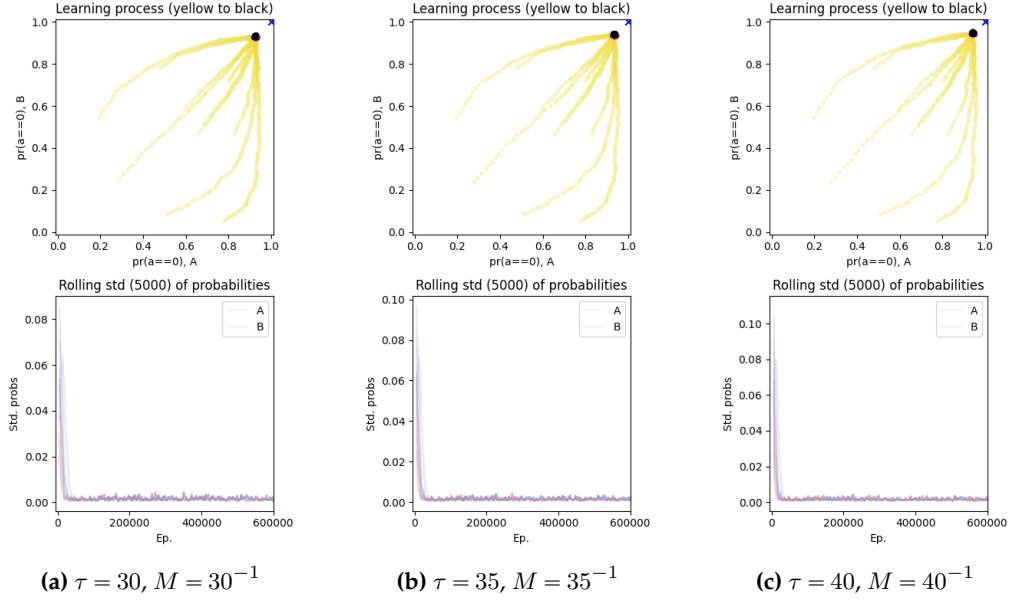


Figure 10: MBL-LC in self-play on the PD game with different values for τ (30, 35, 40) or M (30^{-1} , 35^{-1} , 40^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

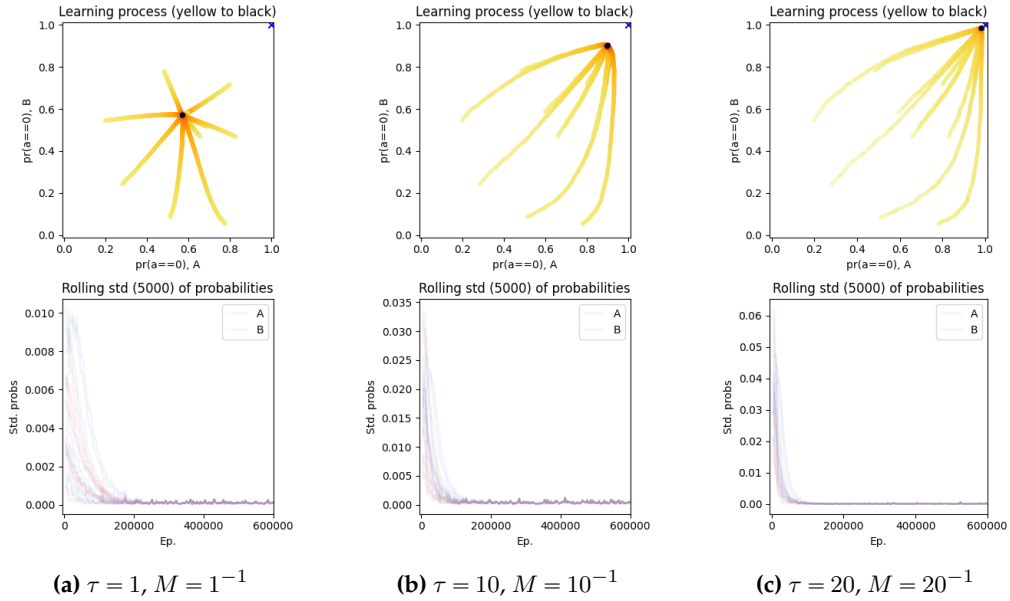


Figure 11: FAQ in self-play on the PD game with different values for τ (1, 10, 20) or M (1 , 10^{-1} , 20^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

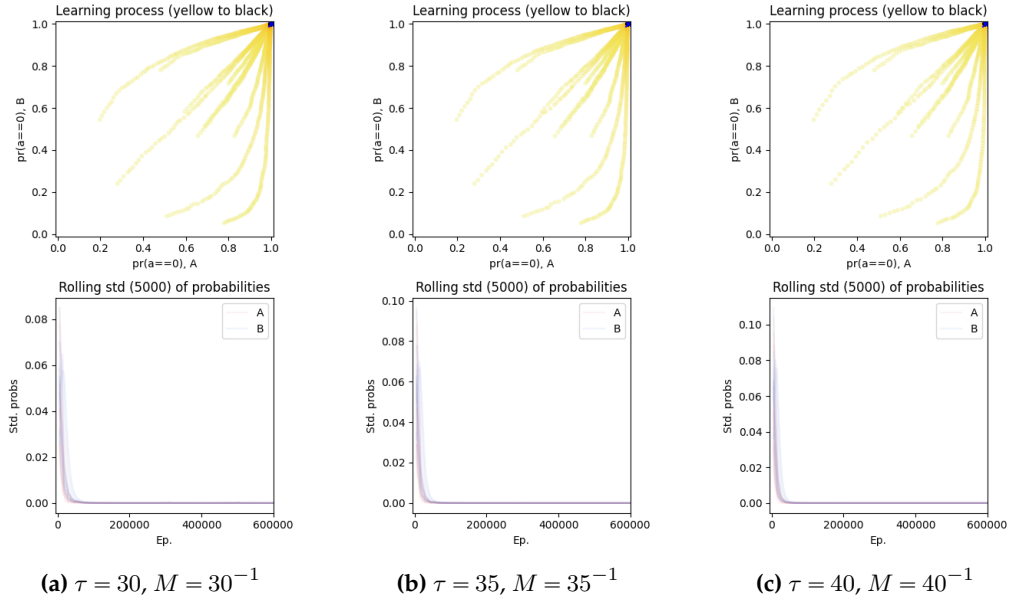


Figure 12: FAQ in self-play on the PD game with different values for τ (30, 35, 40) or M (30^{-1} , 35^{-1} , 40^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

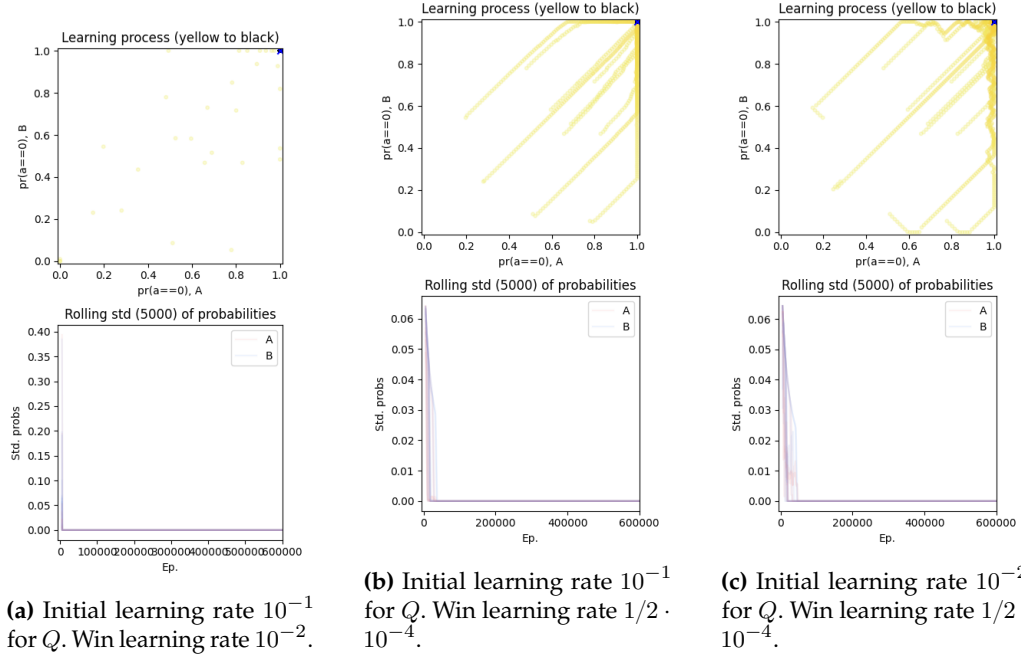


Figure 13: WoLF-PHC in self-play on the PD game with different learning schedules; for 10 different initialisations. Subgraph (a) has a high convergence speed such that only disconnected points can be seen. (See figure 7 for a detailed explanation of the graphs.)

(b) Zero-sum games

For two-player zero-sum games, we have preliminary results showing that the Nash equilibrium is an attracting mutation limit. While RD (and Cross learning) would not converge to interior equilibria (with Cross learning eventually approaching the boundary), RMD converges to the mutation equilibrium for every choice of mutation probabilities, $c \in \mathcal{D}^o$ and $M > 0$, and so does MBL-DPU. Stability is induced by the perturbative terms and their varying strengths have two effects which have to be weighed against each other. We demonstrate the general idea in the simple situation of the Matching Pennies (MP) game. Further, we illustrate the changing behaviour when we grow the strategy space by considering different versions of the Rock-Paper-Scissors game, RPS- n , with $n = 3, 5, 9$, where n denotes the number of strategies available to each player.

(i) Matching Pennies

The experimental results for the Matching Pennies game are based on the following payoff structure:

$$R_1 = \begin{pmatrix} 1 & -23/10 \\ -4/10 & 1 \end{pmatrix} \quad R_2 = \begin{pmatrix} -23/10 & 1 \\ 1 & -4/10 \end{pmatrix}$$

Nash equilibrium x^* at:

$$x_1^* = \begin{pmatrix} 14/47 & 33/47 \end{pmatrix}^T \quad x_2^* = \begin{pmatrix} 33/47 & 14/47 \end{pmatrix}^T$$

The MP game is a particularly simple case of a zero-sum game and hence provides an informative perspective on the basic characteristics of the different algorithms. In general, we see that the location of the mutation equilibrium depends on the mutation strength M , while convergence is slower for lower values of M creating a trade-off between these.

MBL-DPU and MBL-LC. Comparing MBL-DPU and MBL-LC, we see again that the LC-variant (figures 16, 17) approaches the mutation equilibrium more quickly than the DPU-variant (figures 14, 15). However, we see that the DPU-variant exhibits a much smaller variance, more precisely standard deviation, in the vicinity of the mutation equilibrium due to its slower change, with both variants roughly differing by a factor between 5 and 10 (for $M = 40^{-1}$). This illustrates the stronger effect that single larger payoffs have on the LC-variant, producing a larger variance near the mutation equilibrium.

FAQ-learning. For FAQ-learning (figures 18, 19) we see a similar behaviour as MBL-LC, however with a smaller variance near the equilibrium for weaker perturbation (figure 19). As with the MBL variants, FAQ exhibits slower convergence for weaker perturbation with larger variance near its (apparently asymptotically stable) equilibrium. However, we also observe that with FAQ, solutions can get trapped near the boundary (note the trapped solution in the upper left corner in figure 19), which we do not observe for the MBL variants and have proved not to be the case for MBL-DPU.

WoLF-PHC. Similar to the other algorithms, WoLF-PHC (figure 20) follows spiral-like trajectories towards a region close to the Nash equilibrium. It also shows a lower variance near the (apparently asymptotically stable) equilibrium. However, WoLF-PHC employs a learning rate schedule which reduces the learning rate over time and thus reduces variance.⁴ One should note that WoLF-PHC is considerably more complicated as it relies on a reliable way to estimate action-values as well as a long-term population average. It is clear that a player would require more resources for implementing WoLF-PHC than for the other algorithms.

⁴It would be possible to evaluate WoLF-PHC with a fixed learning rate or use a reduction schedule for the other algorithms. However, the former would be a deviation from the canonical formulation of WoLF-PHC while the latter would not be based on a principled approach. Hence, this heterogeneous situation is an appropriate base scenario.

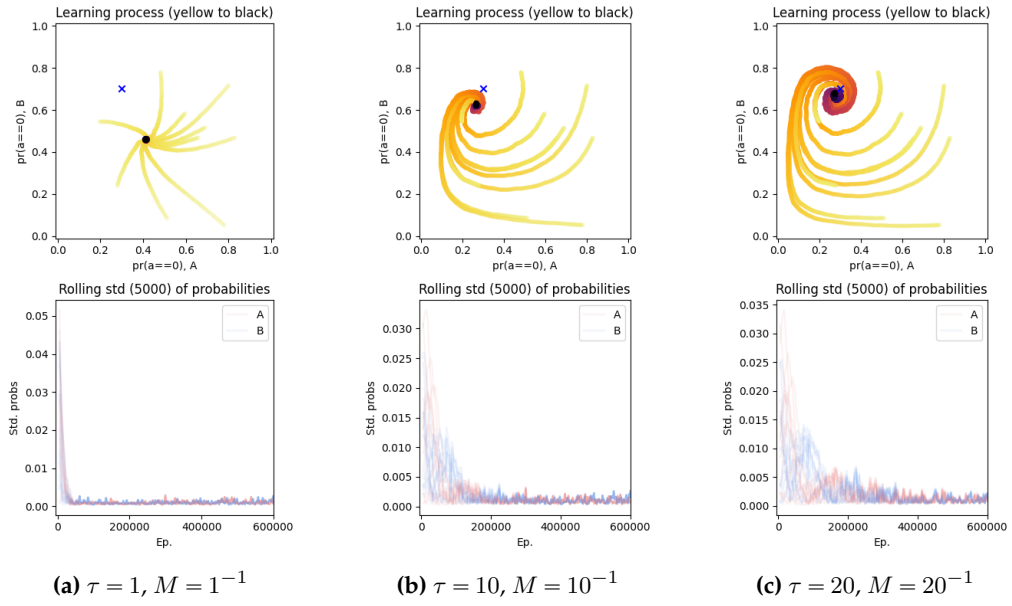


Figure 14: MBL-DPU in self-play on the MP game with different values for τ (1, 10, 20) or M (1, 10^{-1} , 20^{-1}) equivalently; $\theta = 10^{-4}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

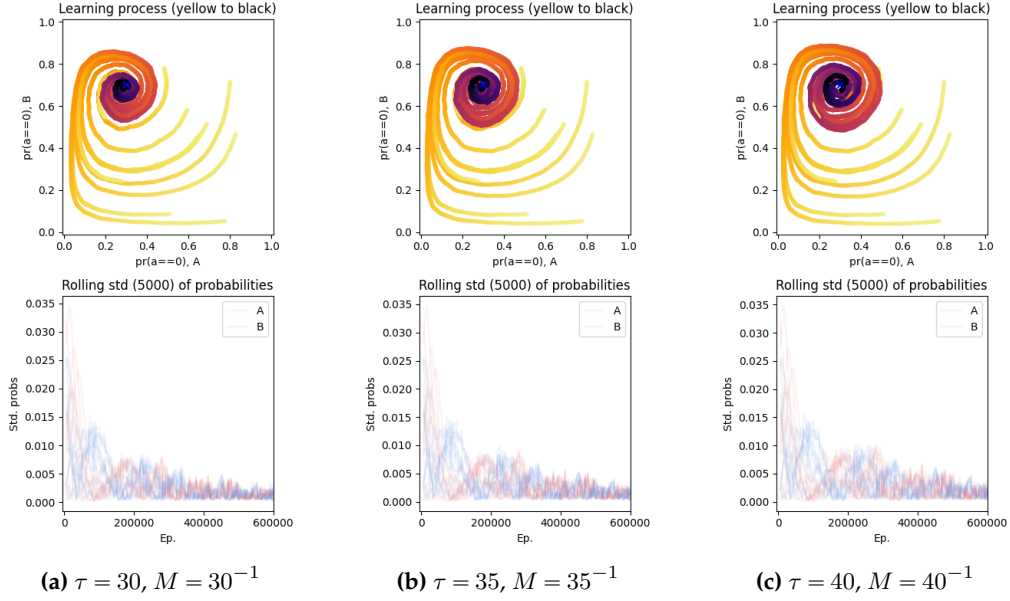


Figure 15: MBL-DPU in self-play on the MP game with different values for τ (30, 35, 40) or M (30^{-1} , 35^{-1} , 40^{-1}) equivalently; $\theta = 10^{-4}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

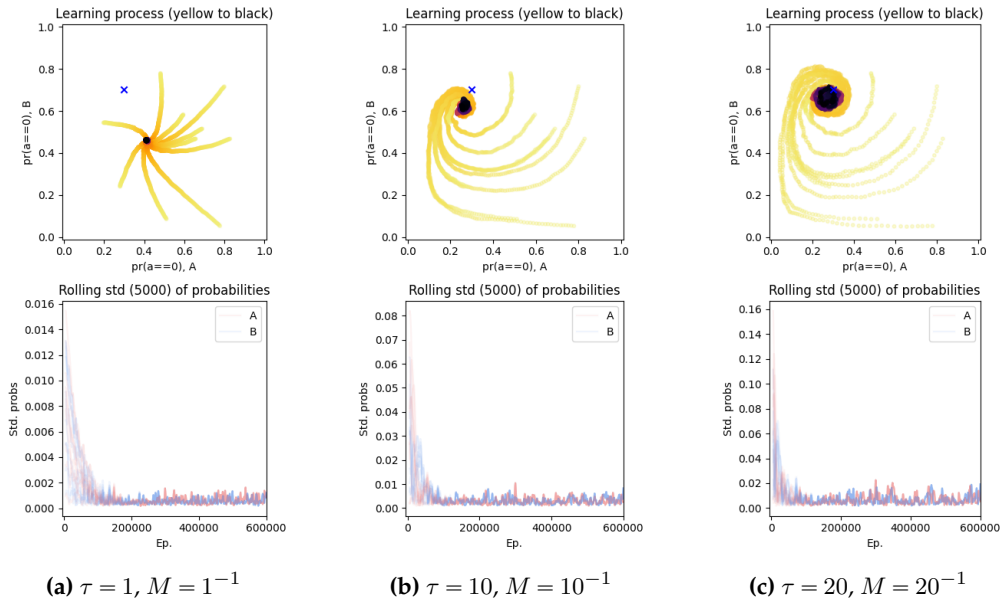


Figure 16: MBL-LC in self-play on the MP game with different values for τ (1, 10, 20) or M (1 , 10^{-1} , 20^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

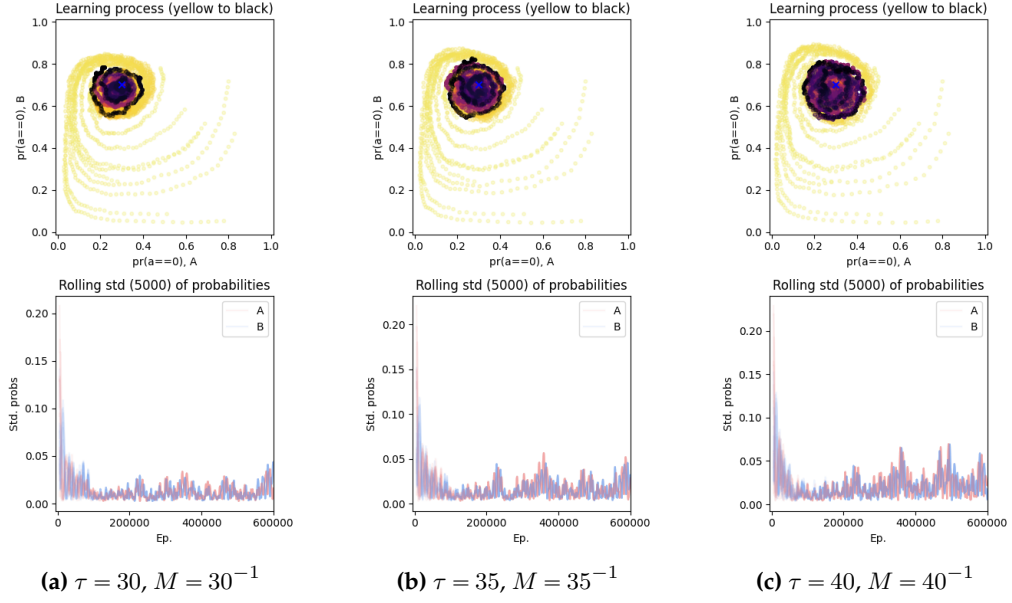


Figure 17: MBL-LC in self-play on the MP game with different values for τ (30, 35, 40) or M (30^{-1} , 35^{-1} , 40^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

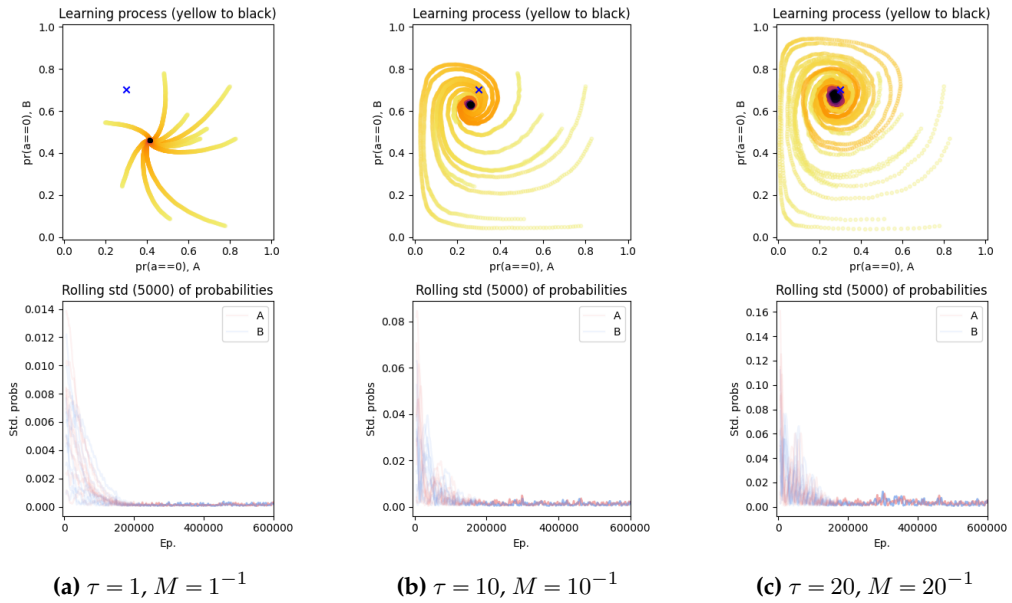


Figure 18: FAQ in self-play on the MP game with different values for τ (1, 10, 20) or M (1 , 10^{-1} , 20^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

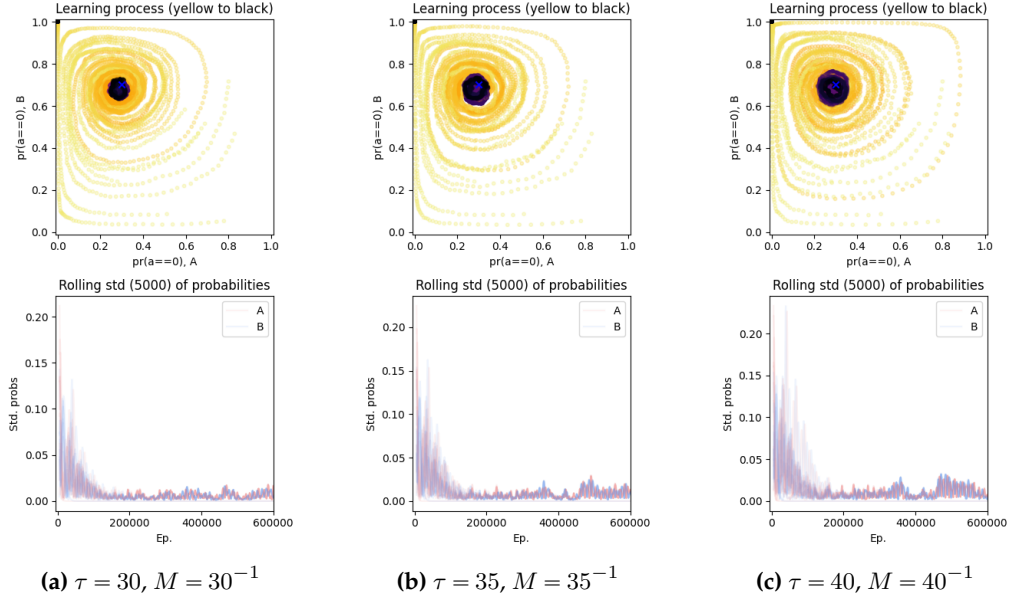


Figure 19: FAQ in self-play on the MP game with different values for τ (30, 35, 40) or M (30^{-1} , 35^{-1} , 40^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

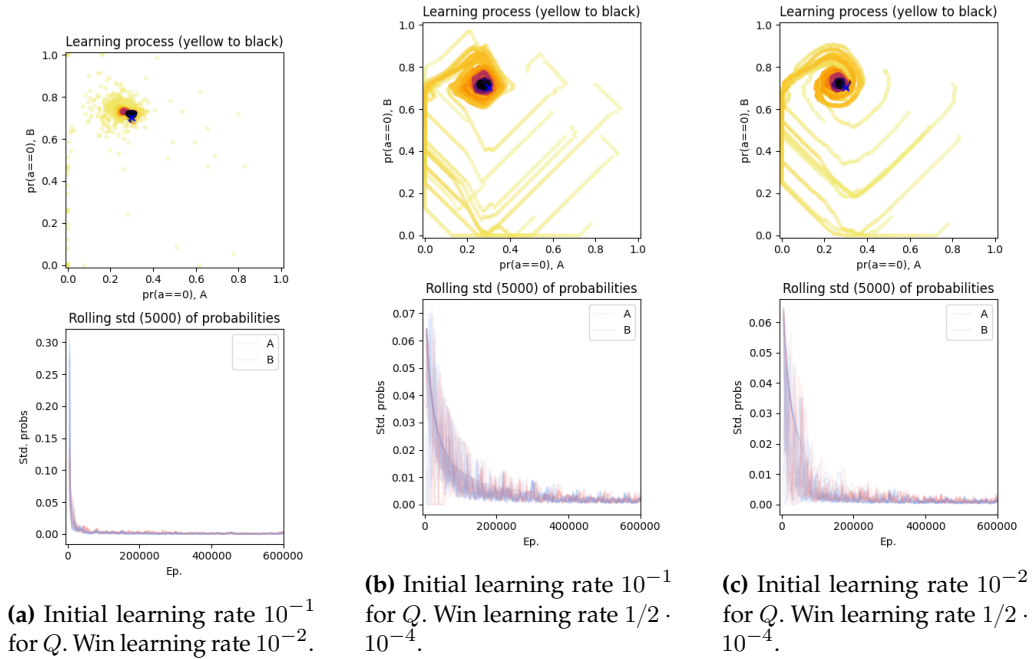


Figure 20: WoLF-PHC in self-play on the MP game with different learning schedules; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

718 (ii) Zero-sum games with larger action spaces

719 The experimental results for the RPS- n games are based on the following payoff structures.

RPS-3.

$$R_1 = \begin{pmatrix} 0 & -2 & 3 \\ 2 & 0 & -2 \\ -1 & 2 & 0 \end{pmatrix} \quad R_2 = -R_1$$

720 Nash equilibrium x^* at:

$$x_1^* = \left(\frac{2}{7} \quad \frac{11}{35} \quad \frac{2}{5} \right)^T \quad x_2^* = \left(\frac{2}{5} \quad \frac{11}{35} \quad \frac{2}{7} \right)^T$$

RPS-5.

$$R_1 = \begin{pmatrix} 0 & 4 & -2 & 2 & -2 \\ -4 & 0 & 2 & -1 & 1 \\ 2 & -4 & 0 & 4 & -1 \\ -4 & 1 & -4 & 0 & 2 \\ 2 & -1 & 1 & -2 & 0 \end{pmatrix} \quad R_2 = -R_1$$

721 Nash equilibrium x^* at:

$$x_1^* = \left(\frac{11}{61} \quad \frac{510}{2989} \quad \frac{8}{61} \quad \frac{50}{427} \quad \frac{1198}{2989} \right)^T$$

$$x_2^* = \left(\frac{1}{7} \quad \frac{68}{427} \quad \frac{6}{49} \quad \frac{502}{2989} \quad \frac{174}{427} \right)^T$$

RPS-9.

$$R_1 = \begin{pmatrix} 0 & 2 & 1 & 3 & 1 & -1 & -1 & -2 & -1 \\ -1 & 0 & 1 & 3 & 1 & 1 & -1 & -2 & -1 \\ -1 & -2 & 0 & 3 & 1 & 1 & 1 & -2 & -1 \\ -2 & -4 & -2 & 0 & 2 & 2 & 2 & 4 & -2 \\ -1 & -2 & -1 & -3 & 0 & 1 & 1 & 2 & 1 \\ 1 & -2 & -1 & -3 & -1 & 0 & 1 & 2 & 1 \\ 2 & 4 & -2 & -6 & -2 & -2 & 0 & 4 & 2 \\ 1 & 2 & 1 & -3 & -1 & -1 & -1 & 0 & 1 \\ 1 & 2 & 1 & 3 & -1 & -1 & -1 & -2 & 0 \end{pmatrix} \quad R_2 = -R_1$$

722 Nash equilibrium x^* at:

$$x_1^* = \left(\frac{1}{8} \quad \frac{1}{8} \quad \frac{1}{8} \quad \frac{1}{16} \quad \frac{1}{8} \quad \frac{1}{8} \quad \frac{1}{16} \quad \frac{1}{8} \quad \frac{1}{8} \right)^T$$

$$x_2^* = \left(\frac{3}{22} \quad \frac{3}{44} \quad \frac{3}{22} \quad \frac{1}{22} \quad \frac{3}{22} \quad \frac{3}{22} \quad \frac{3}{22} \quad \frac{3}{22} \quad \frac{3}{44} \quad \frac{3}{22} \right)^T$$

723 While MP is an informative illustration of the different behaviours, MP reduces to a planar
 724 dynamical system, which does not allow many complex behaviours, as exemplified by the Poincaré-
 725 Bendixson theorem, e.g., [38, theorem 7.16] holding for planar systems. Hence, higher-dimensional
 726 zero-sum games allow a further understanding of the differences between the algorithms and shed
 727 light on the effect of larger state spaces while preserving the neutral stability of interior equilibria.
 728 We consider here the Rock-Paper-Scissors game of different sizes (3, 5 and 9 actions).

729 **MBL-DPU and MBL-LC.** In RPS-3, MBL-DPU (figures 21, 22) shows a similar behaviour to
 730 MP with a marked dependence of the behaviour of the variance on the value of M . In contrast,
 731 MBL-LC (figures 23, 24) shows a much quicker convergence, with the variance dropping after

similar numbers of episodes (around 10^5) for all values of M . As with MBL-DPU, the residual variance increases with weaker mutation. This is in accordance with the neutral stability of the Nash equilibrium, allowing for larger fluctuations.

In RPS-5, both MBL variants (figures 28, 29 for MBL-DPU and figures 30, 31 for MBL-LC) show behaviours similar to their RPS-3 counterparts. In RPS-9, MBL-DPU (figures 35, 36) again shows similar behaviour, with slower convergence compared to its RPS-3 and RPS-5 counterparts. Interestingly, MBL-LC (figures 37, 38) seems to have two distinct regions to which trajectories evolve, suggesting a potentially stronger sensitivity to the choice of θ .

FAQ-learning. Like for MP, we see a quicker convergence for FAQ in RPS-3 (figures 25, 26) compared to the MBL variants, but with trajectories similar to those of MBL-LC when considering low values of M , in which case the replicator dynamics makes a stronger contribution to the trajectories. Similar to MBL-LC, but already in RPS-5, FAQ shows two distinct regions to which trajectories evolve when perturbation is weak (figures 32, 33), whereas the former does not show such a split for RPS-5. In RPS-9, FAQ shows such a split for stronger perturbation levels already and shows even three distinct such regions for weaker perturbation (figures 39, 40).

WoLF-PHC. For WoLF-PHC, we see a still quicker convergence in RPS-3 (figure 27) than for the other algorithms, similar to the MP case. However, the behaviour is much less clear in RPS-5 (figure 34). Here, trajectories do not consistently approach a specific region. It is possible that the reduction schedules for the learning rates, which force each trajectory to converge, lead to trajectories stalling prematurely. This becomes even more pronounced in RPS-9 (figure 41), where WoLF-PHC seems to initially move away from the Nash equilibrium and to get stuck along the boundaries of $\mathcal{D}$.

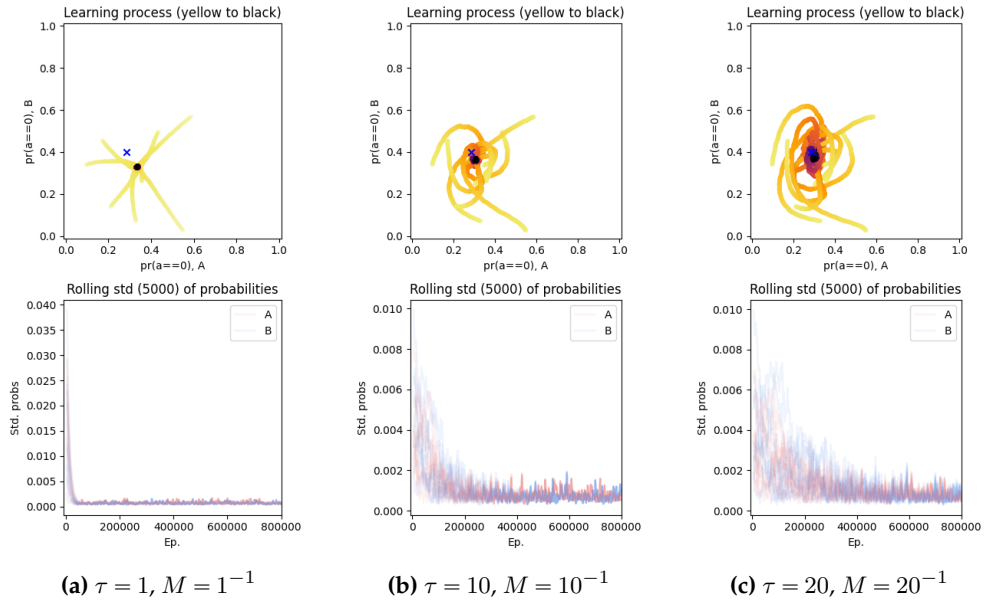


Figure 21: MBL-DPU in self-play on the RPS-3 game with different values for τ (1, 10, 20) or M ($1, 10^{-1}, 20^{-1}$) equivalently; $\theta = 10^{-4}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

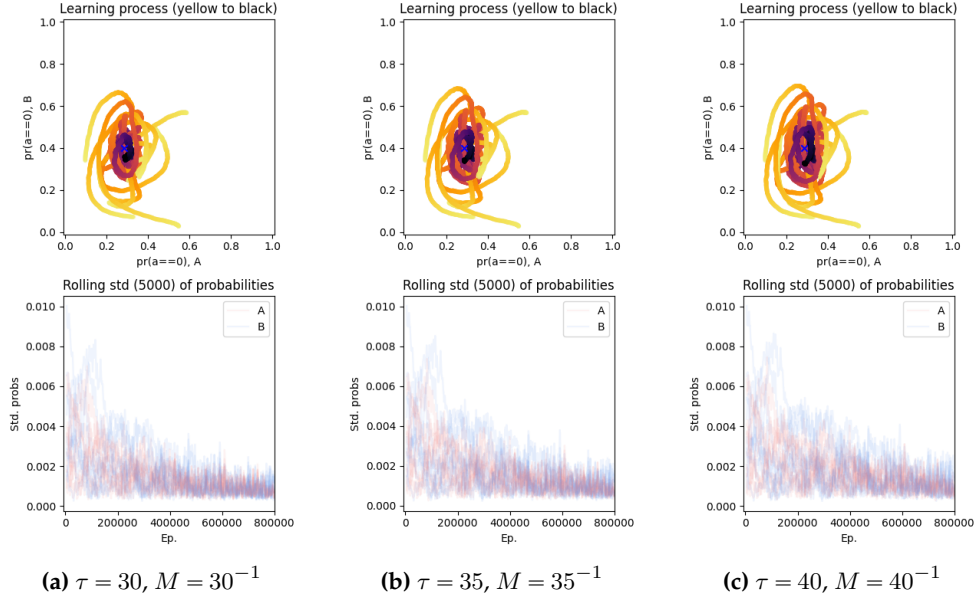


Figure 22: MBL-DPU in self-play on the RPS-3 game with different values for τ (30, 35, 40) or M (30^{-1} , 35^{-1} , 40^{-1}) equivalently; $\theta = 10^{-4}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

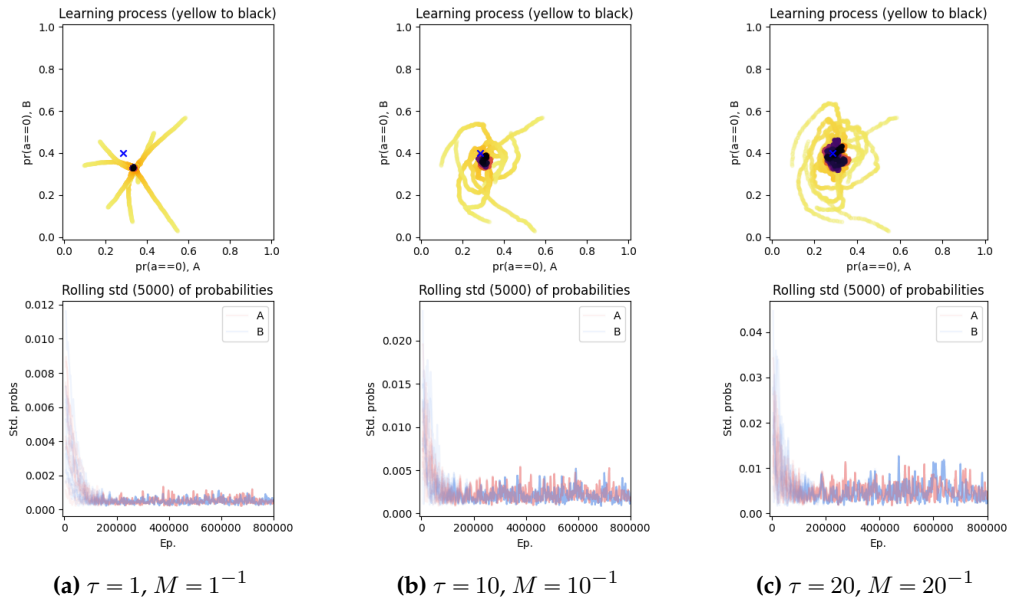


Figure 23: MBL-LC in self-play on the RPS-3 game with different values for τ (1, 10, 20) or M (1 , 10^{-1} , 20^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

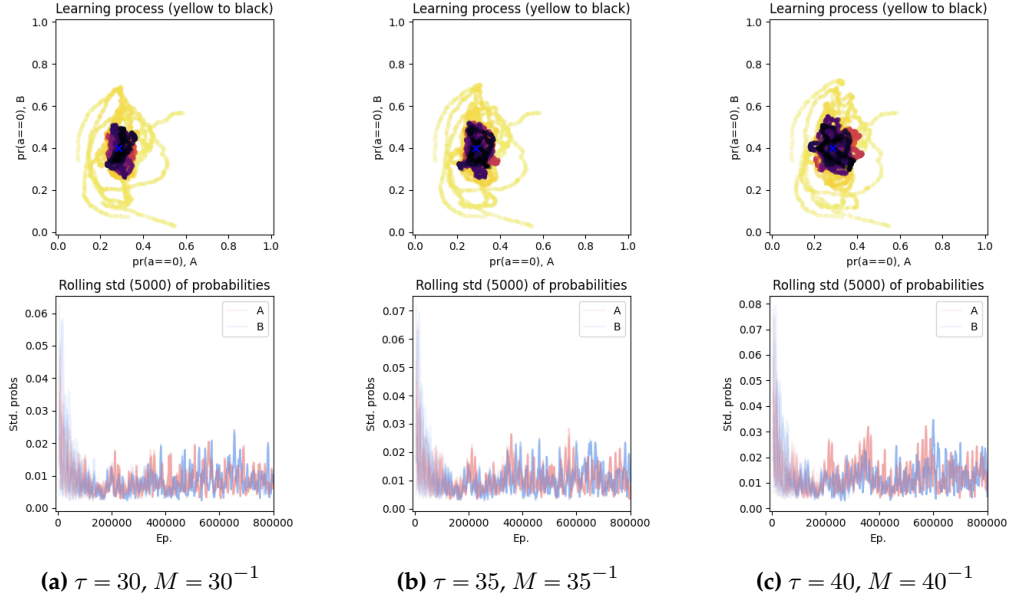


Figure 24: MBL-LC in self-play on the RPS-3 game with different values for τ (30, 35, 40) or M (30^{-1} , 35^{-1} , 40^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

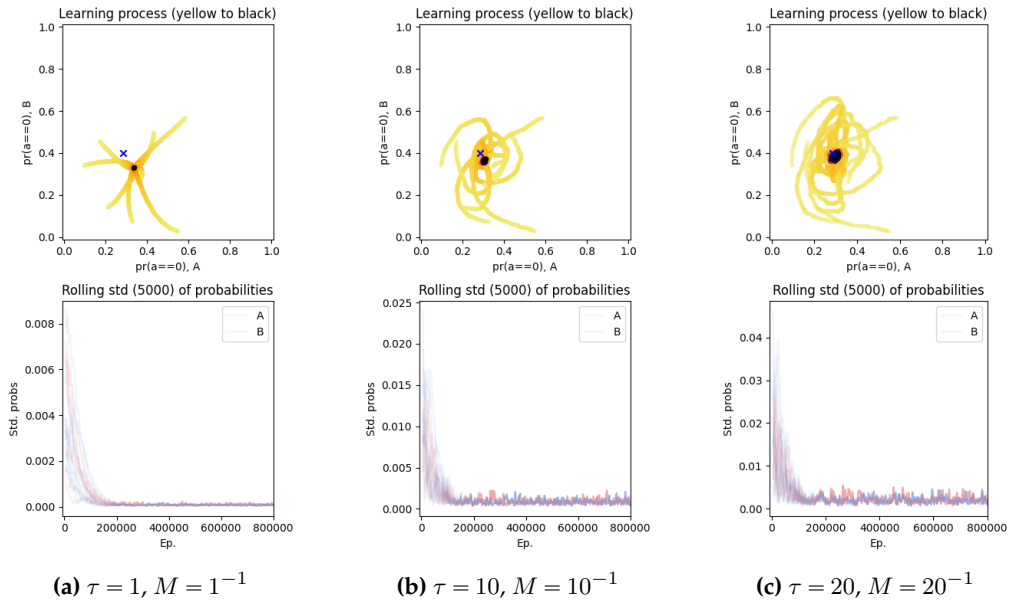


Figure 25: FAQ in self-play on the RPS-3 game with different values for τ (1, 10, 20) or M (1 , 10^{-1} , 20^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

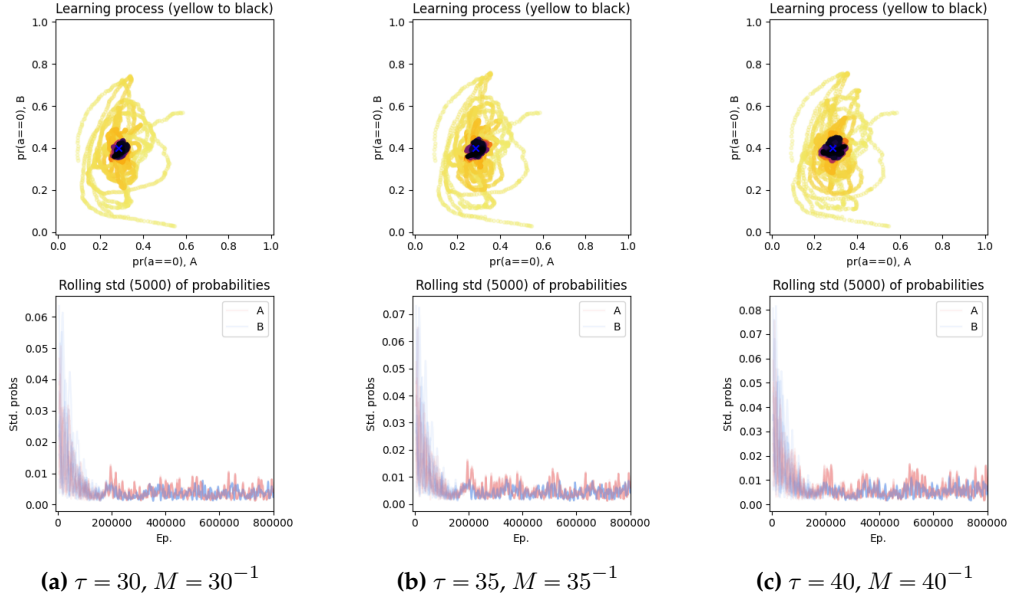


Figure 26: FAQ in self-play on the RPS-3 game with different values for τ (30, 35, 40) or M (30^{-1} , 35^{-1} , 40^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

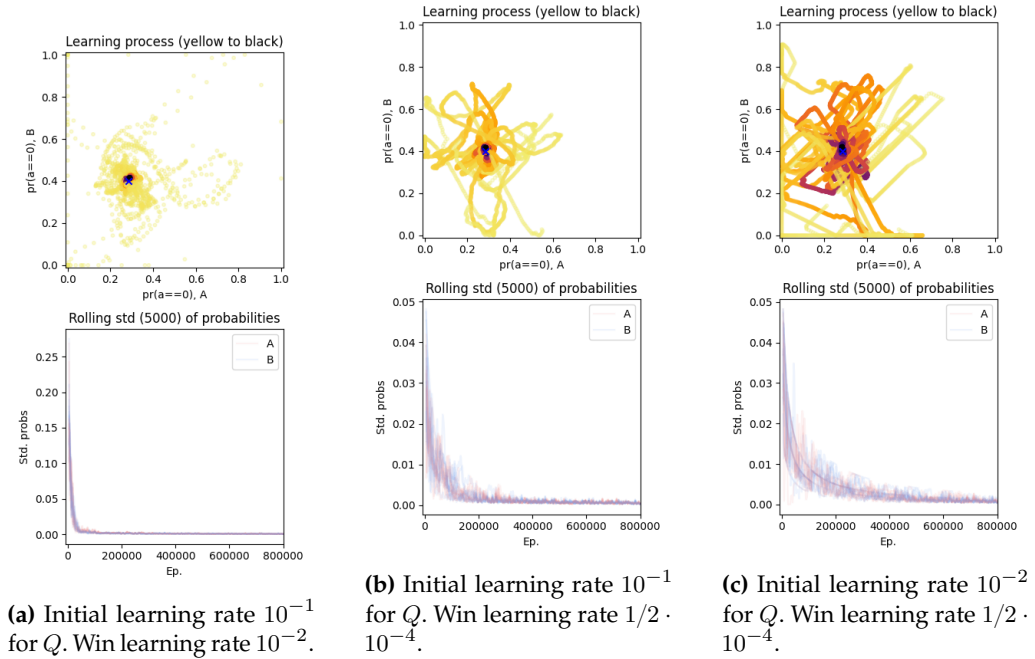


Figure 27: WoLF-PHC in self-play on the RPS-3 game with different learning schedules; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

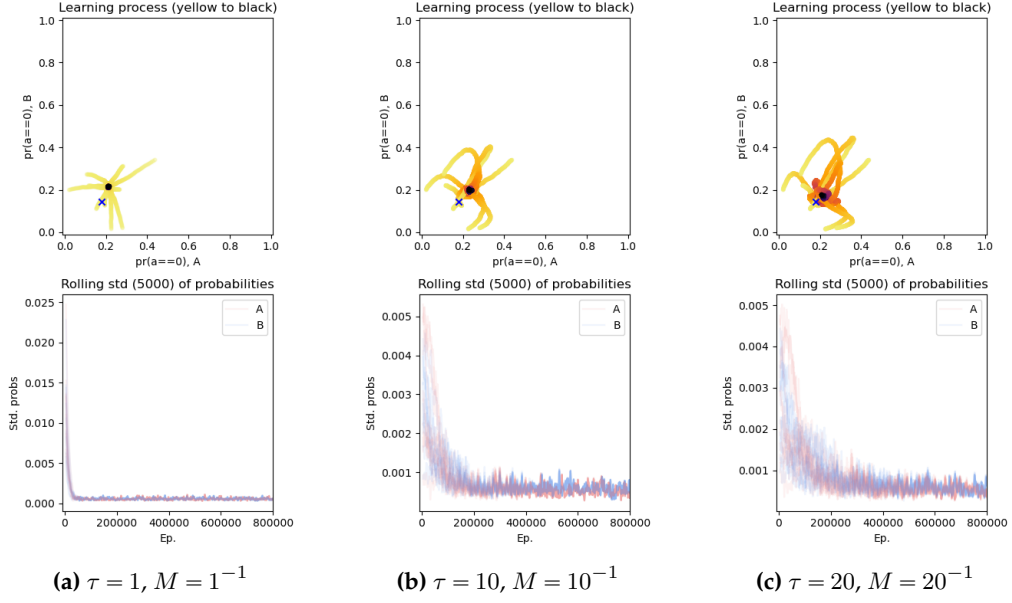


Figure 28: MBL-DPU in self-play on the RPS-5 game with different values for τ (1, 10, 20) or M ($1, 10^{-1}, 20^{-1}$) equivalently; $\theta = 10^{-4}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

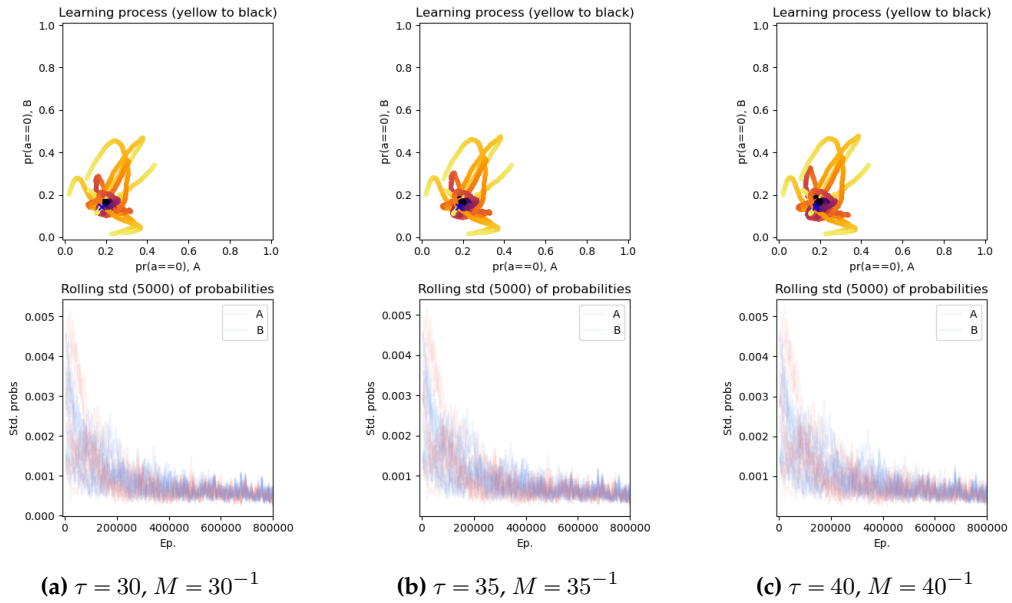


Figure 29: MBL-DPU in self-play on the RPS-5 game with different values for τ (30, 35, 40) or M ($30^{-1}, 35^{-1}, 40^{-1}$) equivalently; $\theta = 10^{-4}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

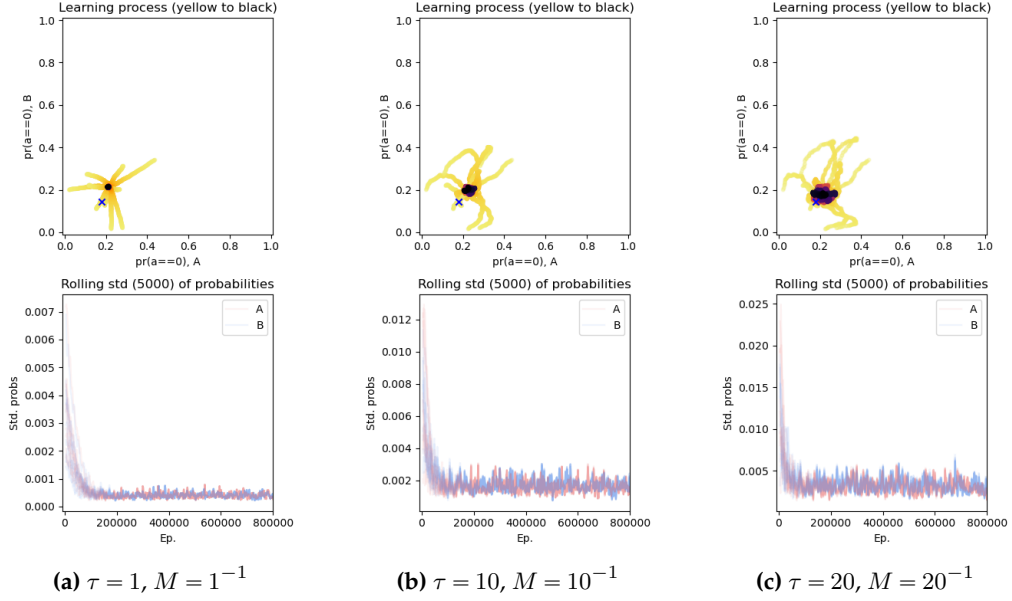


Figure 30: MBL-LC in self-play on the RPS-5 game with different values for τ (1, 10, 20) or M ($1, 10^{-1}, 20^{-1}$) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

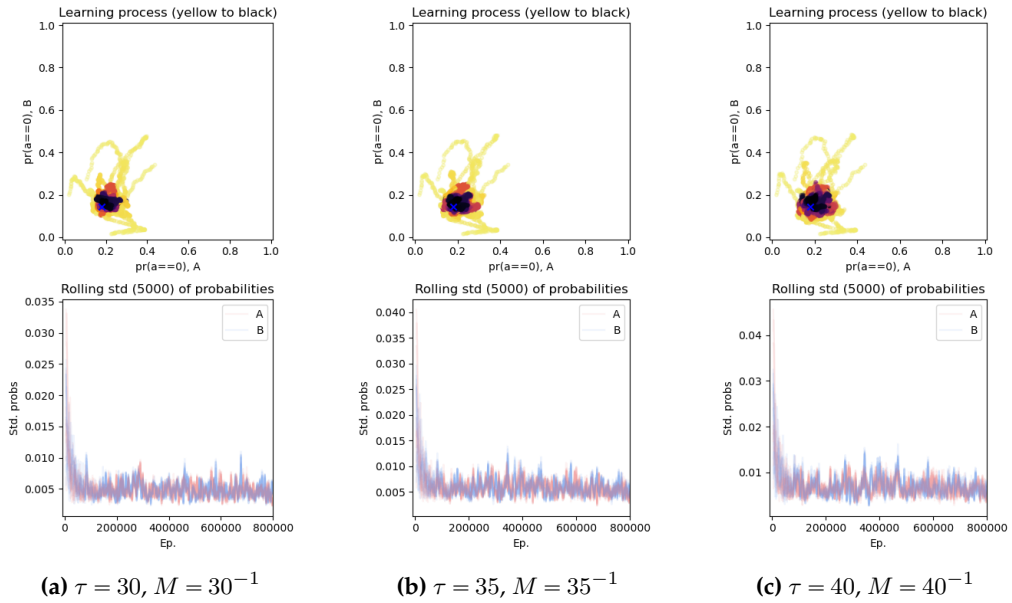


Figure 31: MBL-LC in self-play on the RPS-5 game with different values for τ (30, 35, 40) or M ($30^{-1}, 35^{-1}, 40^{-1}$) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

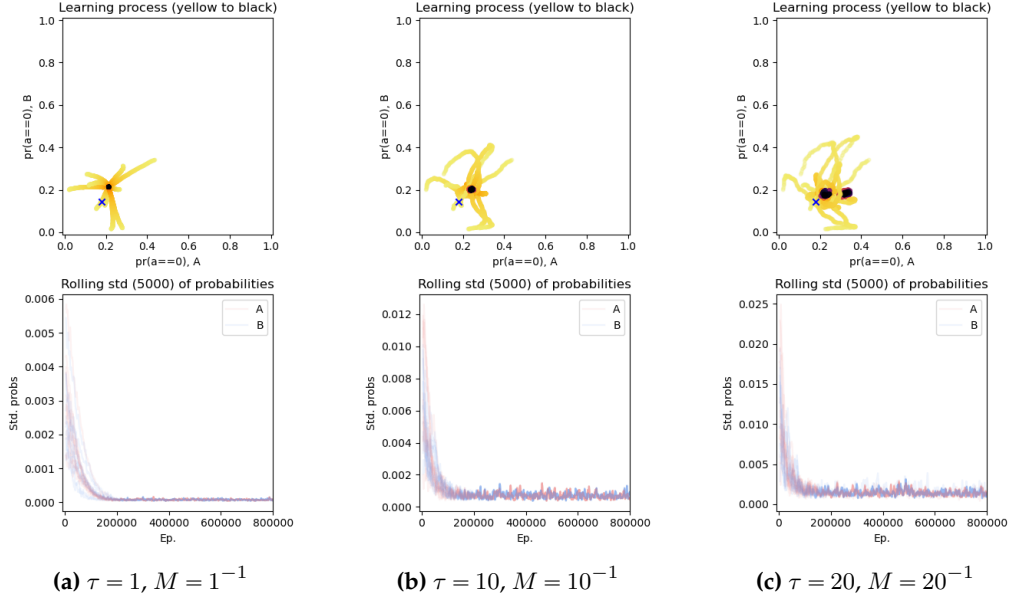


Figure 32: FAQ in self-play on the RPS-5 game with different values for τ (1, 10, 20) or M ($1, 10^{-1}, 20^{-1}$) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

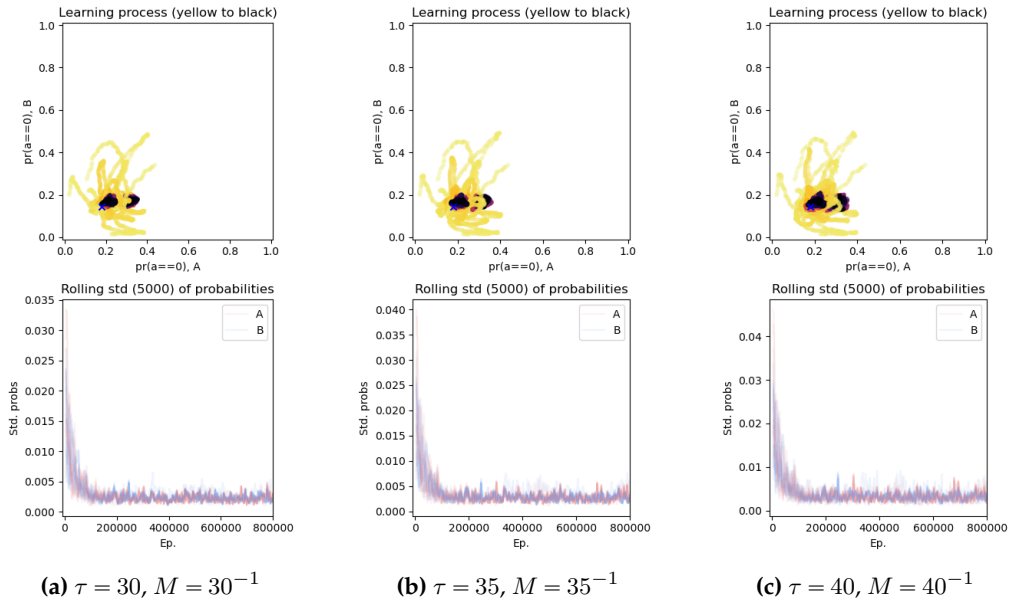


Figure 33: FAQ in self-play on the RPS-5 game with different values for τ (30, 35, 40) or M ($30^{-1}, 35^{-1}, 40^{-1}$) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

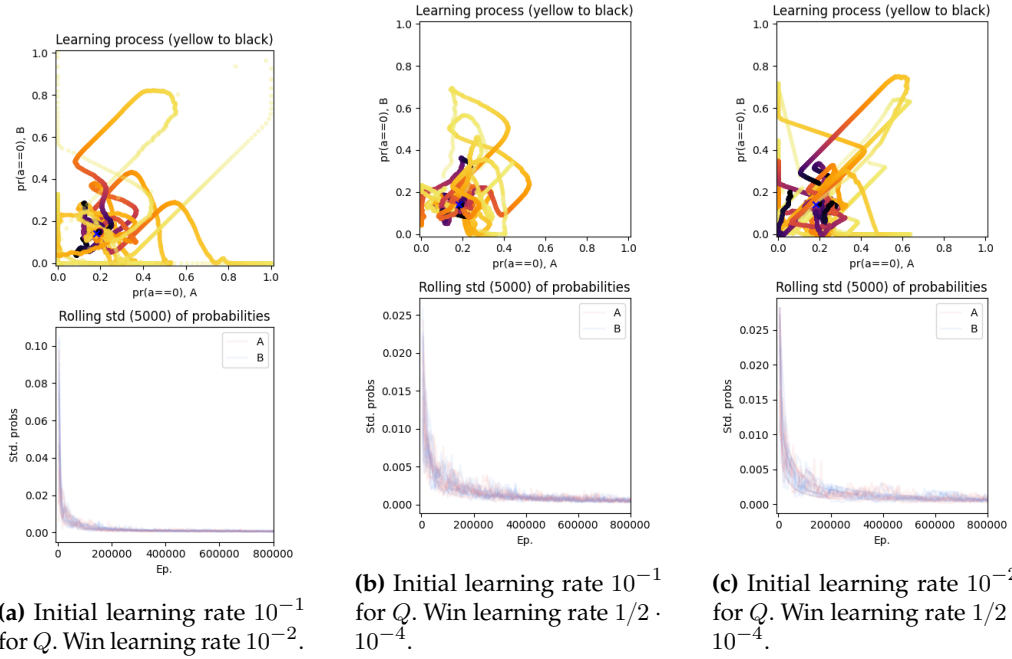


Figure 34: WoLF-PHC in self-play on the RPS-5 game with different learning schedules; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

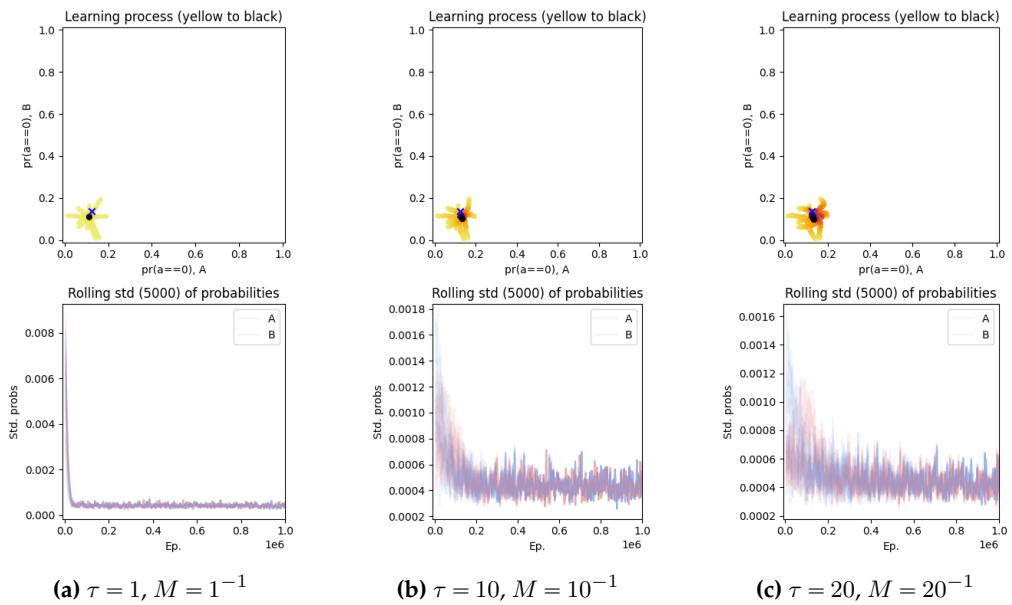


Figure 35: MBL-DPU in self-play on the RPS-9 game with different values for τ (1, 10, 20) or M ($1, 10^{-1}, 20^{-1}$) equivalently; $\theta = 10^{-4}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

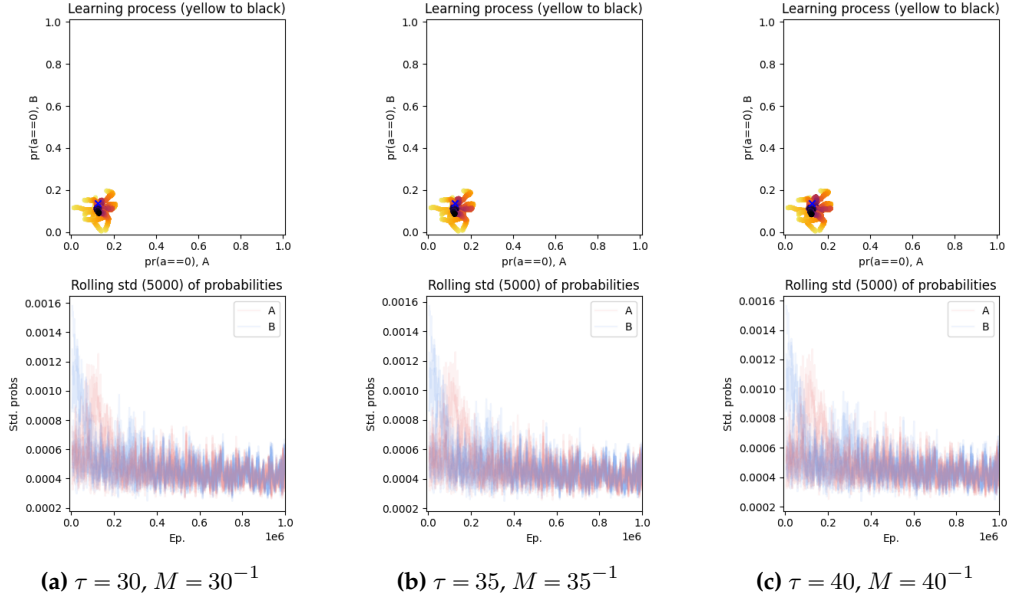


Figure 36: MBL-DPU in self-play on the RPS-9 game with different values for τ (30, 35, 40) or M (30^{-1} , 35^{-1} , 40^{-1}) equivalently; $\theta = 10^{-4}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

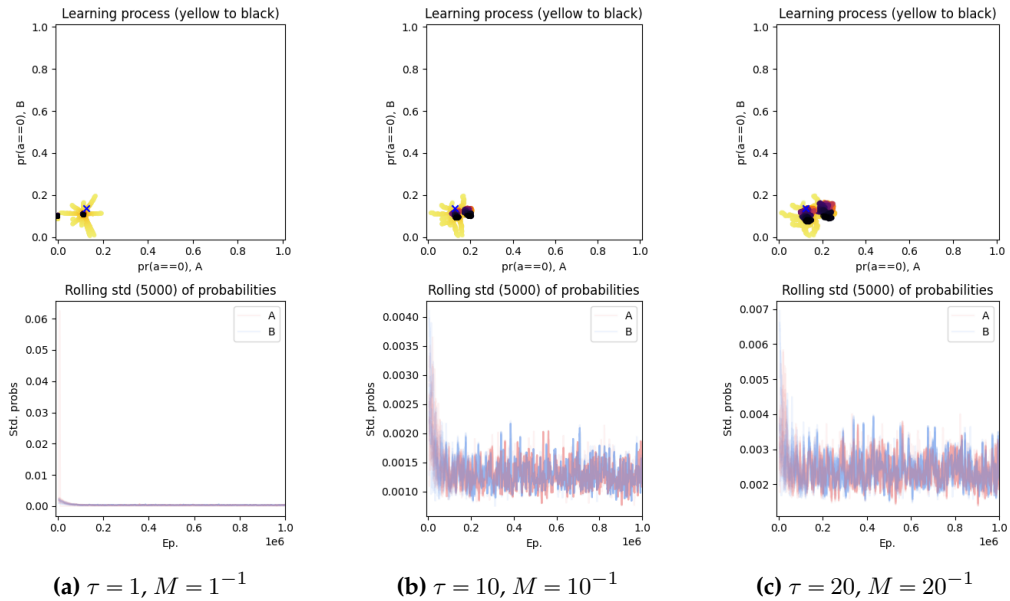


Figure 37: MBL-LC in self-play on the RPS-9 game with different values for τ (1, 10, 20) or M (1 , 10^{-1} , 20^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

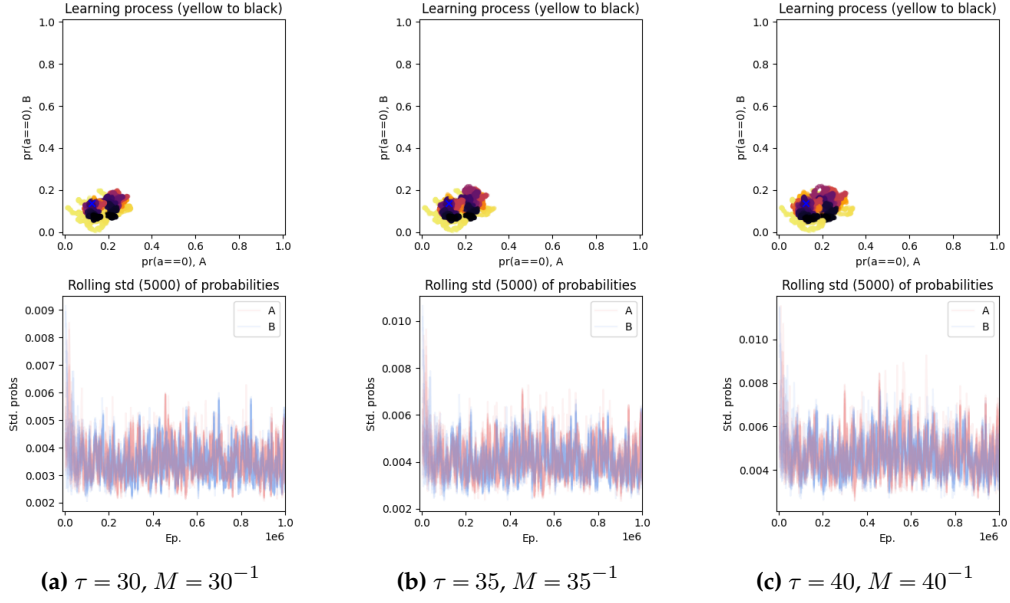


Figure 38: MBL-LC in self-play on the RPS-9 game with different values for τ (30, 35, 40) or M (30^{-1} , 35^{-1} , 40^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

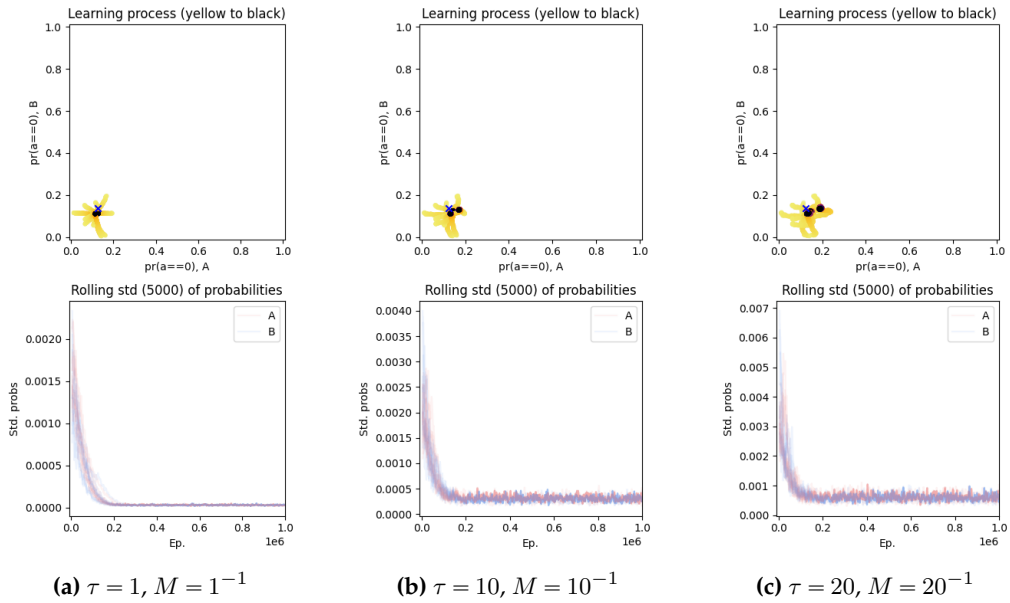


Figure 39: FAQ in self-play on the RPS-9 game with different values for τ (1, 10, 20) or M (1 , 10^{-1} , 20^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

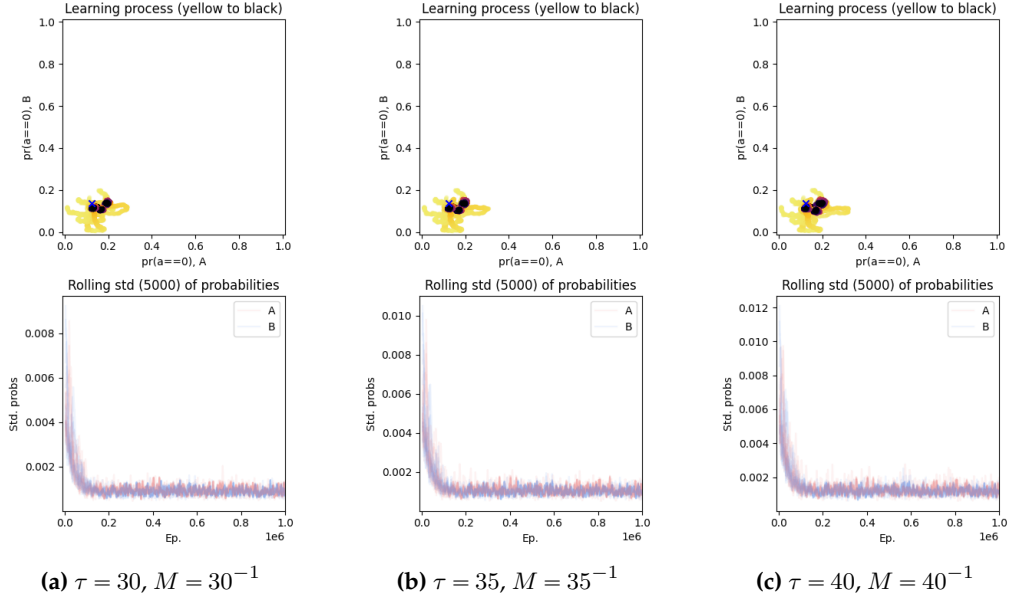


Figure 40: FAQ in self-play on the RPS-9 game with different values for τ (30, 35, 40) or M (30^{-1} , 35^{-1} , 40^{-1}) equivalently; $\theta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

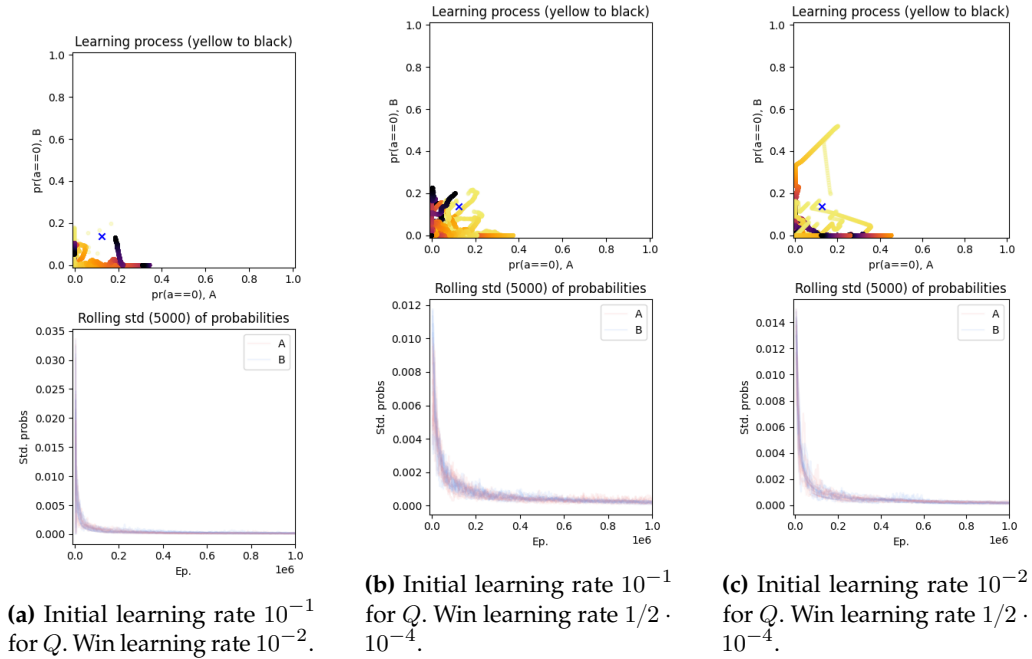


Figure 41: WoLF-PHC in self-play on the RPS-9 game with different learning schedules; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

(c) Three-player Matching Pennies

Further, we consider the behaviour of the MBL variants in comparison to FAQ learning and WoLF-PHC in a three-player Matching Pennies (3MP) game introduced in [37], with payoffs as given in table 1. The similarity to the standard MP game becomes clear when one considers that the payoff structure reflects the following idea: The first player wants to match the second player's action. The second player wants to match the third player's action. However, the third player does not want to match the first player's action. The unique Nash equilibrium for 3MP is located at the centre of $\mathcal{D}$. Note that, as initially proposed, 3MP is not a zero-sum game.

	H	T		H	T
H	(1, 1, -1)	(-1, -1, -1)	H	(1, -1, 1)	(-1, 1, 1)
T	(-1, 1, 1)	(1, -1, 1)	T	(-1, -1, -1)	(1, 1, -1)

(a) Payoffs when the third player chooses 'H'. (b) Payoffs when the third player chooses 'T'.

Table 1: Payoff tuples for the three-player Matching Pennies (3MP) game with the first player's action determining the row, the second player's action the column, and the third player's action the table.

In 3MP, both MBL variants (figures 42, 43) show apparently asymptotically stable periodic limit behaviours, which approach the boundary of $\mathcal{D}$ as mutation diminishes. We further see a very similar behaviour for FAQ (figure 44) with τ^{-1} showing an analogous effect to M in MBL, quite similar to the two-player settings. Likewise, WoLF-PHC (figure 45) exhibits apparently asymptotically stable trajectories, at least in the projection onto the first actions of the first two players. Again, WoLF-PHC shows a reduction of variance over time, presumably due to diminishing learning rates. In [6], the authors show that WoLF-PHC converges to the Nash equilibrium when $\delta_l/\delta_w = 3$ (as opposed to $\delta_l/\delta_w = 2$). Since there is no established ODE approximation of WoLF-PHC that we are aware of, the reasons for this remain unclear. One should also note that we have made sure that the Nash equilibrium is not located at the centre of $\mathcal{D}$ in the two-player games because the perturbation term in FAQ has its equilibrium there and convergence might easily have been coincidental. For 3MP, we have not made any such adaptations and some behaviours might change when the Nash equilibrium is moved away from the centre.

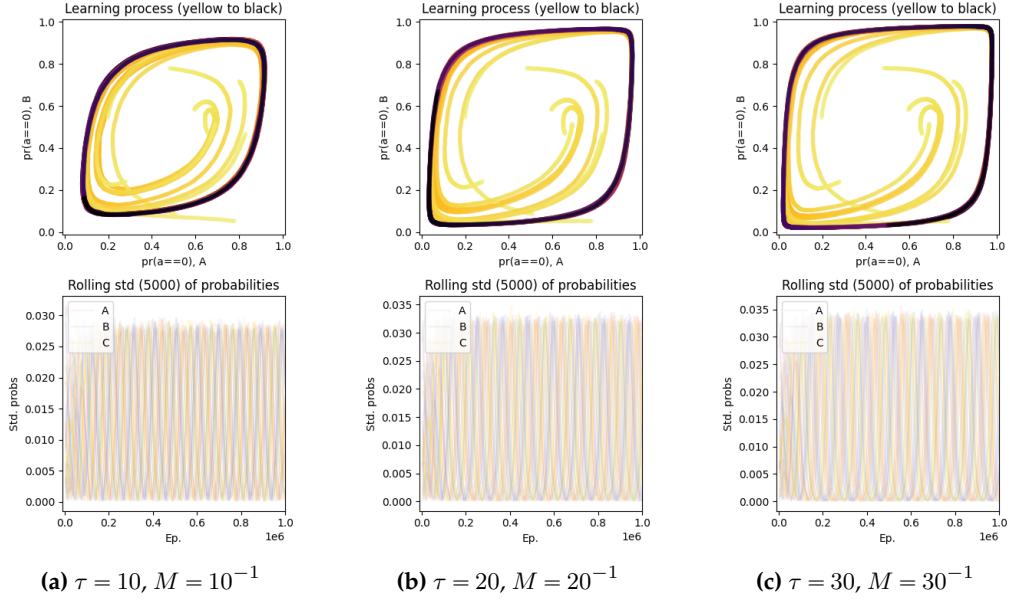


Figure 42: MBL-DPU in self-play on the 3MP game with different values for τ (10, 20, 30) or M (10^{-1} , 20^{-1} , 30^{-1}) equivalently; $\theta = 10^{-4}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

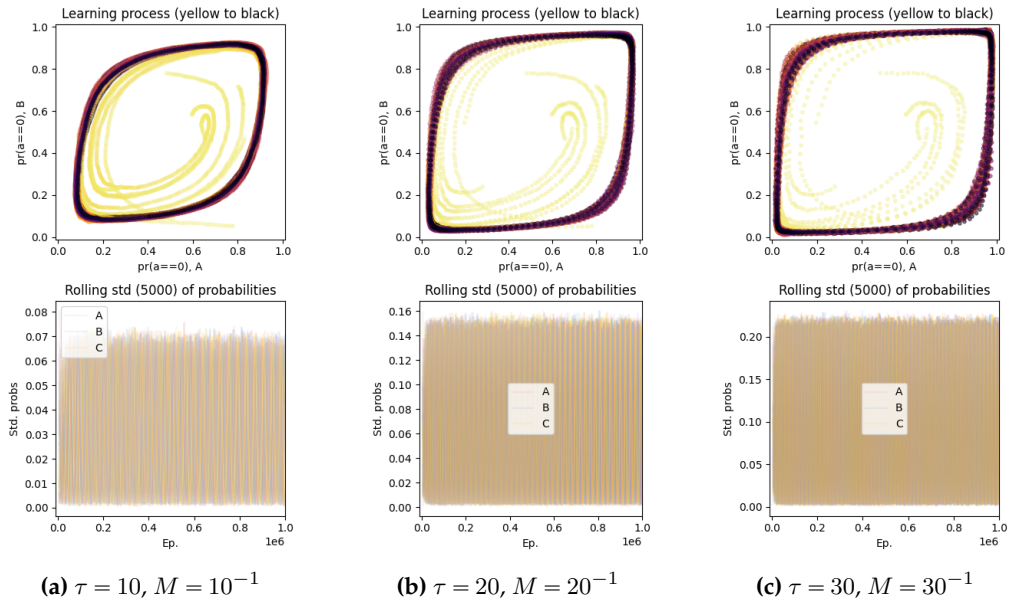


Figure 43: MBL-LC in self-play on the 3MP game with different values for τ (10, 20, 30) or M (10^{-1} , 20^{-1} , 30^{-1}) equivalently; $\theta = 10^{-4}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

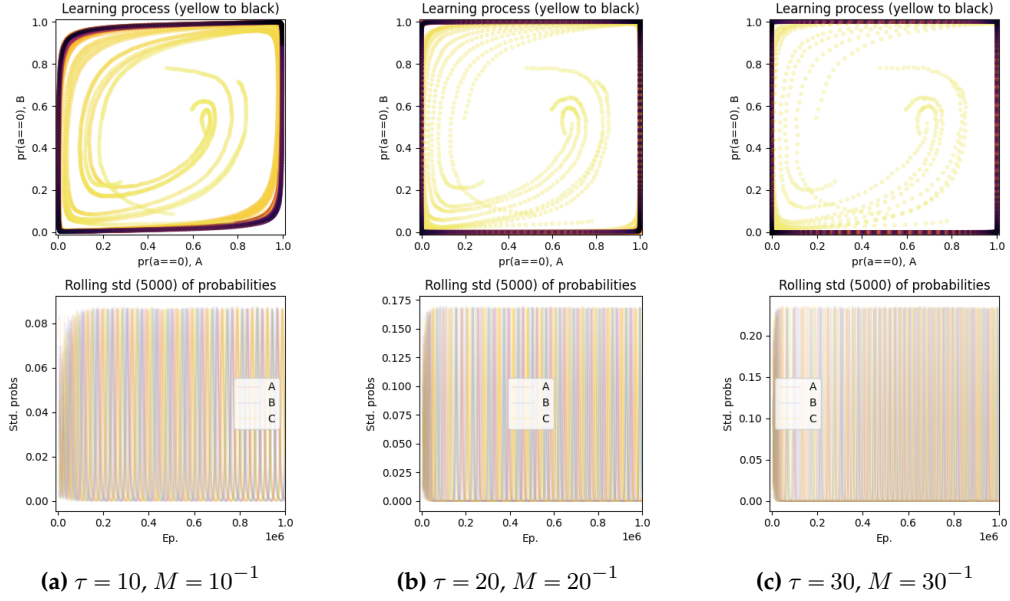


Figure 44: FAQ in self-play on the 3MP game with different values for τ (10, 20, 30) or M (10^{-1} , 20^{-1} , 30^{-1}) equivalently; $\theta = 10^{-4}$; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)

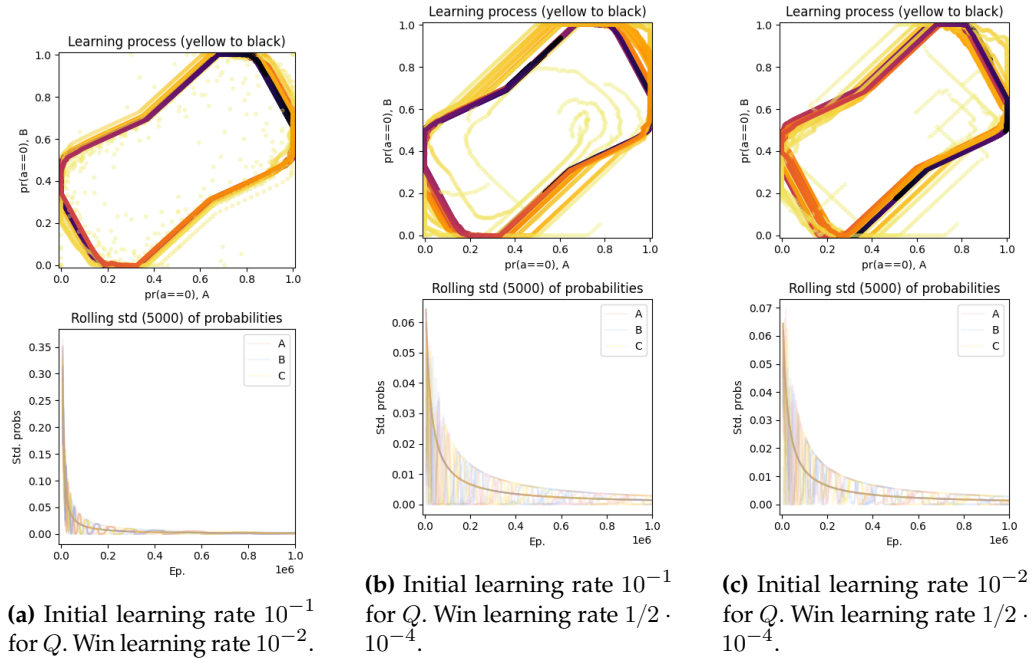


Figure 45: WoLF-PHC in self-play on the 3MP game with different learning schedules; for 10 different initialisations. (See figure 7 for a detailed explanation of the graphs.)