Stochastic stability of a recency weighted sampling dynamic

Alexander Aurell[1] and Gustav Karreskog[2]

[1]Princeton University, ORFE
[2]Stockholm School of Economics

September 29, 2020

Abstract

It is common to model learning in games so that either a deterministic process or a finite state Markov chain describes the evolution of play. Such processes can however produce undesired outputs, where the players' behavior is heavily influenced by the modeling. In simulations we see how the assumptions in Young (1993), a well-studied model for stochastic stability, lead to unexpected behavior in games without strict equilibria, such as Matching Pennies. The behavior should be considered a modeling artifact. In this paper we propose a continuous-state space model for learning in games that can converge to mixed Nash equilibria, the Recency Weighted Sampler (RWS). The RWS is similar in spirit Young's model, but introduces a notion of best response where the players sample from a recency weighted history of interactions. We derive properties of the RWS which are known to hold for finite-state space models of adaptive play, such as the convergence to and existence of a unique invariant distribution of the process, and the concentration of that distribution on minimal CURB blocks. Then, we establish conditions under which the RWS process concentrates on mixed Nash equilibria inside minimal CURB blocks. While deriving the results, we develop a methodology that is relevant for a larger class of continuous state space learning models.

JEL: C72, C73

Keyworks: evolutionary game theory, learning in games, stochastic stability, recency, mixed Nash equilibria, minimal CURB blocks

Contents

1 Introduction

The general setting considered in this paper is the evolution of social conventions as introduced in Young (1993). There are large populations, one for each player role, from which players are randomly drawn to play a normal form game. Before deciding which action to take the players get access to a sample of historical interactions. The players use the sample to form beliefs about the opposite roles' historical behavior, and thereafter responds to the mixed strategy induced by that sample. Once they have played, the history is updated, new players are randomly drawn from the populations, and the process is repeated with the updated history.

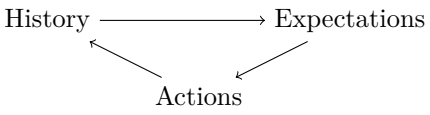


Figure 1: The players form expectations by sampling from historical records of interactions and then act based on those expectations. The realized play is appended to the history.

Social conventions form and evolve in many real life situations. For example, when buying a house each bidder (player) might not have participated in the exact same bidding (game) before, but has knowledge about some, but not all, previous interactions and assumes that the other bidders interacting with her will behave similarly to how bidders have historically behaved. By modeling repeated play based on historical records as diagrammed in Figure 1, one hopes to answer questions about which actions will be taken in the long run, and therefore which stable conventions, if any, will arise. We will refer to a dynamical model for the likelihood of the interactions, interpreted as the social convention, as a *learning process*.

When studying the long run distribution of the (state of the) learning process it is helpful, both theoretically and numerically, if it is a Markov process converging to its unique invariant distribution. In the original formulation of Young (1993) this is achieved by defining the state of the learning process as a vector of size m, a "finite memory" containing the m last interactions, by letting the players form beliefs by sampling $k \in \{1, \ldots, m\}$ strategies from the memory without replacement, and by assuming a small mistake probability $\varepsilon > 0$ with which a random action is taken instead of a best reply. The finite state space and $\varepsilon > 0$ ensures that Young's learning process has a unique invariant distribution to which it converges asymptotically.

Most of the work building on the original model contains the finite memory and noisy action structure, which is well suited for studying the relative stability of different pure (i.e., strict) Nash equilibria or minimal CURB blocks [1]. However, finite memory based learning is ill-suited to answer questions about the players' behavior around mixed Nash equilibria. The approach requires complete information of the order of the history, and exhibits behavior around even simple mixed Nash equilibria that is better viewed as a modeling artifact than as a realistic description of behavior. The purpose of this paper is to define a new learning process with the following features: firstly, it converges to some minimal CURB configuration and secondly, it behaves reasonably also inside minimal CURB and around mixed Nash equilibria.

[1] A subset (block) of strategy profiles C is called Closed Under Rational Behavior (CURB) if the best replies to any strategy profile with support in C is also in C. It is called a minimal CURB block if it does not contain any strictly smaller CURB block Basu and Weibull (1991).

	1	**2**
1	$1, -1$	$-1, 1$
2	$-1, 1$	$1, -1$

Table 1: The Matching Pennies payoff bimatrix. The row player has the "agreeing" role, aiming to match strategy with the column player, who has the "disagreeing" role, and aims to play differently than the row player. The unique mixed Nash equilibrium is $\left(\frac{1}{2}, \frac{1}{2}\right)$, fifty-fifty randomization for both players.

To better understand the limitations of the standard finite memory learning process, consider perhaps the simplest normal form game with a unique mixed Nash equilibrium: Matching Pennies, presented in Table 1. Consider the case where the length of the history is $m = 9$, and both players sample the whole history and play without a mistake, i.e., $k = m$ and $\varepsilon = 0$. Assume that the history contains, reading from the oldest to the latest entry, four interactions where both players took action **1**, followed by five interactions where both took action **2**. The row player will then take action **2** and the column player action **1**. However, since the interaction that falls out of the history is one where the column player played **1**, the sample to which the row player responds will not change until the **1**:s in the end of the history have all fallen out and the first interaction with a **2** falls out of the history. At that point, the history contains five interactions where the column player played **1**, so now the row player wants to play **1** as well. However, by now all the interactions in the history are such that the row player played **2**. So for the coming five interactions they will both take action **1**.

$$\binom{111122222}{111122222} \to \binom{222222222}{222211111} \to \binom{222211111}{111111111} \to \binom{111111111}{111122222} \to \cdots$$

The behavior in the next period depends as much on what falls out of the history as what is added, generating a cycling behavior. The cycling behavior does not only happen in this special case but is a general feature observed when simulating finite memory based learning processes, see Figure 2 for another example.

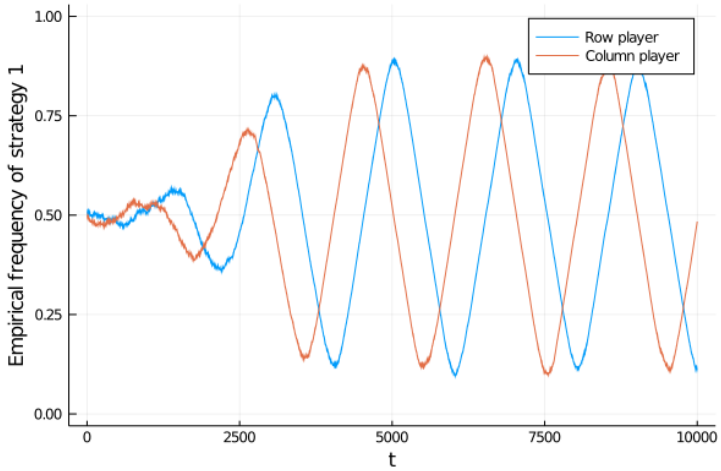


Figure 2: A 10 000 period simulation of Young's finite memory leaning process on Matching Pennies with $m = 1000$, $k = 20$, $\varepsilon = 0.05$. Initiated at the mixed Nash equilibrium.

To address the problem of unwanted cycling and to increase stability of social conventions we introduce a new learning process, the *Recency Weighted Sampler* (RWS). It differs from previous work in the structure of the historical record of plays. The history is assumed to be infinite, but more recent interactions are more likely to be sampled. A total of k samples are drawn with replacement by each player at each period. The probability of sampling the interaction of a past game decreases with a factor β, $0 < \beta < 1$ per game that has been played since. This geometric decrease allows us to use the sampling probabilities for the strategies as the state space of the learning process. The Markovian property of the process is preserved and we can in a meaningful way analyze it at a finer level inside the minimal CURB blocks (and determine properties of the distribution of interactions, i.e., the social convention, inside a minimal CURB block). As an example, a simulation of the RWS on Matching Pennies is presented in Figure 3. The RWS converges to a small neighborhood of the mixed Nash equilibrium and then stays there.

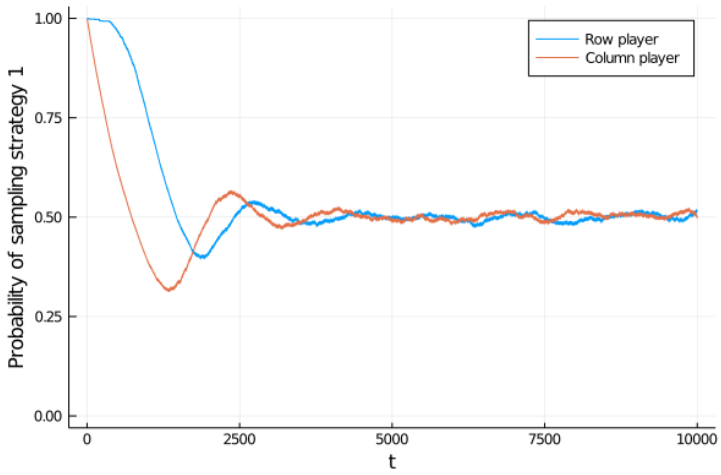


Figure 3: A 10 000 period simulation of the Recency Weighted Sampler on Matching Pennies with $\beta = 0.999$, $k = 20$, $\varepsilon = 0.05$. Initiated at the corner (1,1).

1.1 Related Literature

Already in his dissertation John Nash gave a second interpretation of the Nash equilibrium, the *mass action* interpretation (Nash, 1950). He assumes that a large population is associated to each player role, that one player per role is selected in each period to play the game, and that the individual players accumulate empirical information on the relative advantage of the different available pure strategies. He then argues, informally, that in such a setting, the stable points correspond to Nash equilibria and those points should eventually be reached by the process.

The mass action interpretation is appealing since its assumptions about bounded rationality and repeated interactions are more credible than those underlying the rationalistic interpretation built on assumptions of perfect rationality and common knowledge[2]. Furthermore, experimental evidence clearly favors some kind of learning and adjustment over the rationalistic motivation. The general result is that in a one-shot interaction, play rarely corresponds to a Nash equilibrium, but if the players have a chance to learn and adjust, play often (but far from always) moves to a Nash equilibrium. See (Camerer, 2003, Ch. 6) for an overview of experimental models and results.

Appealing as the motivation might be, the theoretical picture has turned out to be considerably more complicated than indicated by Nash's informal argument

[2]Especially since perfect rationality and common knowledge by itself only leads to rationalizability but not all the way to Nash equilibrium.

and what many researchers might initially have thought. One of the first, and most studied, models formalizing a setting similar in spirit to the mass action interpretation is that of fictitious play in Brown (1951). Even though Brown thought fictitious play would in general converge to a Nash equilibrium, it was shown in Shapley (1964) that even in a game with a unique Nash equilibrium there might only exist a stable cycle and no convergence to the mixed equilibrium. In general, it is the case that if the process has a stationary point, it must be a Nash equilibrium, but the existence of such a stationary point is not guaranteed. See e.g. Fudenberg et al. (1998), Weibull (1997), or Sandholm (2010) for overview of such results. Existing general results do not address convergence to stable points (which normally correspond to Nash equilibria) but convergence to stable sets. Ritzberger and Weibull (1995) show set-convergence results for evolutionary dynamics and Balkenborg, Hofbauer and Kuzmics (2013) for best reply dynamics. Similarly Hurkens (1995) and Young (1998) show set-convergence results for dynamics similar to those studied in this paper.

Smooth fictitious play, first introduced in Fudenberg and Kreps (1993), is a variant of fictitious play where players respond with a perturbed best response. In contrast to the standard version of fictitious play, not only the empirical frequency but also actual play can converge to a Nash equilibrium. In Benaïm and Hirsch (1999) and Hofbauer and Sandholm (2002), global convergence results are shown for some games with unique Nash equilibria, including interior ESS, two-player zero-sum, supermodular and potential games.

A downside with standard versions of fictitious play and smooth fictitious play is that the increments of the learning processes decrease in size over time. Therefore, in practice, the point of initialization is therefore crucial for convergence. Furthermore, if the behavior is cyclic the cycles take longer and longer time to complete. Introducing a bias towards more recent plays, similar to that used in this paper to define the RWS, yields processes with increments of similar size over time, which for many applications is natural. Such processes are studied in Benaïm, Hofbauer and Hopkins (2009), where the time average in unstable games is studied, and in Fudenberg, Levine et al. (2014).

The one class of dynamics for which there exists quite general results for convergence to equilibrium rely on a combination of noisy behavior and satisfaction (Foster and Young, 2003; Young and Foster, 2006; Hart and Mas-Colell, 2006; Block, Fudenberg and Levine, 2019). A given player randomly explores actions until she is satisfied, e.g., her received payoff is higher than some threshold or close enough to the maximum payoff observed. Then she keeps taking that action as long as she is still satisfied. The exact setting and formulation of results vary, but in general models of this category are able to converge to a Nash equilibrium under quite general circumstances. The unsatisfactory aspect is that players are in a sense too unsophisticated, at least if the game is known, and that the path to equilibrium thus might be very long and somewhat unrealistic.

The existing literature building on Young (1993, 1998) has not focused on

questions about convergence to mixed Nash equilibria, but instead focused on questions about speed of convergence of the learning process Kreindler and Young (2013) or improving tools for finding stochastically stable subsets Ellison (2000). To the best of our knowledge no one has more carefully studied convergence of learning processes to mixed Nash equilibria.

1.2 Summary and outline

In Section 2 the proposed learning process, the Recency Weighted Sampler, is formalized and we introduce the tools we need to analyze the process. Since we define a framework different from existing models (most crucially, RWS has a continuous state-space) we cannot rely directly on any existing results and we therefore begin by proving some standard properties of the learning process. We prove weak convergence for a class of learning process, of which the RWS is a member, to their respective unique invariant distribution. Following that, we show that in limit as the error-rate tends to zero, $\varepsilon \to 0$, the invariant distribution of the RWS will concentrate on the minimal CURB blocks of the game. Once we have recovered these crucial results, we turn to the question of behavior inside minimal CURB blocks that are non-singleton, and show that for any generic game where the minimal CURB blocks are at most 2×2 play will eventually concentrate around Nash equilibria or, when the sample size k is small, close to some point on the k-grid spanning the simplices which is also close to the Nash equilibrium. Proofs have been appended in the end of the paper.

2 The Recency Weighted Sampler

We consider a two-player finite game G, iteratively played by two new players drawn from large populations. The game has two asymmetric player roles, 1 and 2. The sets of pure strategies in the game are S_1 and S_2, containing $m_1 \in \mathbb{N}$ and $m_2 \in \mathbb{N}$ pure strategies respectively; the spaces of mixed strategies are thus $\Delta(S_1)$ and $\Delta(S_2)$. Throughout the paper, $-i$ denotes the index $\{1,2\}\backslash\{i\}$, $i \in \{1,2\}$. For $\sigma \in \Delta(S_{-i})$, we denote by $BR_i(\sigma) \subset S_i$ the set of best replies of player i to the mixed strategy σ. We identify $\Delta(S_i)$ with the $(m_i - 1)$-dimensional simplex and denote $\square(S) := \Delta(S_1) \times \Delta(S_2)$, $\square(S)$ being endowed with the usual uniform distance $\|\cdot\|$. We denote by $\mathcal{B}(\square(S))$ and $\mathcal{P}(\square(S))$ the Borel σ-field over $\square(S)$ and the set of probability measures over $\square(S)$, respectively.

2.1 The stochastic best reply of RWS

Each interaction is recorded as a pair (s_1, s_2), with $s_1 \in S_1$ and $s_2 \in S_2$ the strategies played by each player. Denoting $s_1(t)$ and $s_2(t)$ the strategies played at time t, the history is thus a sequence of plays

$$((s_1(t), s_2(t)))_{t\in\mathbb{Z}}.$$

Notice that for $t < 0$, the history is just some infinite history, coding for fictional plays for the purposes of our mechanisms.

At each time t, each player of role $i \in \{1, 2\}$ samples $k \in \mathbb{N}$ plays (with replacement) from the history of the opposing player role $-i$. Each sample is drawn independently and samples are drawn with bias towards more recent plays in a geometric fashion. Namely, players of role i have a bias $\beta \in [0, 1]$, called the *recency parameter*, such that at time t the probability of selecting the time period $t - \tau$, $\tau \in \{1, 2, \dots\}$ is

$$(1 - \beta)\beta^{\tau-1}.$$

Therefore, a play of the strategy $s \in S_{-i}$ will be sampled by player i with total probability

$$p_{-i,s}(t) = (1 - \beta)\sum_{\tau=1}^{\infty} \beta^{\tau-1} 1_s(s_{-i}(t - \tau)),$$

where 1_s is the indicator function on s.

We will call $p_i(t) := (p_{i,1}(t), \dots p_{i,m_i}(t))$ the state process of player role i at time t and $p(t) := (p_1(t), p_2(t))$ for the state process or the learning process, interchangeably. It is a vector of sampling probabilities obtained by player i from player $-i$'s history and is an element of $\Delta(S_{-i})$. The result of player i's sampling is a random vector $\left(n_{-i,1}(t), \dots n_{-i,m_{-i}}(t)\right)$ of integers, multinomially distributed with parameters k and $p_{-i}(t)$. For $s \in S_i$, let $\overrightarrow{1_{i,s}} \in \Delta(S_i)$ be the unit vector representing the pure strategy $s \in S_i$, i.e., a vector of of size m_i with 0 everywhere except at position s, where it is 1. From her sample, player i forms an empirical (average) opposing strategy profile

$$D_{-i}(t) := \frac{1}{k}\sum_{s=1}^{m_{-i}} n_{-i,s}(t)\overrightarrow{1_{-i,s}} \in \Delta(S_{-i}). \tag{1}$$

Player i now deems her opponent will play at turn t accordingly to the mixed strategy $D_{-i}(t)$ and tries to play the best response to it. However, player i can make a mistake. Player i's *error parameter* (or mistake frequency) $\varepsilon \in [0, 1]$ indicates the probability she will fail to play a strategy in $BR_i(D_{-i}(t))$, and instead play a strategy in S_i at random (with uniform probability). If $BR_i(D_{-i}(t))$ is not a singleton, the realized action is sampled uniformly from all

the elements of $BR_i(D_{-i}(t))$. We denote the outcome of the uniform sampling between all best replies to $\sigma \in \Delta(S_{-i})$ by $\widehat{BR}_i(x) \in S_i$. The distinction we want to emphasize with this notation is that $BR_i(x)$ is set-valued (the set of all best replies to x) while $\widehat{BR}_i(x)$ is S_i-valued and random (one of the best replies has been randomly selected).

In the end, player i will play $\widehat{BR}_i(D_{-i}(t))$, with $D_{-i}(t)$ obtained as described above, with a probability of $1-\varepsilon$; and additionally, play any strategy $s \in S_i$ with probability ε/m_i. We complete this section by calling

$$\widetilde{BR}_i(p_{-i}) \in S_i$$

the random choice of strategy obtained by a player i through the following process:

1. Looking at a history where plays of strategies by the opposing role get sampled with probabilities given by p_{-i};
2. Sampling k of them to form the belief $D_{-i} \in \Delta(S_{-i})$;
3. Actually playing the best response $\widehat{BR}_i(D_{-i})$, except in a fraction ε of the time when a randomly selected strategy is played.

2.2 The dynamics of RWS

At $t=0$, an initial history $((s_1(u), s_2(u)))_{u\in\mathbb{Z}_-}$, $s_i(u) \in S_i$, is given. At each time $t \in \mathbb{N}_0$, two new individuals are assigned to the roles. They use same values of the parameters k, β, and ε. After sampling from the history with recency parameter β, they play $s_i(t) = \widetilde{BR}_i(p_{-i}(t))$, $i=1,2$, where $p_{-i}(t)$ is exactly the historical distribution of plays with recency bias. The realized strategy profile $(s_1(t), s_2(t))$ is appended to the history, and the procedure restarts. The exponential nature of sampling leads to the following characterization of the RWS learning process.

Proposition 1. *The state process of player i, $p_i(t) \in \Delta(S_i)$, obeys the equation*

$$p_i(t+1) = \beta p_i(t) + (1-\beta)\overrightarrow{1_{i,s_i(t)}} \tag{2}$$

where $s_i(t) = \widetilde{BR}_i(p_{-i}(t))$ is drawn randomly according to the model.

The order of historical plays is not necessary to characterize the model, all the relevant information is captured by $(p_1(t), p_2(t)) \in \square(S)$. From the position $(p_1(t), p_2(t)) \in \square(S)$, at most $m_1 m_2$ different points $(p_1(t+1), p_2(t+1))$ may be reached. Conditioned on $p(t)$, for any $s_1 \in S_1$ and $s_2 \in S_2$ the point

$$\left(\beta p_1(t) + (1-\beta)\overrightarrow{1_{1,s_1}}, \beta p_2(t) + (1-\beta)\overrightarrow{1_{2,s_2}}\right)$$

will be reached when $s_1(t) = s_1$ and $s_2(t) = s_2$, which happens with probability

$$\prod_{i=1}^{2} \mathbb{P}\left(\widetilde{BR}_i(p_{-i}(t)) = s_i \mid p_{-i}(t)\right),$$

since players sample independently, and

$$\mathbb{P}\left(\widetilde{BR}_i\left(p_{-i}\left(t\right)\right) = s_i\right) = (1-\varepsilon)\,\mathbb{P}\left(\widehat{BR}_i\left(D_{-i}\left(t\right)\right) = s_i\right) + \varepsilon/m_i,$$

where $D_{-i}(t) \in \Delta\left(S_{-i}\right)$ is a multinomial combination of strategies (with parameters k and $p_{-i}(t)$).

2.3 Markovianity

By construction $(p(t); t \in \mathbb{N})$ is a Markov chain taking values in $\square(S)$. Since the state space is the continuous set $\square(S)$, the chain's transition kernel is a function $P : \square(S) \times \mathcal{B}(\square(S)) \to \mathbb{R}$ with the standard Markov kernel properties. The kernel takes a tuple (x, B) and returns the probability of the chain transitioning from x to B in one period. The kernel is the continuous state space equivalent of the transition rate matrix in models with a discrete state space. P is given as the following Markovian kernel: for all $(p_1, p_2) \in \square(S)$ and $B \in \mathcal{B}(\square(S))$,

$$P\left((p_1,p_2),B\right) = \sum_{s_1=1}^{m_1}\sum_{s_2=1}^{m_2} \mathbb{P}\left(\widetilde{BR}_1\left(p_2\right) = s_1, \widetilde{BR}_2\left(p_1\right) = s_2\right) \times$$
$$1_B\left(\beta p_1 + (1-\beta)\overrightarrow{1_{1,s_1}}, \beta p_2 + (1-\beta)\overrightarrow{1_{2,s_2}}\right).$$

Remark 2. *An underlying assumption of the RWS is that there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ carrying all the random variables necessary for defining the learning process. The space is filtered by $\mathbb{F}$, the natural filtration of the state process, and satisfies the usual conditions. The assumption is innocent, it only requires the space to carry a countable number of independent random variables. It is in this filtered space that we subsequently study the learning process as a Markov chain.*

3 Main results

3.1 Ergodicity

Our first result is Theorem 3 which states conditions for when the RWS state process is uniformly ergodic. We use the theory of Markov processes for the proof, the theory can be found in for example Meyn and Tweedie (2012) and the proof is found in the appendices.

Theorem 3. *If $\varepsilon > 0$ and $\beta \in (1 - \max\{m_1, m_2\}^{-1}, 1)$, then the Markov chain with kernel P is uniformly ergodic.*

In other words, for whichever initial distribution $\nu \in \mathcal{P}(\square(S))$ that $p(0)$ is drawn from, the distribution of $p(t)$ will converge "geometrically uniformly" as $t \to \infty$ to the probability measure μ_ε^* which is the unique solution of $\mu_\varepsilon^* P = \mu_\varepsilon^*$. More precisely, for every $\varepsilon \in (0, 1]$ there exists a unique $\mu_\varepsilon^* \in \mathcal{P}(\square(S))$, $c \in \mathbb{R}_+$, and $\lambda \in (0, 1)$ such that for all $p \geq 1$,

$$(W_p(\nu P^n, \mu_\varepsilon^*))^p \leq c\lambda^n, \quad \nu \in \mathcal{P}(\square(S)),$$

where W_p is the Wasserstein distance of order p between measures on $\square(S)$ (Villani, 2008, Def. 6.1) and c is a positive constant depending only on $\max_{x \in \square(S)} |x|$ and p.

The theorem itself is more general than what is needed for the goal of this paper. The result holds for any Markov chain with a compact state space and with a dynamic of the form (2), as long as there is a positive lower bound for the probability that any strategy is played (in any state) and that this probability is Lipschitz as a function of the state. Examples of other best response functions than the one studied here for which Theorem 3 applies are the logit best reply, i.e.,

$$\mathbb{P}\left[\widetilde{BR}_i(p) = s_i\right] = \frac{\exp(\eta \pi_i(s_i, p_{-i}))}{\sum_{a \in S_i} \exp(\eta \pi_i(a, p_{-i}))}$$

for some $\eta > 0$, models where k itself is a random parameter, and models where only robust best responses to the sample are considered.

3.2 Convergence to minimal CURB configurations

Before turning to the convergence to minimal CURB blocks, one minor technical detail most be resolved. A minimal CURB block is a collection of strategy profiles $C_1 \times C_2 \subset S$ such that the best reply to all mixed strategies in in the sub-simplex spanned by those strategies is always inside the set, i.e. $BR(\sigma) \subset C$ for all $\sigma \in \square(C)$, where $\square(C) := \Delta(C_1) \times \Delta(C_2)$. However, since our agents only reply to samples of size k, it might be the case that the mixed strategy from the simplex that has a best reply outside a non-CURB block simply never is sampled. The game below is a simple illustration of this point.

	1	**2**
1	$2, -100$	$-100, 2$
2	$-100, 2$	$2, -100$
3	$1, 0$	$1, 0$

If $k = 1$ only the best replies to pure strategies will ever be considered. If the process initially has support only on the block $\{\mathbf{1}, \mathbf{2}\} \times \{\mathbf{1}, \mathbf{2}\}$, the best reply to

any sample will be inside that block, even though **3** is the best reply to most properly mixed strategies. We could call this smaller set of blocks that are closed under best replies to any strategies on the k-lattice k-CURB blocks. In most settings, a relatively small k is enough for the k-CURB blocks to coincide with the CURB blocks. In the rest of the paper, we will speak of CURB blocks and by that mean k-CURB blocks. Alternatively, one can think of k as sufficiently large so that the notions coincide.

In what follows, we first prove that the RWS concentrates (in probability) on minimal CURB blocks for general two player games. Then we prove the concentration of RWS paths to an approximate mixed Nash equilibrium for games with $m_1 = m_2 = 2$ and a unique mixed Nash equilibrium.

3.2.1 Concentration on minimal CURB blocks

While proving concentration of the RWS on minimal CURB blocks we will partially rely on results for the original finite memory learning process. The RWS dynamics introduces some difficulties that are not present in the original model, mainly that once a strategy has been played it never truly disappears from memory but always has a positive probability of being sampled. However, the probability of sampling that strategy decreases over time as long as the strategy is not played again. A notion well-suited for the RWS is therefore the neighbourhood $B_\delta(C)$, $\delta > 0$, of $C := C_1 \times C_2 \subset S$, defined as all pairs (p_1, p_2) in $\square(S)$ such that each of the components puts at least $1 - \delta$ probability on the block C.

Definition 4. *For all $\delta > 0$,*

$$B_\delta(C) := \left\{ p = (p_1, p_2) \in \square(S) \;\middle|\; \sum_{s=1}^{m_i} p_{i,s} 1_{C_i}(s) \geq 1 - \delta,\ i = 1, 2 \right\}.$$

Let $\mathscr{C}$ denote the union of all minimal CURB blocks in the game. To prove the concentration result Theorem 5, we show that expected time to go from $B_\delta(\mathscr{C})^c$ to $B_\delta(\mathscr{C})$ is always bounded, but the expected time spent inside $B_\delta(\mathscr{C})$ once entered goes to infinity as ε goes to zero. This in turn will imply that as ε goes to zero, the invariant distribution concentrates on the neighbourhood $\mathscr{C}$, the union of all minimal CURB blocks.

Theorem 5. *If $\beta \in (1 - \max\{m_1, m_2\}^{-1}, 1)$, then for all $\delta > 0$ it holds that as $\varepsilon \to 0$, the invariant distribution of the Markov chain p concentrates on $B_\delta(\mathscr{C})$,*

$$\lim_{\varepsilon \to 0} \mu_\varepsilon^* \left(B_\delta(\mathscr{C}) \right) = 1.$$

3.2.2 Behavior inside minimal CURB

The previous section shows that as ε approaches zero, the RWS spends almost all the time inside minimal CURB blocks, possibly with rare excursions between different minimal CURB blocks. In this section, we justify that the RWS can actually concentrate on mixed Nash equilibria inside minimal CURB sets. This property is the main motivation for introducing the RWS and in contrast to similar learning processes.

Consider the deterministic mean-value process x,

$$\dot{x}_i(t) = \mathbb{E}\left[\widetilde{BR}_i(x_{-i}(t))\right] - x_i(t), \quad x_i(0) = p_i(0). \tag{3}$$

The process in (3) is a deterministic process that can be thought of as a continuous-time evolution of the expected value of the RWS state process (2). As a consequence of Lemma 18, if inside a given minimal CURB, the process (3) converges to either a stable point or a stable orbit with constant distance to a stable point, at least for ε small enough.

We show in Lemma 19, found in the appendices, that for a given time horizon T, divided into N time steps of size $(1-\beta)$, $T = N(1-\beta)$, and $\eta > 0$, the probability that the RWS stays closer than η to the deterministic process x during $[0, T]$ goes to 1 as β goes to 1. Taken together, if the deterministic process behaves well in the minimal CURB blocks of a game, we can by tuning β control the RWS and its concentration around stable points or stable orbits. The next theorem states that for a 2×2 minimal CURB block with a unique mixed Nash equilibrium the RWS concentrates around a stable point of (3), which is unique.

Theorem 6. *Let G be a 2×2 normal form game with a unique completely mixed Nash equilibrium. If $\beta > 1/2$, then, for all $\varepsilon, \eta > 0$ there exists a positive constant K such that*

$$\mu_\varepsilon^*\left(x \in \square(S) : \|x - n^*\|_\infty \geq \eta\right) = o\left(\exp\left(-\frac{K\eta^2}{1-\beta}\right)\right)$$

where n^* is the unique stationary point of (3).

The stationary point of (3) naturally depends on k. Under the assumptions in the theorem above, as $k \to \infty$ the equation $(\dot{x}_1(t), \dot{x}_2(t)) = (0, 0)$ is satisfied only by the Nash equilibrium, and we have that $\lim_{k\to\infty} n^* = N^*$. So n^* can be interpreted as a approximation of the Nash equilibrium.

The result of Theorem 6 can be extended to games of any size as long as they contain only minimal CURB blocks that are either 1×1, or are 2×2 and satisfy the assumption in Theorem 6. An argument can be found found in (Aurell, 2019, II.E).

4 Conclusions and outlook

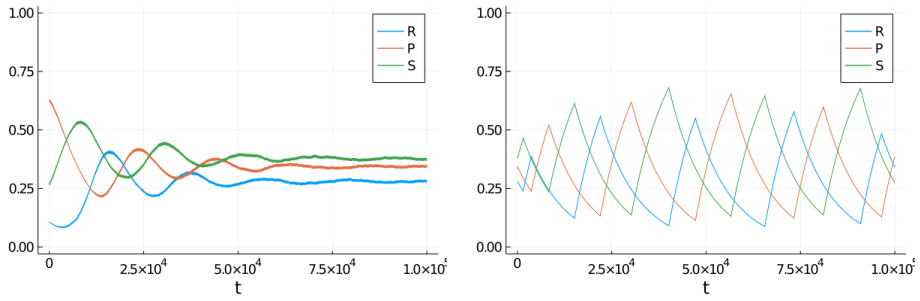
In this paper we have introduced a new process of adaptive play with sampling from history and recency, the RWS, and shown that it has several interesting properties. The invariant distribution of the RWS, which is a Markov process, concentrates on minimal CURB blocks as the mistake probability ε goes to zero. So in the long run, the RWS will almost always be inside a minimal CURB, perhaps with rare transitions between them. While the process is inside a given minimal CURB, the deterministic (mean) dynamics of the RWS will converge to either a stable point or a stable orbit, and the stochastic RWS state process does not deviate far from it during any finite time horizon with a high probability, if β is sufficiently close to 1. Combining these results we see that as ε and β approach 0 and 1, respectively, the RWS almost always is in the neighbourhood of a stable point or a stable orbit inside a minimal CURB. Furthermore, since the sampling best reply function we consider is continuous, this implies that if the state $p(t)$ is close to some stable point, then so is play.

For 2×2 minimal CURB blocks with a unique Nash equilibrium, we have shown that the deterministic process has a unique stable point which is close to the Nash equilibrium for most values of k. For games with minimal CURB blocks larger than 2×2, the picture is more complicated, and it is beyond the scope of this paper to completely map it out. However, for small to intermediate k the RWS behaves well, at least numerically, when other learning dynamics does not. Consider the unstable rock paper scissors game, see Table 2, studied in e.g. Benaïm, Hofbauer and Hopkins (2009).

	R	P	S
R	$0, 0$	$-3, 1$	$1, -3$
P	$1, -3$	$0, 0$	$-2, 1$
S	$-3, 1$	$1, -2$	$0, 0$

Table 2: The payoff in the Unstable Rock Paper Scissors game. The unique symmetric Nash equilibrium is $\left(\frac{9}{32}, \frac{10}{32}, \frac{13}{32}\right)$.

Classical learning processes such as fictitious play or reinforcement learning circles the Nash equilibrium in a stable cycle. In Figure 4 we compare the performance of RWS with $k = 20$ and fictitious play with recency. The RWS remains close to the equilibrium over time, even in this unstable game, while the fictitious play dynamic circles the equilibrium. When k is larger the RWS behaves as fictitious play with recency. This is expected, as k grows the sampled beliefs (D_1, D_2), see (1), become more and more similar to the sampling probabilities by the law of large numbers.



(a) RWS with $\varepsilon = 0$ and $k = 20$.

(b) Fictitious play with recency.

Figure 4: Simulations of behavior in the Unstable Rock Paper Scissors game. *Left:* RWS with a low k-value and no noise. *Right:* fictitious play with recency. The recency parameter was set to $\beta = 0.9999$ in both simulations.

There are still many unanswered questions about the RWS learning process. For example, we have not analyzed which minimal CURB configuration will have positive measure in the long run. It should be possible to conduct such an analysis using standard radius and co-radius arguments as in Ellison (2000) or Benaïm and Weibull (2003), but this is outside the scope of the current paper. On a final note, we expect that our results for 2×2 games can be extended to games with interior ESS, zero-sum games, potential games, and supermodular games, e.g. by following the techniques in Hofbauer and Sandholm (2002).

References

Aurell, Alexander. 2019. "Topics in the mean-field type approach to pedestrian crowd modeling and conventions." PhD diss. KTH Royal Institute of Technology.

Balkenborg, Dieter, Josef Hofbauer, and Christoph Kuzmics. 2013. "Refined best reply correspondence and dynamics." *Theoretical Economics*, 8(1): 165–192.

Basu, Kaushik, and Jörgen W. Weibull. 1991. "Strategy Subsets Closed Under Rational Behavior." *Economics Letters*, 36(2): 141–146.

Benaïm, Michel, and Jörgen W Weibull. 2003. "Deterministic approximation of stochastic evolution in games." *Econometrica*, 71(3): 873–903.

Benaïm, Michel, and Morris W Hirsch. 1999. "Mixed Equilibria and Dynamical Systems Arising from Fictitious Play in Perturbed Games." *Games and Economic Behavior*, 29(1-2): 36–72.

Benaïm, Michel, Josef Hofbauer, and Ed Hopkins. 2009. "Learning in games with unstable equilibria." *Journal of Economic Theory*, 144(4): 1694–1709.

Block, Juan I, Drew Fudenberg, and David K Levine. 2019. "Learning dynamics with social comparisons and limited memory 1." *Theoretical Economics*, 14(1): 135–172.

Brown, George W. 1951. "Iterative solution of games by fictitious play." *Activity analysis of production and allocation*, 13(1): 374–376.

Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton University Press.

Ellison, Glenn. 2000. "Basins of Attraction, Long-Run Stochastic Stability, and the Speed of Step-by-Step Evolution." *Review of Economic Studies*, 67(1): 17–45.

Folland, Gerald B. 1999. *Real analysis: modern techniques and their applications.* Vol. 40, John Wiley & Sons.

Foster, Dean P., and H. Peyton Young. 2003. "Learning, hypothesis testing, and Nash equilibrium." *Games and Economic Behavior*, 45(1): 73–96.

Fudenberg, Drew, and David M. Kreps. 1993. "Learning Mixed Equilibria." *Games and Economic Behavior*, 5: 320–367.

Fudenberg, Drew, David K Levine, et al. 2014. "Learning with recency bias." *Proceedings of the National Academy of Sciences*, 111: 10826–10829.

Fudenberg, Drew, Fudenberg Drew, David K Levine, and David K Levine. 1998. *The theory of learning in games.* Vol. 2, MIT press.

Hart, Sergiu, and Andreu Mas-Colell. 2006. "Stochastic uncoupled dynamics and Nash equilibrium." *Games and Economic Behavior*, 57(2): 286–303.

Hofbauer, Josef, and William H. Sandholm. 2002. "On the Global Convergence of Stochastic Fictitious Play." *Econometrica*, 70(6): 2265–2294.

Hurkens, Sjaak. 1995. "Learning by Forgetful Players." *Games and Economic Behavior*, 11(2): 304–329.

Kreindler, Gabriel E., and H. Peyton Young. 2013. "Fast convergence in evolutionary equilibrium selection." *Games and Economic Behavior*, 80: 39–67.

Meyn, Sean P, and Richard L Tweedie. 2012. *Markov chains and stochastic stability.* London:Springer-Verlag.

Nash, John. 1950. "Non-cooperative games." PhD diss. Princeton University.

Ritzberger, Klaus, and Jorgen W. Weibull. 1995. "Evolutionary Selection in Normal-Form Games." *Econometrica*, 63(6): 1371.

Sandholm, William H. 2010. *Population games and evolutionary dynamics.* MIT press.

Shapley, Lloyd. 1964. "Some topics in two-person games." *Advances in game theory*, 52: 1–29.

Slotine, Jean-Jacques E, Weiping Li, et al. 1991. *Applied nonlinear control.* Vol. 199, Prentice hall Englewood Cliffs, NJ.

Villani, Cédric. 2008. *Optimal transport: old and new.* Vol. 338, Springer Science & Business Media.

Weibull, Jörgen W. 1997. *Evolutionary game theory.* MIT press.

Young, H. Peyton. 1993. "The Evolution of Conventions."

Young, H. Peyton. 1998. *Individual Strategy and Social Structure An Evolutionary Theory of Institutions.* Princeton University Press.

Young, Peyton Hobart, and Dean P. Foster. 2006. "Regret testing: learning to play Nash equilibrium without knowing you have an opponent." *Theoretical Economics*, 1(3): 341–367.

A The basic properties of the learning process: proofs

A.1 Exponential history

Let us prove Proposition 1. Starting from the definition, we have

$$p_{-i,s}(t+1) = (1-\beta)\sum_{\tau=1}^{\infty}\beta^{\tau-1}1_s(s_{-i}(t-\tau+1)).$$

After index substitution $v = \tau - 1$, splitting the term $v = 0$ yields

$$p_{-i,s}(t+1) = (1-\beta)\left(1_s(s_{-i}(t)) + \sum_{v=1}^{\infty}\beta^v 1_s(s_{-i}(t-v))\right).$$

In other words,

$$p_{-i,s}(t+1) = (1-\beta)\beta\sum_{v=1}^{\infty}\beta^{v-1}1_s(s_{-i}(t-v)) + (1-\beta)1_s(s_{-i}(t)).$$

We recognize the first term as $p_{-i,s}(t)$, so we are left for every $s \in S_{-i}$ with

$$p_{-i,s}(t+1) = \beta p_{-i,s}(t) + (1-\beta)1_s(s_{-i}(t)),$$

which is the representation we seek.

A.2 Lipschitz continuity

Lemma 7. *For all $k \in \mathbb{N}$, $i \in \{1,2\}$, and $a \in \{1,\dots,m_i\}$,*

$$\Delta(S_{-i}) \ni p \to \mathbb{P}\left(\widetilde{BR}_i(p) = a\right)$$

is Lipschitz continuous with Lipschitz coefficient at most $(1-\varepsilon)km_{-i}$.

Proof. At the beginning there is a sample with respect to probabilities p, yielding a random vector $N := \left(n_{-i,1}(t), \dots n_{-i,m_{-i}}(t)\right)$ of integers from the (discrete) probability distribution

$$\mathbb{P}\left(N = \left(n_1, \dots n_{m_{-i}}\right)\right) = k! \prod_{j=1}^{m_{-i}} \frac{p_s^{n_s}}{n_s!}.$$

Each N will lead to an empirical opposing strategy profile D, that must belong to some finite 'simplex grid'

$$\Delta^{(-i,k)} := \left\{ \frac{1}{k} \sum_{s \in S_{-i}} n_s \overrightarrow{1_{-i,s}} \; ; \; n_s \in \mathbb{N}_0, \sum_{s \in S_{-i}} n_s = k \right\}.$$

Now let us form m_i subsets from $\Delta^{(-i,k)}$ (which is finite), named $\Delta_s^{(-i,k)}$ for $s \in S_i$, where $x \in \Delta_s^{(-i,k)}$ whenever $s \in BR_i(x)$. Note that $(\Delta_s^{(-i,k)})_s$ is not a disjoint cover of $\Delta^{(-i,k)}$ except in the special case when each $x \in \Delta^{(-i,k)}$ has a unique best response. Also, $\cup_s \Delta_s^{-i,k} = \Delta^{-i,k}$ since the best response set is never empty.

For $a \leq m_i$, the probability that $\widetilde{BR}_i(p) = a$ is going to be played is thus obtained as follows :

- If the player i trembles, which happens a fraction ε of the time, strategy a is played with a probability $1/m_i$, totalling ε/m_i.
- Otherwise the player selects its best response, so it will be a with the probability $\mathbb{P}\left(D \in \Delta_a^{(-i,k)}, \widehat{BR}_i(D) = a\right)$.

In short,

$$\mathbb{P}\left(\widetilde{BR}_i(p) = a\right) = \varepsilon r_a + (1-\varepsilon) \sum_{x \in \Delta_a^{(-i,k)}} \mathbb{P}\left(\widehat{BR}_i(x) = a\right) \mathbb{P}(D = x). \quad (4)$$

However $D = x$ is an event of the shape $N = \left(n_1, \dots n_{m_{-i}}\right)$, so considering $\mathbb{P}(D = x)$ as a function of $p_1, \dots p_n$, we get

$$\frac{\partial \mathbb{P}\left(N = \left(n_1, \dots n_{m_{-i}}\right)\right)}{\partial p_b} = k! \frac{p_b^{n_b - 1}}{(n_b - 1)!} \prod_{j \neq b} \frac{p_j^{n_j}}{n_j!},$$

with the convention $1/(-1)! = 0$ for continuity. So relatively to the norm $\|.\|_\infty$ over $\Delta(S_{-i})$, the Lipschitz constant of the probabilities $\mathbb{P}\left(D \in \Delta_a^{(-i,k)}\right)$ are at most

$$\sum_{b=1}^{m_{-i}} \left| \frac{\partial \mathbb{P}\left(D \in \Delta_a^{(-i,k)}\right)}{\partial p_b} \right| \leq \sum_{b=1}^{m_{-i}} \sum_{x \in \Delta_a^{(-i,k)}} k! \frac{p_b^{n_b - 1}}{(n_b - 1)!} \prod_{\substack{j=1 \\ j \neq b}}^{m_{-i}} \frac{p_j^{n_j}}{n_j!}.$$

However we know that

$$\sum_{x \in \Delta^{(-i,k)}} \frac{p_b^{n_b - 1}}{(n_b - 1)!} \prod_{\substack{j=1 \\ j \neq b}}^{m_{-i}} \frac{p_j^{n_j}}{n_j!} = \frac{1}{(k-1)!},$$

as this is the multinomial formula for $k-1$ draws. Since $\Delta_a^{(-i,k)} \subset \Delta^{(-i,k)}$, the Lipschitz constant of $\mathbb{P}\left(D \in \Delta_a^{(-i,k)}\right)$ is at most

$$\sum_{b=1}^{m_{-i}} k! \frac{1}{(k-1)!} = k m_{-i}.$$

Bounding $\mathbb{P}\left(\widehat{BR}_i(x) = a\right)$ from (4) by 1, the Lipschitz constant for

$$p \mapsto \mathbb{P}\left(\widetilde{BR}_i(p) = a\right)$$

is at most $(1 - \varepsilon) k m_{-i}$. □

A.3 Ergodicity

The proof of ergodicity relies on standard Markov chain theory and a positive lower bound for the probability that the chain, initiated at any point in $\square(S)$, visits any open set in $\square(S)$ after a finite number of time steps. To prove the lower bound, we first need to establish the intermediate result Lemma 8. It is assumed throughout this section that $\varepsilon > 0$.

A.3.1 Approximative history

For $i \in \{1, 2\}$, $j \in \{1, \ldots, m_i\}$, and $t \in \mathbb{N}$, let $\omega_{i,j,t} := 1_j(s_i(t))$ be the indicator of a play j by player i at time t, so that

$$p_{i,j}(t) = (1 - \beta) \sum_{\tau=1}^{\infty} \beta^{\tau - 1} \omega_{i,j,t-\tau}.$$

We will call $\Sigma^{(i)} := \{0,1\}^{m_i \times \mathbb{N}}$ the set of binary arrays, indexed by $s \in \{1,\dots,m_i\}$ and $t \in \mathbb{N}$, such that for every t there is exactly one s such that $\Sigma^{(i)}_{s,t} = 1$. In other words, $\Sigma^{(i)}$ represents a possible history for player i, where a 1 at the entry (s,t) indicates that s was played at time t. Likewise, for $n \in \mathbb{N}$, we will call $\Sigma^{(i,N)} := \{0,1\}^{m_i \times N}$ the set of binary arrays indexed by $s \in \{1,\dots,m_i\}$ and $t \in \{1,\dots N\}$ obeying the same condition, in other words the history up to time N.

Let $p_i \in \Delta(S_i)$. We are going to exhibit a sequence of plays of finite length N, i.e., an $\omega \in \Sigma^{(i,N)}$ for some $N \in \mathbb{N}$, such that the partial sum

$$p^{(N)}_{i,j} := (1-\beta)\sum_{\tau=1}^{N} \beta^{t-1}\omega_{i,j,\tau}$$

falls close to p_i. Namely, we want to prove the following.

Lemma 8. *Let $p_i \in \Delta(S_i)$ and $\delta > 0$. We assume that $(1-\beta)\, m_i \leq 1$. There exists an $N(\delta) \in \mathbb{N}$, independent of i and p_i, such that there is a history $\omega_i^{(N)} \in \Sigma^{(i,N)}$ for each $N \geq N(\delta)$ which satisfies*

$$p^{(N)}_{i,j} = (1-\beta)\sum_{\tau=1}^{N} \beta^{\tau-1}\omega^{(N)}_{i,j,\tau} \in (\max\{p_{i,j}-\delta, 0\}, p_{i,j}] \tag{5}$$

for all $j \in \{1,\dots,m_i\}$.

Proof. The following algorithm provides a proof of Lemma 8. Start by setting $p^{(0)}_{i,j} = 0$ for all $j = 1,\dots,m_i$, and $\omega_i^{(0)}$ to the empty array of dimensions 0 and m_i. Define $N(\delta)$ as the smallest $N \in \mathbb{N}$ such that $\beta^N < \delta$, i.e.,

$$N(\delta) := \inf\{N \in \mathbb{N} : \beta^N < \delta\}.$$

For $t \in \{1,\dots N(\delta)\}$, repeat the following steps :

1. Look for the indices $j \in \{1,\dots,m_i\}$ such that $p_{i,j} - p^{(t-1)}_{i,j}$ is maximal, and call any of these indices a.
2. Append $\overrightarrow{1}_{1,a}$ to $\omega_i^{(t)}$. Now $\omega^{(t)}_{i,a,t} = 1$ and $\omega^{(t)}_{i,j,t} = 0$ for $j \neq a$.
3. Compute $p^{(t)}_{i,j}$ accordingly to (5) and the updated history $\omega_i^{(t)}$.

Return the final history $\omega_i^{(N(\delta))}$ and values $p^{(N(\delta))}_{i,j}$.

We are going to prove inductively that for every $t \in \mathbb{N}$, we always have

$$p^{(t)}_{i,j} \leq p_{i,j}, \quad j = 1,\dots,m \tag{6}$$

and

$$\sum_{j=1}^{m_i} p_{i,j}^{(t)} = 1 - \beta^t. \tag{7}$$

For $t = 0$, (6) is true since $p_{i,j}$ is non-negative and $p_{i,j}^{(0)} = 0$ for all $j = 1, \ldots, m_i$, which also yields that (7) holds at $t = 0$. Now assume that (6)–(7) hold at time t. Since $\sum_{j=1}^{m_i} p_{i,j} = 1$, the maximal difference $\max_{1 \le j \le m_i}(p_{i,j} - p_{i,j}^{(t)})$ must be at least β^t / m_i. By definition, then $\omega_{i,a,t+1} = 1$ for some $a \in \{2, \ldots, m_i\}$ and

$$\begin{aligned}
p_{i,a}^{(t+1)} &= (1-\beta) \sum_{\tau=1}^{t+1} \beta^{\tau-1} \omega_{i,a,\tau}^{(t+1)} \\
&= (1-\beta)\beta^t \omega_{i,a,t+1}^{(t+1)} + (1-\beta) \sum_{\tau=1}^{t} \beta^{\tau-1} \omega_{i,a,\tau}^{(t)} \\
&= (1-\beta)\,\beta^t + p_{i,a}^{(t)} \\
&\le \left((1-\beta) - \frac{1}{m_i} \right) \beta^t + p_{i,a}
\end{aligned}$$

Therefore, since $(1-\beta)\, m_i \le 1$ as assumed, the right-hand side is also bounded by $p_{i,j}$. As for other strategies $j \neq a$, since $p_{i,j}^{(t+1)} = p_{i,j}^{(t)}$ the inequality $p_{i,j}^{(t+1)} \le p_{i,j}$ holds and we have proven the induction step for (6). Now we also know that

$$p_{i,j}^{(t+1)} - p_{i,j}^{(t)} = (1-\beta)\,\beta^t \omega_{i,j,t+1}^{(t+1)},$$

and since exactly one among the m_i entries in $\omega_{i,t+1}^{(t+1)}$ is 1, the other being zero, we have

$$\sum_{j=1}^{m_i} \left(p_{i,j}^{(t+1)} - p_{i,j}^{(t)} \right) = (1-\beta)\,\beta^t.$$

The induction hypothesis thus leads us to

$$\sum_{j=1}^{m_i} p_{i,j}^{(t+1)} = 1 - \beta^t + (1-\beta)\,\beta^t = 1 - \beta^{t+1},$$

which proves (7) by induction. So in particular after time $N(\delta)$, by choice of $N(\delta)$, for every $N \ge N(\delta)$ we have

$$\sum_{j=1}^{m_i} p_{i,j}^{(N)} > 1 - \delta,$$

while $p_{i,j}^{(N)} \le p_{i,j}$ for every j. Since $\sum_j p_{i,j} = 1$, this is possible only if $p_{i,j}^{(N)} > p_{i,j} - \delta$ for each $j = 1, \ldots, m_i$, leading to the result. □

A.3.2 A useful lower bound

Let $x = (x_1, x_2) \in \square(S)$. We apply Lemma 8 with $\delta = \varepsilon$ to x_1 and x_2, yielding play records $\omega_1^{(N(\varepsilon))}$ and $\omega_2^{(N(\varepsilon))}$, and values $p_{i,j}^{(N(\varepsilon))}$ such that for every $1 \leq j \leq m_i$ and $N \geq N(\varepsilon)$,

$$p_{i,j}^{(N)} \in \left(\max\{x_{i,j} - \varepsilon, 0\}, x_{i,j}\right].$$

We know that for any $B \in \mathcal{B}(\square(S))$,

$$P(x, B) = \sum_{s_1=1}^{m_1} \sum_{s_2=1}^{m_2} \sigma(x, s)\, 1_B\left(\Gamma(x, s)\right),$$

and $\sigma(x, s)$ is uniformly bounded from below by $\eta = \varepsilon/(m_1 m_2) > 0$.

The play records $\omega_1^{(N)}$ and $\omega_2^{(N)}$ up to time N from the previous Lemma are now read in reverse time order. At each time step $t \in \{0, \ldots N-1\}$, there is a probability at least η^2 that player 1 chooses the strategy $1 \leq a \leq m_1$ given by $\omega_{1,a,N-t}^{(N} = 1$, and player 2 chooses the strategy $1 \leq b \leq m_2$ given by $\omega_{2,b,N-t}^{(N)} = 1$. Therefore the plays up to time N have a probability at least $\eta^{2N} > 0$ of being dictated by $\omega_1^{(N}$ and $\omega_2^{(N}$. When this happens, thanks to the Proposition 1, a history having started by $(p_1(0), p_2(0)) = p$ will now be at the position

$$\begin{aligned}(p_1(N), p_2(N)) \quad = \quad & \sum_{t=1}^{N} \left((1-\beta)\beta^{t-1}\omega_{1,t}^{(N)}, (1-\beta)\beta^{t-1}\omega_{2,t}^{(N)}\right) \\ & + \left(\beta^N p_1(0), \beta^N p_2(0)\right),\end{aligned}$$

with probability greater or equal to η^{2N}.

By Lemma 8, the choice of the records $\omega_1^{(N)}$ and $\omega_2^{(N)}$ makes j:th component of the sum on the right-hand side of (8) take some value between $(\max\{x_{1,j} - \varepsilon, 0\}, \max\{x_{2,j} - \varepsilon, 0\})$ and $(x_{1,j}, x_{2,j})$. As we also have $\beta^N < \varepsilon$ and $p_{i,j}(0) \leq 1$, we get $p_{i,j}(N) \in (x_{i,j} - \varepsilon, x_{i,j} + \varepsilon)$. We conclude that for all $N \geq N(\varepsilon)$

$$\mathbb{P}\left(|p(N) - x| < \varepsilon\right) \geq \eta^{2N}. \tag{10}$$

In other words, it means that the point $y = (p_1(N), p_2(N))$, which is in an ε-neighbourhood of x, is accessible from p in N steps.

A.3.3 Proof of uniform ergodicity

The path to uniform ergodicity goes through proving that the chain is a so-called T-chain. For Markov chain theory related concepts used below, we follow the definitions of Meyn and Tweedie (2012).

Lemma 9. *The Markov chain p is a T-chain.*

Proof. From (10) we know that for all rectangles $R = R_1 \times R_2 \in \mathcal{B}(\square(S))$ and $N \geq N(\varepsilon_R)$,

$$P^N(x, R) \geq \eta^{2N}, \quad x \in \square(S), \tag{11}$$

where $\varepsilon_R = \frac{1}{2}\min\{\lambda(R_1), \lambda(R_2)\}$ with λ the Lebesgue measure, $N(\varepsilon_R) = \inf\{N \in \mathbb{N} : \beta^N < \varepsilon_R\}$, and $\eta = \varepsilon/(m_1 m_2)$. Note that $\eta < 1$, so if $\varepsilon_R = 0$, which implies that $N(\varepsilon_R) = \infty$, then $\eta^{2N(\varepsilon_R)} = 0$, hence the estimate also covers degenerate rectangles.

Let O be an open subset of $\square(S)$, then O can be written as a countable union of almost disjoint (their boundaries may overlap) closed rectangles $(R_j)_{j=1}^{\infty}$, $R_j = R_{j,1} \times R_{j,2}$. Since O is open, at least one of the rectangles in the cover must have a nonempty interior. The probability to reach O from a point $x \in \square(S)$ is bounded from below by the sum of the probabilities of reaching each of the rectangles covering O, when staring the chain from $x \in \square(S)$. Hence we have the following lower bound

$$\sum_{n=0}^{\infty} P^n(x, O) \geq \sum_{j=1}^{\infty} P^{N(\varepsilon_{R_j})}(x, R_j) \geq \sum_{j=1}^{\infty} \eta^{2N(\varepsilon_{R_j})} > 0, \quad x \in \square(S). \tag{12}$$

In Lemma 10 below, we use (12) to construct a nontrivial measure $\nu_{\square}$ on $(\square(S), \mathcal{B}(\square(S)))$ such that

$$\sum_{n=0}^{\infty} P^n(x, B) \geq \nu_{\square}(B), \quad B \in \mathcal{B}(\square(S)),\ x \in \square(S), \tag{13}$$

hence yielding that $\square(S)$ is a $\nu_{\square}$-petite set.

By (12), all open sets are uniformly accessible from any subset of $\square(S)$ by (12). Since $\square(S)$ is open in the relative topology, the previously stated fact implies that all subsets of $\square(S)$ are petite (Meyn and Tweedie, 2012, Prop. 5.5.3). In particular, every compact set is petite and it follows that p is a T-chain (Meyn and Tweedie, 2012, Thm. 6.0.1). $\square$

Lemma 10. *There exists a nontrivial measure $\nu_{\square}$ on $(\square(S), \mathcal{B}(\square(S))$ that satisfies* (13)*.*

Proof. Define $\mathscr{R}$ to be the collection of all half-open rectangles $\left(\times_{j=1}^{m_1-1}[a_{1,j}, b_{1,j})\right) \times \left(\times_{j=1}^{m_2-1}[a_{2,j}, b_{2,j})\right)$ in $\mathbb{R}^{m_1-1} \times \mathbb{R}^{m_2-1}$. Let the function $\bar{\eta} : \mathscr{R} \to [0, \infty]$ be given by $\bar{\eta}(R) = \eta^{2N(\varepsilon_R)}$ (clearly, if $R \cap \square(S) = \emptyset$ then $\bar{\eta}(R) = 0$). We define, for any $A \subset \mathbb{R}^{m_1-1} \times \mathbb{R}^{m_2-1}$,

$$\nu^*(A) := \inf\left\{\sum_{j=1}^{\infty} \bar{\eta}(R_j) : R_j \in \mathscr{R}, A \subset \cup_{j=1}^{\infty} R_j\right\}.$$

Then ν^* is a countably additive pre-measure on the semi-ring $\mathscr{R}$, and an outer measure on $\mathbb{R}^{m_1-1}\times\mathbb{R}^{m_2-1}$. We denote by ν the restriction of ν^* to its measurable sets. Carathéodory's extension theorem says that ν is a measure on the smallest σ-algebra containing $\mathscr{R}$, which is $\mathcal{B}(\mathbb{R}^{m_1-1}\times\mathbb{R}^{m_2-1})$ since the half-open rectangles generate the Borel σ-algebra. Furthermore, since ν^* is σ-finite, ν is the unique extention of ν^*, and ν agrees with $\bar{\eta}$ on $\mathscr{R}$.

Let, for each $x \in \square(S)$, $\bar{\eta}_\Delta(x,\cdot)$ be a set-function on $\mathbb{R}^{m_1-1}\times\mathbb{R}^{m_2-1}$ that satisfies

$$\begin{aligned}\bar{\eta}_\Delta(x,R) &= \sum_{n=1}^{\infty} \bar{P}^n(x,R) - \nu(R) \\ &= \sum_{n=1}^{\infty} \bar{P}^n(x,R) - \bar{\eta}(R), \quad R \in \mathscr{R}.\end{aligned}$$

where $\bar{P}(x,R) := P(x, R\cap\square(S))$ extends P to $\mathcal{B}(\mathbb{R}^{m_1-1}\times\mathbb{R}^{m_2-1})$. By (11), $\bar{\eta}_\Delta(x,\cdot)$ is non-negative for all $x \in \square(S)$. For each $x \in \square(S)$, define for any $A \subset \mathbb{R}^{m_1-1}\times\mathbb{R}^{m_2-1}$

$$\nu^*_\Delta(x,A) := \inf\left\{\sum_{j=1}^{\infty}\bar{\eta}_\Delta(x,R_j) : R_j \in \mathscr{R}, A \subset \cup_{j=1}^{\infty} R_j\right\}.$$

Then $\nu^*_\Delta(x,\cdot)$ is, for each $x \in \square(S)$, a countably additive pre-measure on $\mathscr{R}$ and an outer measure on $\mathbb{R}^{m_1-1}\times\mathbb{R}^{m_2-1}$, and with the same argument used to constructed ν, we construct the measures $(\nu_\Delta(x,\cdot))_{x\in\square(S)}$, $\nu_\Delta(x,\cdot)$ being the unique extension of $\nu^*_\Delta(x,\cdot)$ (the σ-finite part follows from the definition of $\bar{\eta}_\Delta(x,\cdot)$; a countable sum of σ-finite measures is again a σ-finite measure). The following fact is essentially (Folland, 1999, Theorem 1.14), and follows as a corollary to Carathéodory's extension theorem: Since

$$\nu_\Delta(x,R) = \sum_{n=1}^{\infty}\bar{P}^n(x,R) - \nu(R),$$

for all $R \in \mathscr{R}$,

$$\nu_\Delta(x,B) = \sum_{n=1}^{\infty}\bar{P}^n(x,B) - \nu(B)$$

for all $B \in \mathcal{B}(\mathbb{R}^{m_1-1}\times\mathbb{R}^{m_2-1})$. Then, in particular,

$$\nu_\Delta(x,B) = \sum_{n=1}^{\infty}P^n(x,B) - \nu(B), \quad B \in \mathcal{B}(\square(S)),\ x \in \square(S)$$

from which it follows that

$$\sup_{x\in\square(S)}\sum_{n=1}^{\infty}P^n(x,B) \geq \nu(B), \quad B \in \mathcal{B}(\square(S)).$$

Defining $\nu_\square$ as the restriction of ν to $\mathcal{B}(\square(S))$ completes the proof. □

Lemma 11. *The Markov chain p is open set irreducible.*

Proof. A point $x \in \square(S)$ is called reachable if for every open set $O \in \mathcal{B}(\square(S))$ containing x,

$$\sum_{n=1}^{\infty} P^n(y, O) > 0, \quad y \in \square(S).$$

We know that all $x \in \square(S)$ are reachable by (10). A Markov chain is open set irreducible if every point is reachable. □

Proposition 12. *The chain p is ψ-irreducible.*

Proof. We know that p is an open set irreducible T-chain. By (Meyn and Tweedie, 2012, Prop. 6.2.1), p is ψ-irreducible. □

Remark 13. *The measure $\nu_{\square}$ is an irreducibility measure for p (Meyn and Tweedie, 2012, Prop. 5.5.4 (ii)).*

Remark 14. *The state space $\square(S)$ is a petite set, but not a small set, since there are sets in the corners of $\square(S)$ which the chain needs arbitrarily long time to reach.*

We move on towards showing uniform ergodicity for p. The argument is based on (Meyn and Tweedie, 2012, Thm. 16.2.5), which says that if p is a ψ-irreducible and aperiodic T-chain, and if the state space $\square(S)$ is compact, then p is uniformly ergodic.

Lemma 15. *The Markov chain p is aperiodic.*

Proof. The negative implication of (Meyn and Tweedie, 2012, Prop. 5.4.6) says that if there exists no absorbing state for $p^{(d)}$, the chain corresponding to the transition kernel P^d, for any $d \geq 2$, then p is aperiodic.

Assume that D is an absorbing state for $p^{(d)}$, that is $\inf_{x \in D} P^d(x, D) = 1$. By (11), D must contain all rectangles $R \subset \square(S)$, since $\inf_{x \in D} P^{Nd}(x, R) \geq \eta^{2Nd} > 0$ for any $N \geq N(\varepsilon_R)$. This implies that $D = \square(S)$ is the only absorbing state for $p^{(d)}$ and we conclude that the chain is aperiodic. □

We have proven the following result:

Proposition 16. *The chain p is uniformly ergodic.*

Proof. This follows by ψ-irreducibility (Proposition 12) and aperiodicity (Lemma 15), see (Meyn and Tweedie, 2012, Thm. 16.2.5). □

By (Meyn and Tweedie, 2012, Thm. 15.0.1), ψ-irreducibility and aperiodicity implies that p has an invariant probability measure μ_ε^*. By (Meyn and Tweedie, 2012, Thm. 16.0.2), uniform ergodicity of p is equivalent to the existence of $r > 1$ and $R < \infty$ such that for all x,

$$d_{TV}(P^n(x,\cdot), \mu_\varepsilon^*) \leq Rr^{-n},$$

where d_{TV} is the total variation distance on $\mathcal{P}(\square(S))$. Clearly, μ_ε^* is the unique invariant probability measure of p (in $(\mathcal{P}(\square(S)), d_{TV})$). By (Villani, 2008, Thm. 6.18), the p-Wasserstein distance W_p is for all $p \geq 1$ controlled by the total variation on bounded sets, and

$$\sup_{x \in \square(S)} W_p^p(P^n(x,\cdot), \mu_\varepsilon^*) \leq CRr^{-n},$$

where C depends on p and $|\square(S)|$. The last inequality implies the statement of Theorem 3.

A.4 Proof of Theorem 5

Proof. The proof consists of four steps.

Step 1. Bounding the probability of reaching $B_\delta(\mathscr{C})$ in finite time.
To find a lower bound for the probability to go from an arbitrary point $p(t) \in B_\delta(\mathscr{C})^c$ to $B_\delta(\mathscr{C})$ in finite time we create a particular path of positive probability that does exactly that. Let $p(t) \in \square(S)$ be given and let $s^1 \in S_1 \times S_2$ be the strategy profile played at in period t. Either s^1 is a CURB block, or the best reply set to s^1 contains a strategy not in s^1, $BR(\overrightarrow{1_{s^1}}) \not\subset s^1$. If the former statement is true this step of the proof is complete. That is not always the case, therefore assume that we are in the case of the latter statement, i.e. that the best reply set to s^1 contains a strategy not in s^1. Then, the probability of both players only sampling s^1 at time $t+1$ is bounded from below by $(1-\beta)^{2k}$. Hence the probability of a strategy profile $s^2 \in BR(\overrightarrow{1_{s^1}}), s^2 \neq s^1$, being played is bounded from below by

$$\mathbb{P}\left(\widetilde{BR}(p(t)) = s^2 \mid p(t)\right) \geq \frac{(1-\beta)^{2k}}{m_1 m_2}(1-\varepsilon)^2.$$

Now let F_2 be the smallest block $F_2 \in S_1 \times S_2$ that contains $\{s^1, s^2\}$. Either F_2 is a CURB block or $BR(\Delta(F_2)) \not\subset F_2$, in which case there is at least one sample D of size k from F_2 such that $BR(D) \not\subset F_2$. The probability of sampling that particular D, and the best replies to D being such that at least one of them is not in F_2, is again bounded away from zero. Until we have sampled a sequence of strategy profiles, each extending the set F_i, such that F_i is a CURB block, there is always some sample with positive sampling probability such that $BR(D) \not\subset F_i$. The probability of playing a strategy s^i which is a best reply to

D which is not in F_i, $s^i \in BR(D) \cap (F_i)^c$, is therefore bounded from below by

$$\mathbb{P}\left(\widetilde{BR}(p(t+i-1)) = s^i \mid p(t+i-1)\right) \geq \frac{\left(\beta^{i-1}(1-\beta)\right)^{2k}}{m_1 m_2}(1-\varepsilon)^2.$$

Keep filling $F_i, F_{i+1}, F_{i+2}, \ldots$ with strategies from the CURB block in this fashion, so that F_T spans a CURB block and $T \leq m_1 + m_2$ (Hurkens, 1995, Lemma 1). To get a uniform lower bound, assume that $T = m_1 + m_2$ and that once F_i is a CURB block the following $T - i$ strategy profiles are inside the CURB block. The probability of this progression of plays is bounded from below: let $\mathcal{E}$ be the event that $p(t+T)$ puts at most β^{T+1} mass outside the CURB block spanned by F_T, then

$$\mathbb{P}(\mathcal{E}) \geq \frac{\left(\beta^{2k}\right)^{(T-1)!}(1-\beta)^{2Tk}}{m_1^T m_2^T}(1-\varepsilon)^{2T}.$$

Inside the CURB block spanned by F_T, there is a minimal CURB block which we denote by $C = C_1 \times C_2$. The probability of both players sampling from C given the state $p(t+T)$ (as described above) is greater or equal to

$$\mathbb{P}\left((D_1/k, D_2/k) \in \square(C) \mid D \text{ from } p(t+T)\right) \geq \left(\beta^T(1-\beta)\right)^{2k}(1-\varepsilon)^2.$$

Starting from $p(t) \in B_\delta(C)^c$, a sequence of plays that results in $p(t+T+T^*) \in B_\delta(C)$ is to play T strategies to fill F_T followed by T^* strategies from the minimal CURB block C. Conditional on $p(t) \in B_\delta(\mathscr{C})^c$ and the aforementioned event $\mathcal{E}$, the probability that $p(t+T+T^*) \in B_\delta(C) \subset B_\delta(\mathscr{C})$ is bounded from below by

$$\begin{aligned}&\mathbb{P}\Big((D_1, D_2)(t+T+i) \in \square(C), i = 0, \ldots, T^* - 1 \mid p(t+T) \text{ as above}\Big)\\ &\geq \left(\beta^T(1-\beta)(1-\varepsilon)\right)^{2kT^*} =: \gamma(\varepsilon, T, T^*).\end{aligned}$$

Now $p(t+T+T^*)$ gives at most β^{T^*} probability to all strategy profiles outside $\square(C)$. Therefore, we pick $\delta > 0$ and let $T^* \in \mathbb{N}$ be such that $\beta^{T^*} < \delta$ and, summarizing the analysis in this step, we have derived a bound on the probability of moving from any point $p(t) \in B_\delta(\mathscr{C})^c$ to $B_\delta(\mathscr{C})$ in $T + T^*$ steps. We denote this bound by $\underline{K}$ and it is given by

$$\begin{aligned}&P^{T+T^*}(p(t), B_\delta(\mathscr{C}))\\ &\geq \frac{\left(\beta^{2k}\right)^{(T-1)!}(1-\beta)^{2TK}(1-\varepsilon)^{2T}}{m_1^T m_2^T}\gamma(\varepsilon, T, T^*) =: \underline{K}.\end{aligned}$$

Step 2. Expected exit time from $B_\delta(\mathscr{C})$.
Once in $B_\delta(\mathscr{C})$, one of two things must happen for the process to leave. Either one player makes a mistake or one player samples at least one strategy from outside the minimal CURB block C the process is currently centered around. So instead of calculating the time to the first exit, denoted τ_ε, we calculate the

expected time until one of these two things happen the first time. Let τ_ε^* denote the time, starting from $t = 0$, until either a strategy is sampled outside C or one player makes an ε-tremble. We denote the expression for the probability that $\tau_\varepsilon^* > t^*$, $t^* \in \mathbb{N}$, with $Q_\varepsilon(t^*)$,

$$Q_\varepsilon(t^*) := \mathbb{P}\left(\tau_\varepsilon^* > t^* \mid p(0) \in B_\delta(C)\right) = \prod_{t=0}^{t^*}(1-\beta^t\delta)^{2k}(1-\varepsilon)^2.$$

For the case $\varepsilon = 0$, we use the fact that $\sum_{t=0}^{\infty}\beta^t\delta$ is convergent to conclude that $\prod_{t=0}^{\infty}(1-\beta^t\delta)^{2k}$ approaches a non-zero limit. Since Q_ε is decreasing and non-negative,

$$\lim_{t^*\to\infty} Q_\varepsilon(t^*) = \begin{cases} Q^* \in (0,1), & \text{if } \varepsilon = 0, \\ 0, & \text{if } \varepsilon > 0. \end{cases}$$

We can now derive a bound for τ_ε, the expected time to exit from $B_\delta(\mathscr{C})$,

$$\begin{aligned}
\mathbb{E}\left[\tau_\varepsilon\right] &\geq \mathbb{E}\left[\tau_\varepsilon^*\right] \\
&\geq \mathbb{E}\left[\tau_\varepsilon^* \mid \tau_\varepsilon^* \geq t^*, p(0) \in B_\delta(C)\right] \\
&\qquad \times \mathbb{P}(\tau_\varepsilon^* \geq t^* \mid p(0) \in B_\delta(C))\mathbb{P}(p(0) \in B_\delta(C))) \\
&\geq t^* Q_\varepsilon(t^*)\nu(B_\delta(C)),
\end{aligned}$$

where ν is the initial distribution of the state process and $\nu(B_\delta(C))$ is the probability that $p(0) \in B_\delta(C)$. We know that the state process converges weakly to the invariant distribution for all initial distributions and therefore ν is any distribution on $\square(S)$ of our choice. Choosing ν as the distribution of the constructed $p(t + T + T^*)$ from above,

$$\begin{aligned}
E[\tau_\varepsilon] &\geq t^* \prod_{t=0}^{t^*}(1-\beta^t\delta)^{2k}(1-\varepsilon)^2 \\
&= t^*(1-\varepsilon)^{2t^*}Q_0(t^*) \\
&\geq t^*(1-\varepsilon)^{2t^*}Q^*,
\end{aligned}$$

where t^* is any positive integer. For a fixed ε, the function $t^* \mapsto t^*(1-\varepsilon)^{2t^*}$ is maximized by $t^*(\varepsilon) = -(2\ln(1-\varepsilon))^{-1}$. There is therefore a decreasing sequence of positive numbers $(\varepsilon_j)_{j=1}^{\infty}$, tending to zero as $j \to \infty$, such that $t^*(\varepsilon_j)$ is an integer and

$$\mathbb{E}[\tau_\varepsilon] \geq -\frac{Q^*}{2e\ln(1-\varepsilon_j)},$$

which diverges to ∞ as $j \to \infty$.

Step 3. Bounding $\mu_\varepsilon^*(B_\delta(\mathcal{C})^c)$ from above.
We know that for any $\varepsilon > 0$ there exists a unique invariant probability measure μ_ε^*. We also have a lower bound for $P(x, B_\delta(\mathscr{C}))$ uniform over $x \in B_\delta(\mathscr{C})^c$, and

a lower bound for the expected time the process stays in $B_\delta(\mathscr{C})$ once it has entered.

The probability given by the invariant distribution to the set $B_\delta(\mathscr{C})$ is at least the sum over n of the probability of: the state process not being in it $(n+1)(T+T^*)$ steps ago, but in it $n(T+T^*)$ steps ago, and then staying there for at least $n(T+T^*)$ time steps,

$$\begin{aligned}
1 \geq \mu_\varepsilon^*(B_\delta(\mathscr{C})) &\geq \sum_{n=0}^{\infty} \left(\int_{B_\delta(\mathscr{C})^c} P^{T+T^*}\left(x, B_\delta(\mathscr{C})\right) d\mu_\varepsilon^*(x) \right) \mathbb{P}\left(\tau_\varepsilon \geq n(T+T^*)\right) \\
&\geq \mu_\varepsilon^*\left(B_\delta(\mathscr{C})^c\right) \underline{K} \left(\sum_{n=0}^{\infty} \mathbb{P}\left(\frac{\tau_\varepsilon}{T+T^*} \geq n \right) \right) \\
&\geq \mu_\varepsilon^*\left(B_\delta(\mathscr{C})^c\right) \frac{\underline{K}}{T+T^*} \, \mathbb{E}\left[\tau_\varepsilon^*\right].
\end{aligned}$$

Step 4. Putting it all together.
The collection $(\mu_\varepsilon^*)_{\varepsilon>0}$ is tight because $\square(S)$ is compact. So there exists a subsequence that converges weakly to $\mu^* \in \mathcal{P}(\square(S))$. The limit μ^* is not necessarily unique, however, by the Portmanteau theorem,

$$\liminf_{\varepsilon \to 0} \mu_\varepsilon^*(U) \geq \mu^*(U)$$

for all open sets U of $\square(S)$. Note that $B_\delta(\mathscr{C})^c$ is open, and

$$\mu_\varepsilon^*(B_\delta(\mathscr{C})^c) \leq \frac{T+T^*}{\underline{K}\mathbb{E}[\tau_\varepsilon^*]}.$$

Since $\underline{K} > 0$ increases as $\varepsilon \to 0$, $\mathbb{E}\left[\tau_\varepsilon^*\right] \to \infty$ as $\varepsilon \to 0$, and $T+T^*$ does not depend on ε,

$$\mu^*(B_\delta(\mathscr{C})^c) \leq \liminf_{\varepsilon \to 0} \mu_\varepsilon^*(B_\delta(\mathscr{C})^c) \leq (T+T^*) \liminf_{\varepsilon \to 0} \frac{1}{\underline{K}\mathbb{E}[\tau^*]} = 0.$$

We conclude that that $\mu_\varepsilon^*\left(B_\delta(\mathscr{C})\right) \to 1$ as $\varepsilon \to 0$. □

B Concentration around approximate Nash equilibrium: proofs

Parts of this appendix relies on the assumption that the game is of size 2×2 and has a unique mixed Nash Equilibrium. Generically, all 2×2 games without pure Nash equilibria must have the basic Matching Pennies structure. One player will be 'agreeing' and the other 'disagreeing' in the sense that the best reply of the agreeing player is to play the same strategy (0 or 1) as the disagreeing player. On the other hand, the disagreeing player's best reply is to not play the same strategy as the agreeing player. Any other situation will generically yield at least one pure equilibrium, and generically a strict pure equilibrium.

B.1 Unique fixed point to the expected best reply

Lemma 17. *Let G be a 2×2 game with a unique mixed Nash equilibrium N^* and let k, the number of samples, be an integer such that $N_1^* k \notin \mathbb{N}$ and $N_2^* k \notin \mathbb{N}$. Then there exists a unique fixed point $n^* = (n_1^*, n_2^*) \in int(\square(S))$ to the system*

$$\begin{cases} \mathbb{E}\left[\widetilde{BR}_1(n_2^*)\right] = n_1^*, \\ \mathbb{E}\left[\widetilde{BR}_2(n_1^*)\right] = n_2^*. \end{cases} \tag{14}$$

Proof. We will refer to the player 1 and 2 as the agreeing and the disagreeing player, respectively. The Nash equilibrium $N^* = (N_1^*, N_2^*)$ defines the 'cut-off' $M_i := \lfloor N_i^* k \rfloor$, $i = 1, 2$. The cut-off is such that if more than M_1 of the agreeing player's k samples from the disagreeing player's history are 1, he plays 1. The disagreeing player will play strategy 1 if more than M_2 of his k samples from the agreeing player's history of plays are 0. Consider the function

$$p_{k,M}(x) := (1-\varepsilon) \sum_{i=M+1}^{k} \binom{k}{i} x^i (1-x)^{k-i} + \varepsilon/2.$$

Given that player history is in state (a, d), the probability that the agreeing and disagreeing player plays strategy 1 is $p_a(d) := p_{k,M_2}(d)$ and $p_d(a) := 1 - p_{k,M_1}(a)$, respectively. We can now rewrite (14) as

$$p_a(n_2^*) = n_1^*, \qquad p_d(n_1^*) = n_2^*.$$

The range of p_a and p_b is $I_\varepsilon := [\varepsilon/2, 1-\varepsilon/2]$. Therefore, by the strict monotonicity and the continuity of p_a and p_d, we may rewrite (14) again, now as

$$\begin{aligned} (p_a \circ p_d)(n_1^*) &= n_1^*, \quad n_1^* \in I_\varepsilon, \\ (p_d \circ p_a)(n_2^*) &= n_2^*, \quad n_2^* \in I_\varepsilon. \end{aligned}$$

Note that since p_a and p_d are strictly increasing and decreasing, respectively, both $p_a \circ p_d$ and $p_d \circ p_a$ are strictly decreasing functions from $[0, 1]$ to $[p_d(1 - \varepsilon/2), p_d(\varepsilon/2)]$ and $[p_a(\varepsilon/2), p_a(1 - \varepsilon/2)]$, respectively. Therefore

$$\begin{aligned} \min\{p_a \circ p_d(\varepsilon/2), p_d \circ p_a(\varepsilon/2)\} &\geq \min\{p_d(1-\varepsilon/2), p_a(\varepsilon/2)\} > \varepsilon/2, \\ \max\{p_a \circ p_d(1-\varepsilon/2), p_d \circ p_a(1-\varepsilon/2)\} &\leq \max\{p_d(\varepsilon/2), p_a(1-\varepsilon/2)\} < 1 - \varepsilon/2. \end{aligned}$$

Hence, since $p_a \circ p_d$ and $p_d \circ p_a$ are continuous, they intersect the straight line $x = y$ at a (function-wise) unique point in their respective images and these intersection points are n_1^* and n_2^*. □

B.2 Global exponential stability of mean-field dynamics

Denote by ξ the solution mapping of $\dot{x}(t) = F(x(t))$, $x(0) = p$, where $F(x) := \mathbb{E}[\widetilde{BR}(x)] - x$. Then

$$\xi(t,p) = p + \int_0^t F(\xi(s,p))ds.$$

Lemma 18. *Let Σ contain all points $x \in \square(S)$ such that $F(x) = 0$ or such that $\xi(t,x)$ satisfies $(\xi(t,x) - y)^* F(\xi(t,x)) = 0$ for all $t \geq 0$ and some y, such that $F(y) = 0$. The mapping $t \mapsto \xi(t,p)$ is globally asymptotically stable, with $\lim_{t\to\infty} \xi(t,p) \in \Sigma$. Furthermore, if the game is 2×2 with a unique mixed Nash equilibrium, then $\Sigma = \{n^*\}$, the unique root of F.*

Proof. Let $V(x) := \frac{1}{2}\|x - n^*\|_2^2$ where n^* is a root of F. The existence of n^* is granted by Brouwer's fixed point theorem; $\square(S)$ is compact and convex and F is continuous. Differentiating V with respect to time at the solution mapping $\xi(t,p)$, we get

$$\begin{aligned}
-\dot{V}(\xi(t,p)) &= -\nabla V(\xi(t,p))\dot{\xi}(t,p) \\
&= -(\xi(t,p) - n^*)^T F(\xi(t,p)) \\
&= -(\xi(t,p) - n^*)^T \left(\mathbb{E}[\widetilde{BR}(\xi(t,p)) \mid \xi(t,p)] - \xi(t,p)\right) \\
&= 2V(\xi(t,p)) - (\xi(t,p) - n^*)^T \left(\mathbb{E}[\widetilde{BR}(\xi(t,p) \mid \xi(t,p)] - n^*\right) \\
&= V(\xi(t,p)) - V(\mathbb{E}[\widetilde{BR}(\xi(t,p)) \mid \xi(t,p)]) \\
&\quad + \frac{1}{2}\|\xi(t,p) - \mathbb{E}[\widetilde{BR}(\xi(t,p)) \mid \xi(t,p)]\|_2^2
\end{aligned}$$

where in the last step we used the identity $2y^T z = \|y\|_2^2 + \|z\|_2^2 - \|y - z\|_2^2$, $y, z \in \mathbb{R}^d$. We notice that

$$\begin{aligned}
&V(\mathbb{E}[\widetilde{BR}(\xi(t,p))\|\xi(t,p)]) \\
&= \frac{1}{2}\|\mathbb{E}[\widetilde{BR}(\xi(t,p)) \mid \xi(t,p)] - \xi(t,p) + \xi(t,p) - n^*\|_2^2 \\
&\leq \frac{1}{2}\|\mathbb{E}[\widetilde{BR}(\xi(t,p)) \mid \xi(t,p)] - \xi(t,p)\|_2^2 + V(\xi(t,p)),
\end{aligned}$$

hence $\dot{V}(\xi(t,p)) \leq 0$. Furthermore, V is radially unbounded. Let $R := \{x \in \square(S) : (x - n^*)^T F(x) = 0\}$, then $R = \{x \in \square(S) : \dot{V}(x) = 0\}$ and R contains n^*, any other point solution to $F(x) = 0$, and all x such that the vectors $(x - n^*)$ and $F(x)$ are orthogonal. By a global invariant set theorem (Slotine, Li et al., 1991, Thm. 3.5), $\xi(t,p)$ converges to the largest invariant set of R, which is Σ.

Next, for 2×2 games with a unique mixed Nash equilibrium, we show the points in R different from n^* (now unique) cannot be in Σ. First note that if

$x \in R\backslash\{n^*\}$, then $x_i \neq n_i^*, i = 1, 2$. Without loss of generality, assume that player 2 has the disagreeing role and that $x_0 > n^*$. If $x_0 \in R\backslash\{n^*\}$ then $F(x_0) \neq 0$ and a trajectory starting in x_0 will evolve according to the dynamic system $\dot{x}(t) = F(x(t))$, $x(0) = x_0$. Assume, towards a contradiction, that $x(t) \in R\backslash\{n^*\}$ for all $t \geq 0$. After some finite positive time, call it t^*, the path must cross the line $(x, n_2^*; x \in [0, 1])$ (because the trajectory starts at at $x_0 > n^*$ and player 2 is disagreeing, it will move "south-east" in $\square(S)$). This crossing contradicts $x(t^*) \in R\backslash\{n^*\}$ since $x(t^*) \in R\backslash\{n^*\}$ would require both components of $x(t^*)$ to be different from n^*. The same argument can be carried out for all other possible initial positions ($x_0 - n^* < 0$ or mixed signs) and for switched player roles. It follows that $\{n^*\}$ is the only invariant set in R. □

B.3 Trajectories over bounded time intervals

By (Benaïm and Weibull, 2003, Lemma 1), the state process $p(\cdot)$ and its mean-field approximation $\xi(\cdot, p(0))$ lie close to each other (over bounded time intervals) with high probability. We have to do one modification to apply the result: we re-scale size of the time steps taken by our learning process. This has no effect on previous results since we will always (for a fixed β) have a fixed positive step size. The original proof of Benaïm and Weibull (2003) can be used to prove the lemma below.

Lemma 19. *Scale the step size of t by $(1 - \beta)$. Let $T = N(1 - \beta)$ for some $\mathbb{N} \in \mathbb{N}$ and let $(\hat{p}(t); t \in [0, T])$ be the linear interpolation of the path $(p(t); t = 0, 1 - \beta, \ldots, (1 - \beta)N)$. Then, for all $\eta > 0$,*

$$\mathbb{P}\left(\max_{t\in[0,T]} \|\hat{p}(t) - \xi(t, p(0))\|_\infty \geq \eta\right) \leq 2(m_1 + m_2 - 2)e^{-\eta^2 c}$$

where c is a positive constant and proportional to $e^{-\gamma T}(T(1-\beta))^{-1}$, where $\gamma > 0$ depends only on the size of the game.

B.4 Proof of Theorem 6

Let $t \geq s \geq 0$. Below, K will denote a generic positive constant. Whenever $\eta > \|\xi(t, \hat{p}(t - s)) - \xi(t, p(0))\|_\infty$, Lemma 19 yields that

$$\begin{aligned}
&\mathbb{P}(\|\hat{p}(t) - \xi(t, 0)\|_\infty \geq \eta) \\
&= \mathbb{P}\left(\|\hat{p}(t) - \xi(t, \hat{p}(t - s))\|_\infty \geq \eta - \|\xi(t, \hat{p}(t - s) - \xi(t, p(0))\|_\infty\right) \\
&\leq K \exp\left(-(\eta - \|\xi(t, \hat{p}(t - s)) - \xi(t, p(0))\|_\infty)^2 K \frac{e^{-\gamma s}}{s(1 - \beta)}\right).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\mathbb{P}(\|\hat{p}(t) - n^*\|_\infty \geq \eta) \\
&= \mathbb{P}\left(\|\hat{p}(t) - \xi(t, p(0))\|_\infty \geq \eta - \|\xi(t, p(0)) - n^*\|_\infty\right),
\end{aligned}$$

so we have that

$$\begin{aligned}
&\mathbb{P}(\|\hat{p}(t) - n^*\|_\infty \geq \eta) = \mathbb{P}\Big(\|\hat{p}(t) - \xi(t, \hat{p}(t-s))\|_\infty \geq \eta \\
&- \|\xi(t, \hat{p}(t-s)) - \xi(t, p(0))\|_\infty - \|\xi(t, p(0)) - n^*\|_\infty\Big) \\
&\leq K \exp\Bigg(- (\eta - \|\xi(t, \hat{p}(t-s)) - \xi(t, p(0))\|_\infty - \|\xi(t, p(0)) - n^*\|_\infty)^2 \\
&\qquad\qquad \times K \frac{e^{-\gamma s}}{s(1-\beta)}\Bigg).
\end{aligned}$$

Letting $t \to \infty$, we know from Lemma 18 that $\xi(t, p(0)) \to n^*$, so

$$\begin{aligned}
&\lim_{t\to\infty} \mathbb{P}(\|\hat{p}(t) - n^*\|_\infty \geq \eta) \\
&\leq \sup_{x \in \square(S)} K \exp\left(-(\eta - \|\xi(s,x) - n^*\|_\infty)^2 K \frac{e^{-\gamma s}}{s(1-\beta)}\right).
\end{aligned}$$

Choosing σ large enough, so that for all $s \geq \sigma$: $\|\xi(s,x) - n^*\|_\infty \leq \eta/2$ uniformly in x. Then

$$\lim_{t\to\infty} \mathbb{P}\left(\|\hat{p}(t) - n^*\|_\infty^2 \geq \eta\right) = o\left(\exp\left(-\frac{K\eta^2}{1-\beta}\right)\right),$$

proving the theorem.