SIMULTANEOUS BEST-RESPONSE DYNAMICS IN RANDOM POTENTIAL GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper examines the convergence behaviour of simultaneous best-response dynamics in random potential games. We provide a theoretical result showing that, for two-player games with sufficiently many actions, the dynamics converge quickly to a cycle of length two. This cycle lies within the intersection of the neighbourhoods of two distinct Nash equilibria. For three players or more, simulations show that the dynamics converge quickly to a Nash equilibrium with high probability. Furthermore, we show that all these results are robust, in the sense that they hold in non-potential games, provided the players' payoffs are sufficiently correlated. We also compare these dynamics to gradient-based learning methods in near-potential games with three players or more, and observe that simultaneous best-response dynamics converge to a Nash equilibrium of comparable payoff substantially faster.

1 INTRODUCTION

Strategic interactions between agents are typically modelled as games, with the *Nash equilibrium* (NE) serving as the central solution concept. However, this concept requires strong assumptions, including player rationality and, in the presence of multiple equilibria, a principled method for equilibrium selection (Aumann and Brandenburger (1995), Harsanyi and Selten (1988)). Moreover, computing a NE often requires knowledge of the opponents' payoffs. These requirements are rarely met in practice, where often solutions are instead found heuristically.

The increasing use of learning agents to address complex optimisation problems raises a central question in game theory and multi-agent learning: **do learning agents interacting repeatedly converge to a NE?** Such agents adapt their strategies through learning rules designed to improve individual rewards, thereby generating a dynamic process over the space of strategy profiles. These dynamics are called learning dynamics or adaptive dynamics.

A particularly intuitive and widely studied class of learning dynamics is the *best-response dynamics* (see, e.g. Swenson et al. (2018b)), in which players update their strategies to their best response against the current strategy profile of their opponents. There are two primary variants. In the **sequential** variant, players revise their strategies one at a time, either in a fixed order or according to a stochastic rule (Swenson et al. (2018b)). In contrast, the **simultaneous** variant (from now on SBRD) has all players updating their strategies simultaneously, and is the focus of this work. While the sequential variant requires coordination on the update order, the simultaneous version does not.

Moreover, SBRD is a *deterministic uncoupled dynamic*, meaning that each player updates their strategy based solely on their own payoffs, without knowledge of others' payoffs or any need for coordination. While it is known that dynamics of this family do not always converge to NE in general games (Hart and Mas-Colell (2003)), this paper investigates the behaviour of SBRD in the specific context of *random potential games*.

Potential games have been extensively employed to model a variety of strategic environments, including congestion games Voorneveld et al. (1999); Sandholm (2010), Cournot Competition Dragone et al. (2012), and they have applications in theoretical computer science Nisan et al. (2007), wireless communication Lasaulce and Tembine (2011) and evolutionary biology Hofbauer and Sigmund (2003), among others. Potential games, introduced by Rosenthal (1973) and further developed by Monderer and Shapley (1996) and Voorneveld (2000), are games in which there exists a common potential function, mapping action profiles to real or ordinal values, such that each player's optimisation aligns with the optimisation of this global function. In other words, although players optimise their own payoffs, unilateral deviations correspond to changes in a common potential function, so their incentives are aligned through this shared structure- even if maximising the potential does not necessarily maximise any individual player's payoff. In order to estimate typical behaviour, we consider random potential games, i.e. we sample randomly the potential game in which the dynamic occurs.

The simplicity of SBRD, coupled with the broad applicability of potential games, motivates our central question:

If all players in a potential game follow a simultaneous best-response rule,
is the resulting dynamic likely to converge to a Nash equilibrium?

Our results indicate that this is the case for three or more players. However, interestingly, we prove that this is not the case for two-player games. Moreover, we show that the same phenomena occur if the potential game assumption is weakened.

Our goal is to estimate the probability with which SBRD converges to a NE in random potential games. The model relies on two assumptions: (i) the values of the potential function are sampled independently for each action profile from a common distribution. We use the normal distribution in our experiments, although our theoretical results only require sampling from any continuous distribution; and (ii) all players have the same number of actions. This assumption is made purely for notational convenience and computational simplicity, rather than necessity.

Under these conditions, the resulting probability distribution over best-response trajectories is equivalent to that induced by uniformly sampling an ordering of the action profiles Collevecchio et al. (2024). This equivalence allows us to study convergence behaviour within the broader framework of random ordinal potential games (see again Collevecchio et al. (2024) for further discussion).

Paper outline and summary of results We obtain results in three directions. *First*, we characterise the limiting behaviour of SBRD in random potential games as the number of actions per player increases. We do so by providing a formal proof for the two-player case and by giving strong numerical evidence for the cases with three or more players. To our knowledge, ours is the first theoretical result of its kind. *Second*, we verify the robustness of our results by numerically testing whether similar behaviour holds in games that are 'close' to potential games —specifically, games with highly correlated payoffs. *Third*, we compare SBRD with the widely used and well-understood (Zhang et al. (2022)) softmax policy gradient dynamic, examining both convergence rates and the quality of the resulting payoffs. We now elaborate further.

Firstly, in Section 3.1, we reveal an interesting difference between games with two players and games with at least three players. We prove, for random potential games with two players, that with high probability SBRD ends up cycling over a cycle of length two, and thus, not converging to a NE. To the best of our knowledge, no theoretical result analysing convergence of SBRD has been obtained before ours. Furthermore, the convergence to the cycle takes place within a constant number of steps, with a small proportion of the action-profiles being played. This two-cycle consists of two action-profiles (a, b) and (a', b') such that both (a, b') and (a', b) are NE. These results are presented in Section 3.1 and experiments in Section 4.2. For random potential games with at least three players, we find in Section 4.3 that as the number of actions increases, the probability with which SBRD converges to a NE increases to one.

Secondly, throughout Section 4, we numerically test the robustness of our results to the assumption of the game being a potential game. We simulate random games with various levels of correlation for the payoffs of the players, and find that, in the highly-correlated regime, the results obtained for potential games still hold. With this, we provide strong evidence that highly correlated games behave similarly to potential games with respect to SBRD.

Thirdly, also in Section 4.4, we compare SBRD to the softmax policy gradient dynamic (SPGD). We choose SPGD as our benchmark due to its desirable combination of properties: it updates in the direction of the best response while introducing smoothness to the learning dynamics, enjoys strong theoretical convergence guarantees, is well-suited for practical model-free implementation, and incorporates inherent exploration. These features have led to the widespread adoption of softmax policy gradient methods, and their variants, in contemporary reinforcement learning (Mei et al. (2020), Klein et al. (2023), Bernasconi et al. (2025), Chen et al. (2022) and Shi et al. (2019)). We observe that, for three or more players near-potential games, SBRD converges significantly quicker to an equilibrium and scales better to large action sets. We also find that, while SPGD tends to converge to equilibria with moderately higher payoffs, the average payoff along the dynamics is higher for SBRD. The case of three or more players is presented in Section 4.

In summary, we show that SBRD cycles around two NE in the case of two players, and converges to a NE in the case of three or more players. We show that this happens quickly, and is robust, meaning that the same holds for games with highly correlated payoffs. Hence, SBRD is a quick and highly-rewarding learning method.

Related work Our research is closely related to two branches of research: learning dynamics in potential games, and learning dynamics in games with random payoffs.

Learning dynamics in potential games In recent years, learning dynamics in potential games, and their Markovian extensions, have been extensively studied. Convergence guarantees are of interest: for instance, Sakos et al. (2024) analyse q-replicator dynamics, Heliou et al. (2017) prove convergence under no-regret learning with the exponential weights algorithm and minimal information, and Fox et al. (2022) show convergence for natural policy gradient learning. Other works focus on the complexity of these dynamics: Leonardos et al. (2021) study projected gradient dynamics, Cen et al. (2022) analyse softmax policy gradient descent with entropy regularisation, Zhang et al. (2022) examine gradient and natural gradient with log-barrier regularisation, Ding et al. (2022) consider projected gradient under various informational assumptions, and Sun et al. (2023) investigate natural policy gradient descent methods. More recent contributions include Dong et al. (2024), who analyse a variant of the Frank-Wolfe algorithm, and Alatur et al. (2024), who study independent policy mirror descent.

Learning dynamics in games with random payoffs The behaviour of learning dynamics in games with randomly generated payoffs has been the subject of increasing interest. Rather than studying specific game instances that may be cherry-picked or have special structure, random games let us understand what happens in 'generic' cases - crucial for assessing when algorithms will work reliably in practice and common method in the literature (see Durand and Gaujal (2016)). In the two-player setting, Galla and Farmer (2013) show that experience-weighted attraction learning can lead to a range of outcomes, from convergence to fixed points to complex chaotic behaviour. Chan et al. (2025) generalises these results in the many player limit. Assuming the ability for players to coordinate, Mimun et al. (2024) demonstrate that, under payoff correlation and a growing number of actions, sequential best-response dynamics converge to a pure NE with high probability. In a similar setting, Collevecchio et al. (2024) study two-player random potential games and show that the basin of attraction of each equilibrium is effectively determined by the identity of the player that first updates their strategy. Durand and Guajal Durand and Gaujal (2016) consider sequential best response dynamics in random potential games and obtain worse case and average complexity results.

In games with many players or actions, structural properties of the dynamics are nuanced. Amiet et al. (2021b) examine large-player games where each player has two actions and payoffs are randomly drawn with a small probability of ties; they show that sequential best-response dynamics typically reach a pure NE as the number of players grows. Amiet et al. (2021a) contrast best and better-response dynamics in two-player games with many actions, finding that while better-response dynamics (with randomly selected updating players) reliably converge to equilibrium when one exists, best-response dynamics tend to enter cycles. This sensitivity to update rules is further emphasised by Heinrich et al. (2023), who show that sequential best-response dynamics converge only under random turn-taking; cyclic update orders generally fail to reach equilibrium. Finally, Johnston et al. (2023) prove that in large random games, any non-equilibrium action profile can be connected via a best-response path to a pure equilibrium, if one exists, with high probability as the action space grows.

We notice that SBRD is sensitive to perturbations of the payoff matrix. This means that for all $\varepsilon > 0$, there is a positive probability that by an ε-perturbation of payoff we obtain a different trajectory than the non-perturbed one. We point out however that this is not always the case, and that a small perturbation of most of the payoffs would typically not influence the trajectory.

2 THE SETUP

An n-player normal-form game is a triple $(N, (A_i)_{i\in N}, (u_i)_{i\in N})$, where $N = \{1, \ldots, n\}$ is a finite set of players, each A_i is the finite action set of player i, and $u_i : \prod_{j\in N} A_j \to \mathbb{R}$ is the payoff function of player i. For ease of exposition, we assume that all players have the same number m of actions, i.e. for all players $i \in N$ and some $m \in \mathbb{N}$, we have $|A_i| = m$. Let $A = \prod_{i\in N} A_i$ denote the set of action profiles, and, for $i \in N$, let $A_{-i} := \prod_{j\in N\setminus\{i\}} A_j$.

Players may randomise over their actions by playing a strategy $x_i \in \Delta(A_i)$, where $\Delta(A_i)$ is the simplex over the set A_i. It is standard to extend the payoff function u_i to strategy profiles. And so, for $x = (x_1, \ldots, x_n) \in \prod_{i\in N} \Delta(A_i)$, the expected payoff to player i is

$$u_i(x) = \sum_{a\in A} u_i(a) \prod_{j=1}^{n} x_{j,a_j} \,. \tag{1}$$

With a slight abuse of notation, we sometimes write $a = (a_i, a_{-i})$ and $x = (x_i, x_{-i})$ to denote the combination of player i's action or strategy with the actions or strategies of their opponents. A strategy profile $x^* \in \prod_{i \in N} \Delta(A_i)$ is a NE if there are no profitable unilateral deviations from x^*, namely, if for any player $i \in N$ and any strategy $x_i \in \Delta(A_i)$ of this player, it holds that $u_i(x_i^*, x_{-i}^*) \geq u_i(x_i, x_{-i}^*)$. A strategy profile x is *pure* if each player plays one action with probability 1. In this case, we often refer to it as an *action profile*.

2.1 POTENTIAL GAMES

A game is a potential game if there exists a single function $\Psi \colon A \to \mathbb{R}$ ('the potential') that captures the players' incentives. Formally,

Definition 2.1. *A normal-form game* $(N, (A_i)_{i \in N}, (u_i)_{i \in N})$ *is a potential game if there is a function* $\Psi : A \to \mathbb{R}$ *such that for each player* $i \in N$ *and each possible action profile of the opponents* $a_{-i} \in \prod_{j \neq i} A_j$*, there is a constant* $c_i(a_{-i}) \in \mathbb{R}$ *such that, for every* $a_i \in A_i$*, we have,*

$$u_i(a_i, a_{-i}) = \Psi(a_i, a_{-i}) + c_i(a_{-i}) \,,$$

When all c_i *are equal to zero, we simply write* $(N, (A_i)_{i \in N}, \Psi)$*.*

This is equivalent to the classical definition of a potential game (see Monderer and Shapley (1996)).

Thus, a change in player i's payoff from switching actions exactly equals the change in the global potential. Consequently, in a potential game, an action profile is a NE if and only if it is a local maximum of the potential function Ψ.

Without loss of generality, here and in the following we assume all the c_i are equal to 0. This is not restrictive in our setting as in SBRD players only considers pairwise comparisons of rewards.

2.1.1 RANDOM POTENTIAL GAMES

To study typical behaviour, we introduce the notion of random potential game with n players and m actions.

Definition 2.2. *Let* F *be a continuous real-valued distribution, and* n *and* m *positive integers. An* n*-player* m*-actions* F*-random potential game is a potential game* $G = (N, (A_i)_{i \in N}, \Psi)$ *in which* $|N| = n$*, and* $|A_i| = m$*, and moreover we have that for each* $a \in A$*, the value* $\Psi(a)$ *is sampled independently at random from* F*.*

When N, A *and* F *are clear from context, we just refer to* G *as a random potential game.*

2.2 THE SIMULTANEOUS BEST RESPONSE ALGORITHM

One of the simplest learning dynamics is the simultaneous best response dynamic (SBRD). Given a game, starting from an initial action profile $a^0 \in A$, SBRD proceeds as follows: at each round $t \geq 1$ every player $i \in N$ myopically best-responds to the previous action profile a^{t-1}. Formally,

$$a_i^t = \arg\max_{a_i \in A_i} u_i(a_i, a_{-i}^{t-1}) \,. \tag{2}$$

If, at some time t, we have $a^t = a^{t+1}$, then every player must be playing a best-response to their opponents' strategies, which means a^t is a NE.

We can assume without loss of generality that a^0 is some arbitrary fixed action profile, up to reordering. Once a^0 is fixed, since best-response update depends only on the realised potential function Ψ, the sequence $(a^t)_{t \geq 0}$ is a random process.

3 RESULTS

In this section, we present our main findings on the convergence of SBRD in random potential games. We begin by establishing a theoretical result for two-player games. We show that, for large enough number of actions, SBRD almost surely reaches a two-cycle in a constant number of steps. This two-cycle consists of two action-profiles (a, b) and (a', b') such that (a, b') and (a', b) are both NE. We then consider the case of three players or more. Here, we demonstrate via simulations that SBRD converges to a pure NE with probability tending to one as A tends to infinity.

3.1 TWO PLAYERS

Our main theoretical result is that, in two-player games with sufficiently large action sets, SBRD almost surely converges to a two-cycle in a constant number of steps.

Theorem 3.1. *Let* $\varepsilon \in (0,1)$*,* F *be a continuous real distribution, and* G *be a two-player* m*-actions* F*-random potential game. If* m *is large enough, then SBRD converges to a two-cycle in at most* $\frac{\log \varepsilon}{\log(3/4)}$ *steps with probability at least* $1-\varepsilon$*.*

The main steps of the proof are explained below. All lemmas are proved in Appendix A. The proof of Theorem 3.1 works by comparing the SBRD to another dynamic that converges to a two-cycle, and showing that these two processes coincide up to the termination time with high probability.

We view the SBRD as a random process over the set of action profiles, where only the payoffs needed are sampled at each time. In this sense, the SBRD for two players proceeds as follows:

- At period 0, the initial action profile $(a^0, b^0) \in A$ is arbitrarily chosen. Also, the following payoffs are sampled (i.i.d. from F): $\Psi(a^0, b^0)$, $\Psi(a, b^0)$ for $a \in A_1 \setminus \{a^0\}$, and $\Psi(a^0, b)$ for $b \in A_2 \setminus \{b^0\}$.
- At period 1, the current action profile is $(a^1, b^1) \in A$ where $a^1 := \arg\max_{a \in A_1} \Psi(a, b^0)$, $b^1 := \arg\max_{b \in A_2} \Psi(a^0, b)$. As the realised potential values are drawn from a continuous distribution, ties occur with probability zero, and best responses are almost surely unique. Furthermore, the following payoffs are sampled independently from F (if they have not already been sampled): $\Psi(a^1, b^1)$, $\Psi(a, b^1)$ for $a \in A_1 \setminus \{a^1\}$, and $\Psi(a^1, b)$ for $b \in A_2 \setminus \{b^1\}$.
- In general, at period t, the current action profile is $(a^t, b^t) \in A$ where $a^t := \arg\max_{a \in A_1} \Psi(a, b^{t-1})$, and $b^t := \arg\max_{b \in A_2} \Psi(a^{t-1}, b)$. Additionally, the following payoffs are sampled independently from F (if they have not already been sampled before): $\Psi(a^t, b^t)$, $\Psi(a, b^t)$ for $a \in A_1 \setminus \{a^t\}$, and $\Psi(a^t, b)$ for $b \in A_2 \setminus \{b^t\}$.
- This process terminates when there is a repetition, i.e. if at some time T there exists some earlier time $s < T$ such that $(a^T, b^T) = (a^s, b^s)$, then the process terminates at time T in a cycle of length $T - s$. Since the action space is finite, the process must eventually cycle and thus terminate.

The first step of our proof is to observe that no cycle of length greater than two can occur.

Lemma 3.2. *With probability one, the SBRD process terminates at a cycle of length one or two.*

We can further characterise the two-cycle as follows:

Remark 3.3. *Suppose that the SBRD process does not converge to a NE. By Lemma 3.2, there exists some time* T*, such that* $(a^{T-2}, b^{T-2}) = (a^T, b^T)$*. Consider the two action profiles* (a^{T-1}, b^T) *and* (a^T, b^{T-1})*. As* $b^T = b^{T-2}$*, we have that* a^{T-1} *is a best response of player one to* b^T*, and clearly* b^T *is a best-response of player two to* a^{T-1}*. Hence,* (a^{T-1}, b^T) *is a NE. Similarly,* (a^T, b^{T-1}) *is also a NE.*

The rest of the proof works by comparing the SBRD with a restricted version of our dynamic, which we call the Independent Dynamic (INDD). In INDD, at each time t, players do not necessarily move to the current best response but rather select the best response amongst the actions they have not yet played, or the action they played in the previous period. While counter-intuitive, because m is large and INDD quickly converges, the set of actions excluded is insignificant compared to the whole set of available actions and therefore the dynamics behave in the same way with high probability.

The reason we consider INDD is that, in this dynamic, at each time t, each player's next action is chosen as the maximiser of a set of potential values that are either independent of the history of the process or whose dependence can be carefully controlled. In contrast, under SBRD, any previously sampled payoff that was not the maximiser at the time it was observed becomes less likely to be the maximiser at a later time. This introduces a form of path dependence, thereby breaking the independence structure of the process.

Formally, the INDD is defined as follows.

- At time 0, the initial action profile is (a^0, b^0), and the following payoffs are sampled (i.i.d. from F):

  $$\left\{\Psi(a, b^0) : a \in A_1 \setminus \{a^0\}\right\} \quad \text{and} \quad \left\{\Psi(a^0, b) : b \in A_2 \setminus \{b^0\}\right\}.$$

  Note that the value $\Psi(a^0, b^0)$ is not sampled.

- At time 1, the action profile is (a^1, b^1) where:
$$a^1 := \operatorname*{arg\,max}_{a \in A_1 \setminus \{a^0\}} \Psi\big(a, b^0\big) \qquad \text{and} \qquad b^1 := \operatorname*{arg\,max}_{b \in A_2 \setminus \{b^0\}} \Psi\big(a^0, b\big).$$
Furthermore, all payoffs of the form $\Psi(a, b^1)$ and $\Psi(a^1, b)$ that are not known yet are sampled, besides $\Psi(a^1, b^1)$. Note that the set of payoffs for player one that need to be sampled is $R_1^1 := \{\Psi(a, b^1), a \notin \{a^\tau, \tau < 1\}\} = \{\Psi(a, b^1), a \neq a^0\}$ and likewise for player two it is $R_2^1 := \{\Psi(a^1, b), b \notin \{b^\tau, \tau < 1\}\} = \{\Psi(a^1, b), b \neq b^0\}$.
- At time $t \geq 2$, the action profile is (a^t, b^t) where
$$a^t = \operatorname*{arg\,max}_{a \in \{a^{t-2}\} \cup (A_1 \setminus \{a^\tau : \tau < t\})} \Psi(a, b^{t-1}), \tag{3}$$
$$b^t = \operatorname*{arg\,max}_{b \in \{b^{t-2}\} \cup (A_2 \setminus \{b^\tau : \tau < t\})} \Psi(a^{t-1}, b). \tag{4}$$
Additionally, all payoffs of the form $\Psi(a, b^t)$ and $\Psi(a^t, b)$ that are not known yet are sampled, besides $\Psi(a^t, b^t)$. The set of payoffs for player one that need to be sampled is $R_1^t := \{\Psi(a, b^t), a \notin \{a^\tau, \tau < t\}\}$ and likewise for player two it is $R_2^t := \{\Psi(a^t, b), b \notin \{b^\tau, \tau < t\}\}$.
- We define this process to terminate when there is a repetition, i.e. if at some time T there exists some earlier time $s < T$ such that $(a^T, b^T) = (a^s, b^s)$. Then we say that the process terminates at time T in a cycle of length $T - s$. Since the action space is finite, the process must eventually terminate.

In formal statements, we refer to this process as a two-player m-actions F-INDD.

Since, at time t, each player can only play the action that was played at time $t-2$ or one of the actions that they have not played before, the only cycles that can occur are of length two. As the set of action profiles is finite, INDD must cycle, and thus INDD must converge to a cycle of length two.

The dynamics INDD and SBRD are different only if in SBRD one of the players plays at time t an action that they already played at time s with $s \neq t-2$.

We argue that this occurs with small probability. To this end, we prove first that the INDD process terminates quickly with a high probability.

Lemma 3.4. *Let $\varepsilon \in (0,1)$, and F a continuous real-valued distribution. Let us consider a two-player m-actions F-INDD. If m is large enough, then the probability that the INDD process has not terminated by period $\frac{\log \varepsilon}{\log(3/4)}$ is at most ε.*

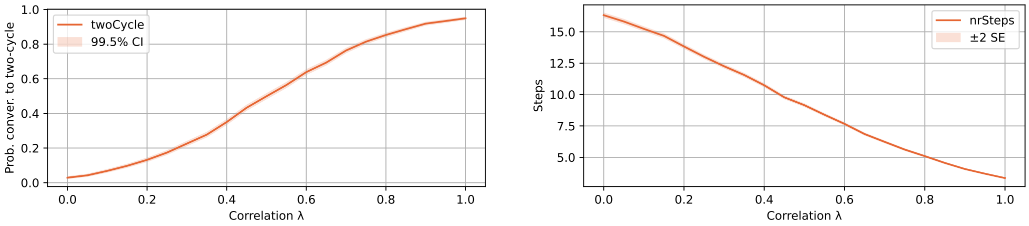
We then show that the probability that INDD and SBRD differ at any step tends to zero as $m \to \infty$. A difference between SBRD and INDD can only appear if, for some time t, the best response of player one to b^{t-1} is either a^{t-1} or a^τ for some $\tau < t-2$ or the analogous happens for player two. The following two lemmas bound the probability of any of these events occurring.

Lemma 3.5. *Let $\varepsilon \in (0,1)$, and let T be a positive integer. Consider a two-player m-actions F-SBRD. If m is large enough, with probability at least $1 - \frac{\varepsilon}{2}$ there is no $t \leq T$ with either $a^t \in \{a^0, \dots, a^{t-3}\}$ or $b^t \in \{b^0, \dots, b^{t-3}\}$.*

Lemma 3.6. *Let $\varepsilon \in (0,1)$, and let T be a positive integer. Consider a two-player m-actions F-SBRD. If m is large enough, with probability at least $1 - \frac{\varepsilon}{2}$ there is no $t \leq T$ with either $a^t = a^{t-1}$ or $b^t = b^{t-1}$.*

Combining Lemmas 3.5 and 3.6, we conclude that for any $\varepsilon > 0$, for any fixed T, there exists $\bar{m}$ such that whenever $m \geq \bar{m}$,
$$\mathbb{P}\big(\text{INDD and SBRD coincide up to time } T\big) \geq 1 - \varepsilon. \tag{5}$$
Using this result, and that INDD converges quickly to a two-cycle, we obtain Theorem 3.1.

It can be shown that modifying SBRD to the inertial best-response dynamics (Swenson et al. (2018a)), where players keep their current strategy with probability ε, guarantees convergence to a NE with probability 1, as eventually one of the players will break from the cycle, and NE are absorbing for this dynamic. Our result is novel in showing that without inertia the deterministic SBRD does not converge, but instead stabilises in a persistent two-cycle orbiting in the neighbourhood of two equilibria. From a learning perspective, this is not a theoretical curiosity: agents updating with noisy or approximate value estimates may effectively behave without inertia in some strategy updates. Our findings therefore clarify why inertia matters, and show how its absence produces qualitatively different and potentially undesirable dynamics, complementing the existing literature.



(a) Probability of converging to a two-cycle.

(b) Number of steps to convergence.

Figure 1: SBRD in a two-player 50-actions game. 10000 samples were drawn. Runtime: 22 seconds.

4 EXPERIMENTAL RESULTS

We run extensive simulations for random potential and near-potential games with up to four players.

Key findings. (i) In 4.2, we show that the behaviour proved in Theorem 3.1 persists even in games where player payoffs are highly correlated but not identical. (ii) In 4.3 we show that, in the three-player settings, SBRD converges to a Nash equilibrium quickly and with high probability. (iii) In 4.4, we give evidence that SBRD is considerably faster than SPGD, while obtaining comparable rewards.

Technical details. All experiments were executed locally on an Apple MacBook Air with M3 chip with 16 GB RAM with no use of GPU. Code and data are publicly available in supplementary material. Metrics on continuous-valued variables are plotted with ± 2 standard errors (SE); binomial metrics are presented with 99.5% Clopper–Pearson confidence intervals.

4.1 NUMERICAL SETUP

Let n denote the number of players, m the number of actions, s the number of samples, and $\lambda \in [0, 1]$ the correlation parameter. For each experiment, we generate s independent n-player m-action games. For each action profile $a \in A$, the payoff $u_i(a)$ is drawn from a standard normal distribution with pairwise correlation λ between any two players $i \neq j$. Samples are taken independently for each a. As in Galla and Farmer (2013), we argue that this is the natural choice because, given the first and second degree moments, it is entropy maximising.

We vary λ over `[0.05 * i for i in range(21)]` to cover the full $[0, 1]$ range, and over `[0.85 + 0.025 * i for i in range(7)]` to test robustness to the potential game assumption. While a finer discretisation is possible, we find these values sufficient to illustrate the trends. The choice of m and n are described for each experiment.

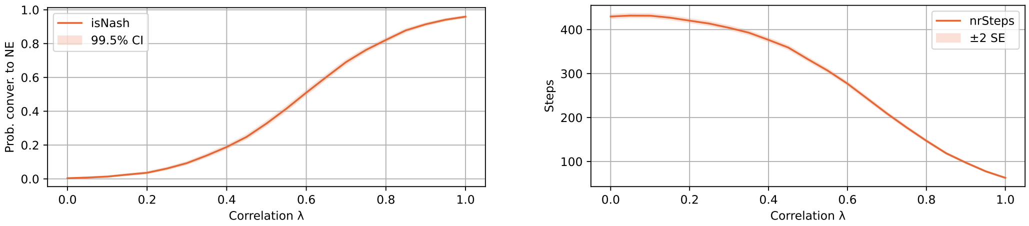
4.2 SBRD IN TWO-PLAYER GAMES

Our main findings, supported by Figure 1, are to support and show robustness of Theorem 3.1. Figure 1a illustrates the findings regarding two-player 50-actions games, and shows that, for high values of correlation λ, SBRD is likely to quickly converge to a two-cycle. For $\lambda = 1$ we rediscover the statement of our Theorem 3.1. Figure 1b also shows that the number of steps to convergence diminishes drastically with higher values of λ.

We address the assumption $m = 50$ in Appendix B.1, where we show that the same behaviour occurs for $m = 500$ (and therefore it is not reductive to assume $m = 50$ in this case).

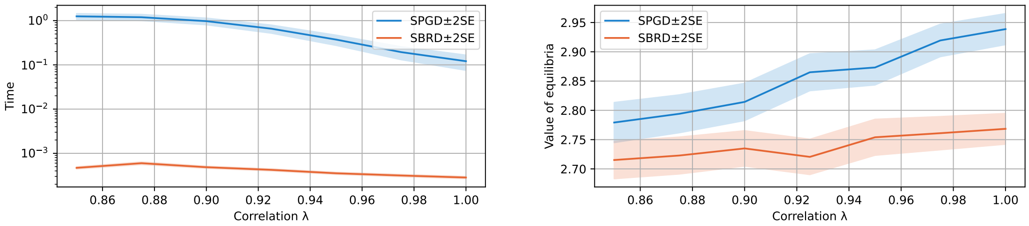
4.3 SBRD IN THREE (OR MORE)-PLAYER GAMES

Figure 2a provides strong empirical evidence that, in contrast to the two-player case, SBRD is likely to converge to a NE in three-player random potential games. This behaviour is not only prevalent in potential games, but also persists in games with sufficiently high payoff correlation λ. As for the two-player case, Figure 2b shows that convergence happens in a number of steps that diminishes for higher values of correlation λ.



(a) Probability of converging to a NE. (b) Number of steps to convergence.

Figure 2: SBRD in a three-player 50-actions game. 1000 samples were drawn. Runtime: 20 minutes.



(a) Time to convergence in SBRD and SPGD. (b) Final value of equilibrium.

Figure 3: Comparison of SPGD and SBRD in a three-player 50-actions game. 1000 samples were drawn. Runtime: 80 minutes.

As before, we postpone to Appendix B.2 to show that the assumption $m = 50$ is not reductive, and that similar behaviour occurs for $m = 100$.

Also in Appendix B.3 we address the case with four players, showing for the case $n = 4$, $m = 50$ that the same behaviour occurs in this setting as well. We conjecture that this trend extends to games with more than four players. However, we did not pursue this direction further, as we believe that the three- and four-player cases already provide strong evidence.

4.4 COMPARISON OF SBRD AND SPGD IN THREE-PLAYER NEAR-POTENTIAL GAMES

We now consider near-potential games with $\lambda \geq 0.85$ and compare SBRD with SPGD. As previously discussed in the paper outline, we selected SPGD as a natural baseline due to its smooth best-response updates, convergence guarantees, and model-free applicability.

As in the previous section, we focus on three-player games with 50 actions. Figure 3a shows that SBRD converges drastically faster than SPGD. Empirically, the time from start to convergence under SBRD is roughly three orders of magnitude lower than for SPGD.

In terms of achieved payoffs, SPGD tends to attain marginally higher equilibrium payoffs, but the difference remains small (Figure 3b). Crucially, as shown in Section B.4, SPGD often requires several thousand iterations to converge, and during its trajectory, the average payoff is much lower. On the other hand, as shown above in Figure 2b, SBRD consistently converges in under 100 iterations when $\lambda \geq 0.9$. In Section B.4 we quantify precisely the number of steps needed on average for SPGD to converge, and the average payoff of SPGD compared to the equilibrium value attained by SBRD. This speed advantage and the payoff comparison persist in the 100-actions setting as shown in Section B.5. We thus claim that SBRD provides a favourable trade-off, especially in online settings, delivering comparable payoffs at a small fraction of the computational cost. Furthermore, variations of SBRD, for example using some form of regularisation, might reveal a “sweat spot”: i.e. an algorithm that might enjoy both the fast convergence of SBRD and the high rewards of SPGD.

5 DISCUSSION AND LIMITATIONS

In this work, we have analysed Simultaneous Best Response Dynamics (SBRD) in the setting of random potential games. In contrast to sequential best response dynamics, SBRD requires no centralised coordination on the order of updates: at each round, every player updates their action to a best response against the joint profile of their opponents. This feature causes SBRD to be a more plausible model of strategic adaptation in decentralised multi-agent systems.

Our findings exhibit an interesting dependence on the number of players. In the two-player case, SBRD enters a two-cycle with high probability in games with highly correlated payoffs. In particular, the players alternate between two action profiles involving mismatched actions from two distinct Nash equilibria. Although such oscillatory behaviour prevents convergence, introducing a small random perturbation to each best-response update would break the cycle and restore convergence to a Nash Equilibrium (NE). By contrast, in games with three or four players, our simulations suggest that SBRD tends to converge very quickly to a NE. Moreover, when benchmarking against Softmax Policy Gradient Dynamics (SPGD), we observe that SBRD achieves higher learning-phase payoffs, even if SPGD tends to perform slightly better in terms of final payoffs at convergence. We conjecture that this also holds for n-player potential games where $n \geq 5$, though we have not conducted experiments to test this hypothesis. The underlying intuition is that, as the number of players grows, each player is increasingly unlikely to encounter the same profile of opponents' actions twice, restoring the approximate independence that underpins our INDD analysis. This line of inquiry is ongoing, and we have found that many of the conceptual tools developed in the two-player analysis can be applied for the n-player potential games where $n \geq 3$.

Moreover, since best-response updates depend solely on the ordinal ranking of payoffs, all results carry over to ordinal potential games in the sense of Monderer and Shapley (1996). We further demonstrate empirically that our conclusions are robust when the payoff-correlation assumption is relaxed: games with highly correlated payoffs exhibit the same convergence behaviours.

Assumptions and Limitations Our theoretical analysis focuses on two-player random potential games, hence we assume a perfectly correlated payoff structure. Although exact payoff alignment is uncommon in practical settings, the potential-game framework encompasses a broad class of models (see Section 1), and our empirical investigations indicate that the core convergence behaviour persists when payoffs are merely highly, rather than perfectly, correlated.

All experimental findings are derived from simulations in which payoff entries are drawn from a normal distribution. As previously argued, this is the natural entropy-maximising choice. However, this choice may not capture the diversity of strategic environments; exploring alternative distributions (e.g. uniform, heavy-tailed or bimodal) could reveal new phenomena.

We believe that our INDD method to approximate SBRD can be extended to work in many-player games to obtain a guarantee of convergence of the dynamics to a NE. This is aligned with the experimental results provided in this paper. The main technical difficulty of this approach is to study the 3-player case, which we believe we have addressed experimentally in this paper to a satisfactory degree but we hope will be addressed theoretically in future studies. Likewise, while we benchmark SBRD only against SPGD, other adaptive schemes, such as Q-learning, replicator dynamics or fictitious play, may exhibit different performance characteristics and merit systematic comparison.

Finally, our model assumes that each player has complete knowledge of their own payoffs and full observability of opponents' actions. Relaxing these assumptions to allow for partial observability or payoff estimation through exploration would make the model more realistic, but at the cost of substantially greater analytical complexity. We defer the study of such extensions to future research.

Summary The Simultaneous Best-Response Dynamic is a simple yet powerful learning rule, with provable convergence behaviour in two-player potential games and promising empirical performance in potential and near-potential games with more players. Its key limitation is the current gap between numerical conjectures and formal proofs for games with more than two-players. Addressing this challenge will deepen our theoretical understanding and broaden the applicability of SBRD.

ETHICS STATEMENT

We are not aware of ethical implications of our research, as it is an analysis of pre-existing concepts.

REPRODUCIBILITY STATEMENT

We aim to make our experimental results fully reproducible. All code used to produce the figures and tables in this paper is included as anonymised supplementary material (see Appendix B) and will remain available to reviewers. The main text and appendices contain detailed descriptions of the experimental setup.

LLM USAGE DISCLOSURE

We used LLMs only for minor copy-editing and occasional code comments. They did not contribute to the design of methods, proofs, or experiments, all of which were produced and validated by the authors.

REFERENCES

Pragnya Alatur, Anas Barakat, and Niao He. Independent policy mirror descent for markov potential games: Scaling to large number of players. *arXiv preprint arXiv:2408.08075*, 2024.

Ben Amiet, Andrea Collevecchio, and Kais Hamza. When "better" is better than "best". *Operations Research Letters*, 49(2):260–264, 2021a.

Ben Amiet, Andrea Collevecchio, Marco Scarsini, and Ziwen Zhong. Pure nash equilibria and best-response dynamics in random games. *Mathematics of Operations Research*, 46(4):1552–1572, 2021b.

Robert J. Aumann and Adam Brandenburger. Epistemic conditions for nash equilibrium. *Econometrica*, 63(5): 1161–1180, 1995.

Martino Bernasconi, Federico Cacciamani, Simone Fioravanti, Nicola Gatti, and Francesco Trovò. The evolutionary dynamics of soft-max policy gradient in multi-agent settings. *Theoretical Computer Science*, 1027:115011, 2025.

Shicong Cen, Fan Chen, and Yuejie Chi. Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 2833–2838. IEEE, 2022.

Desmond Chan, Bart De Keijzer, Tobias Galla, Stefanos Leonardos, and Carmine Ventre. Asymptotic extinction in large coordination games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13683–13690, 2025.

Dingyang Chen, Qi Zhang, and Thinh T. Doan. Convergence and price of anarchy guarantees of the softmax policy gradient in markov potential games. *arXiv preprint arXiv:2206.07642*, 2022.

Andrea Collevecchio, Hlafo Alfie Mimun, Matteo Quattropani, and Marco Scarsini. Basins of attraction in two-player random ordinal potential games. *arXiv preprint arXiv:2407.05460*, 2024.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pages 5166–5220. PMLR, 2022.

Jing Dong, Baoxiang Wang, and Yaoliang Yu. Convergence to nash equilibrium and no-regret guarantee in (markov) potential games. In *International Conference on Artificial Intelligence and Statistics*, pages 2044–2052. PMLR, 2024.

Davide Dragone, Luca Lambertini, and Arsen Palestini. Static and dynamic best-response potential functions for the non-linear cournot game. *Optimization*, 61(11):1283–1293, 2012.

Stéphane Durand and Bruno Gaujal. Complexity and Optimality of the Best Response Algorithm in Random Potential Games. In *SAGT 2016 - Symposium on Algorithmic Game Theory (SAGT) 2016*, pages 40–51, Liverpool, United Kingdom, September 2016. doi: 10.1007/978-3-662-53354-3_4. URL `https://hal.science/hal-01404643`.

Roy Fox, Stephen M Mcaleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in markov potential games. In *International Conference on Artificial Intelligence and Statistics*, pages 4414–4425. PMLR, 2022.

Tobias Galla and J Doyne Farmer. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, 110(4):1232–1236, 2013.

John C. Harsanyi and Reinhard Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press, 1 edition, 1988.

Sergiu Hart and Andreu Mas-Colell. Uncoupled dynamics do not lead to nash equilibrium. *American Economic Review*, 93(5):1830–1836, 2003.

Torsten Heinrich, Yoojin Jang, Luca Mungo, Marco Pangallo, Alex Scott, Bassel Tarbush, and Samuel Wiese. Best-response dynamics, playing sequences, and convergence to equilibrium in random games. *International Journal of Game Theory*, 52(3):703–735, 2023.

Amélie Heliou, Johanne Cohen, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. *Advances in Neural Information Processing Systems*, 30, 2017.

Josef Hofbauer and Karl Sigmund. Evolutionary game dynamics. *Bulletin of the American mathematical society*, 40(4): 479–519, 2003.

Tom Johnston, Michael Savery, Alex Scott, and Bassel Tarbush. Game connectivity and adaptive dynamics. *arXiv preprint arXiv:2309.10609*, 2023.

Sara Klein, Simon Weissmann, and Leif Döring. Beyond stationarity: Convergence analysis of stochastic softmax policy gradient methods. *arXiv preprint arXiv:2310.02671*, 2023.

Samson Lasaulce and Hamidou Tembine. *Game theory and learning for wireless networks: fundamentals and applications*. Academic Press, 2011.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 6820–6829. PMLR, 2020.

Hlafo Alfie Mimun, Matteo Quattropani, and Marco Scarsini. Best-response dynamics in two-person random games with correlated payoffs. *Games and Economic Behavior*, 145:239–262, 2024.

Dov Monderer and Lloyd S Shapley. Potential games. *Games and economic behavior*, 14(1):124–143, 1996.

Noam Nisan, Tim Roughgarden, and Eva Tardos. *Algorithmic Game Theory*. Cambridge University Press, 2007.

Robert W Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2:65–67, 1973.

Iosif Sakos, Stefanos Leonardos, Stelios Andrew Stavroulakis, William Overman, Ioannis Panageas, and Georgios Piliouras. Beating price of anarchy and gradient descent without regret in potential games. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

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

Wenjie Shi, Shiji Song, and Cheng Wu. Soft policy gradient method for maximum entropy deep reinforcement learning. *arXiv preprint arXiv:1909.03198*, 2019.

Youbang Sun, Tao Liu, Ruida Zhou, PR Kumar, and Shahin Shahrampour. Provably fast convergence of independent natural policy gradient for markov potential games. *Advances in Neural Information Processing Systems*, 36: 43951–43971, 2023.

Brian Swenson, Ceyhun Eksin, Soummya Kar, and Alejandro Ribeiro. Distributed inertial best-response dynamics. *IEEE Transactions on Automatic Control*, 63(12):4294–4300, December 2018a. ISSN 0018-9286. doi: 10.1109/TAC.2018.2817161. Publisher Copyright: © 2018 IEEE.

Brian Swenson, Ryan Murray, and Soummya Kar. On best-response dynamics in potential games. *SIAM Journal on Control and Optimization*, 56(4):2734–2767, 2018b.

Mark Voorneveld. Best-response potential games. *Economics letters*, 66(3):289–295, 2000.

Mark Voorneveld, Peter Borm, Freek Van Megen, Stef Tijs, and Giovanni Facchini. Congestion games and potentials reconsidered. *International Game Theory Review*, 1(03n04):283–299, 1999.

Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. On the global convergence rates of decentralized softmax gradient play in markov potential games. *Advances in Neural Information Processing Systems*, 35:1923–1935, 2022.

A THEORETICAL APPENDIX

In this appendix, we provide the proofs of section 3.1.

A.1 OBSERVATIONS REGARDING INDD

We start from some observations regarding INDD:

- Since, at time t, each player can only play the action that was played at time $t-2$ or one of the actions that they have not played before, the only cycles that can occur are of length two. As the set of action profiles is finite, INDD must cycle, and thus INDD must converge to a cycle of length two.
- At each time t, we have $\Psi(a^t, b^{t-1}) = \max(\Psi(a^{t-2}, b^{t-1}), \max R_1^{t-1})$ and $\Psi(a^{t-1}, b^t) = \max(\Psi(a^{t-1}, b^{t-2}), \max R_2^{t-1})$.
- Once either player repeats their previous but one action, there is always one player repeating their previous but one action, in an alternating manner. For example, suppose that at period t, Player 1 chooses $a^t = a^{t-2}$, then at period $t+1$ Player 2 chooses $b^{t+1} = b^{t-1}$.

We are now ready to follow the proofs.

A.2 PROOF OF LEMMA 3.2

Proof. Define two sequences (M_ℓ) and (N_ℓ) for $\ell = 1, 2, \ldots$ by:

$$M_\ell = \begin{cases} \Psi(a^\ell, b^{\ell-1}) & \text{if } \ell \text{ is odd,} \\ \Psi(a^{\ell-1}, b^\ell) & \text{if } \ell \text{ is even.} \end{cases} \tag{6}$$

$$N_\ell = \begin{cases} \Psi(a^{\ell-1}, b^\ell) & \text{if } \ell \text{ is odd,} \\ \Psi(a^\ell, b^{\ell-1}) & \text{if } \ell \text{ is even.} \end{cases} \tag{7}$$

Hence,

$$(M_\ell)_{\ell \geq 1} = (\Psi(a^1, b^0), \Psi(a^1, b^2), \Psi(a^3, b^2), \ldots) \tag{8}$$

$$(N_\ell)_{\ell \geq 1} = (\Psi(a^0, b^1), \Psi(a^2, b^1), \Psi(a^2, b^3), \ldots) \tag{9}$$

Observe that each transition $M_\ell \to M_{\ell+1}$ is a best-response transition by one of the players, so almost surely $M_{\ell+1} > M_\ell$, unless the opponent's action does not change (i.e. $b^\ell = b^{\ell+2}$ when ℓ is even, or $a^\ell = a^{\ell+2}$ when ℓ is odd). But note that, if $b^\ell = b^{\ell+2}$ for some even ℓ, then one obtains

$$a^{\ell+1} = a^{\ell+3}, \quad b^{\ell+2} = b^{\ell+4}, \quad \ldots \tag{10}$$

and the same holds if $a^\ell = a^{\ell+2}$ when ℓ is odd, which makes (M_ℓ) a one-cycle. So, with probability one, either $M_1 < M_2 < \ldots$ or (M_ℓ) is a one-cycle.

An identical argument applies to (N_ℓ). But then both players' actions have period at most 2, and so $T \leq 2$. Therefore, with probability one, no cycle of length greater than 2 can occur.

Since the space is finite, neither of the sequences can increase indefinitely, and therefore will eventually cycle. □

A.3 PROOF OF LEMMA 3.4

Proof. We split the argument into two parts.

Fix $\varepsilon > 0$ and a horizon $T \in \mathbb{N}$. The first part is to show that for all periods $t \leq T$, the probability that at least one player repeats their action from period $t-2$ is at least $1/2$, provided m large enough. Equivalently, at each t, at least one of the events

$$E_1^{t-1} : \Psi\big(a^{t-2}, b^{t-1}\big) \; > \; \max R_1^{t-1}\,, \tag{11}$$

$$E_2^{t-1} : \Psi\big(a^{t-1}, b^{t-2}\big) \; > \; \max R_2^{t-1}\,, \tag{12}$$

occurs with probability at least $1/2$.

The second part is to show that for large enough m, if exactly one of these events takes place, then from that period on, the probability that both events happen is at least $\frac{1}{2}$.

Once these parts are done, for m large enough to satisfy both conditions, it holds that for $T \geq \frac{\log \varepsilon}{\log(3/4)}$, the probability that INDD lasts more than T periods is less than ε.

For the first part, we focus on path of comparisons through the space of action profiles, where:

- Player 1 chooses a^1 to maximise $\Psi(a, b^0)$.
- Player 2 compares $\Psi(a^1, b^0)$ to the newly revealed values to choose b^2.
- Player 1 compares $\Psi(a^1, b^2)$ to the newly revealed values to choose a^3.
- And so on …

The choice of starting with Player 1 is arbitrary. There is an equivalent path that begins with Player 2 choosing b^1.

In the first step of the path, Player 1 selects

$$a^1 \;=\; \arg\max_{a \neq a^0} \Psi(a, b^0)\,, \tag{13}$$

so that

$$\Psi(a^1, b^0) = \max\bigl\{\Psi(a, b^0) : a \in A_1 \setminus \{a^0\}\bigr\}\,, \tag{14}$$

is the maximum of $m-1$ independent draws from F.

In the second step of the path, Player 2 only knows $\Psi(a^1, b^0)$ and draws $m-2$ new payoffs

$$R_2^1 = \{\Psi(a^1, b) : b \in A_2 \setminus \{b^0, b^1\}\}\,. \tag{15}$$

and they return to playing b^0 precisely if

$$\Psi(a^1, b^0) \;>\; \max R_2^1\,. \tag{16}$$

By symmetry of i.i.d. sampled from F,

$$\mathbb{P}(\Psi(a^1, b^0) \;>\; \max R_2^1) = \frac{m-1}{(m-1)+(m-2)} > \frac{1}{2} \quad (\text{for } m \geq 3)\,. \tag{17}$$

Hence, the event E_2^1 occurs with probability at least $1/2$.

In the event that E_2^1 does not occur, then at period 2 Player 2 is playing b^2, and $\Psi\bigl(a^1, b^2\bigr)$ is the maximum of $2m-3$ i.i.d. sampled from F. Then, $m-3$ new payoffs are randomised, and by the same symmetry argument, we have

$$\mathbb{P}(E_1^2) = \mathbb{P}(\Psi\bigl(a^1, b^2\bigr) \;>\; \max R_1^2) = \frac{2m-3}{(2m-3)+(m-3)} > \frac{1}{2}\,. \tag{18}$$

In general, consider period t. If t is odd, then, in the event that $E_2^1, E_1^2, E_2^3, \dots, E_1^{t-1}$ all did not occur, then Player 1 is playing a^t, and $\Psi\bigl(a^t, b^{t-1}\bigr)$ is the maximum of $\sum_{\tau=1}^{t-1}(m-\tau) = (t-1)m - \frac{t(t-1)}{2}$ i.i.d. from F. The realisations of $m-t$ variables are observed, and hence:

$$\mathbb{P}(E_2^t) = \mathbb{P}(\Psi(a^t, b^{t-1}) \;>\; \max R_2^t) = \frac{(t-1)m - \frac{t(t-1)}{2}}{tm - \frac{t(t+1)}{2}} > \frac{1}{2}\,. \tag{19}$$

For t even, analogously we can show that $\mathbb{P}(E_1^t) > \frac{1}{2}$. Then, for m large enough, this holds for all $t < T$.

For the second part, we suppose that at some period t exactly one of the events E_1^{t-1} or E_2^{t-1} occurs; without loss of generality assume

$$E_1^{t-1} :\; \Psi\bigl(a^{t-2}, b^{t-1}\bigr) \;>\; \max R_1^{t-1} \quad \text{and} \quad \neg E_2^{t-1} :\; \Psi\bigl(a^{t-1}, b^{t-2}\bigr) \;\leq\; \max R_2^{t-1}\,. \tag{20}$$

Then Player 1 re-plays action a^{t-2}, so $a^t = a^{t-2}$, while Player 2 plays a new action $b^t \neq b^{t-2}$. Hence the action profile at time t is

$$(a^t, b^t) = \left(a^{t-2}, b^t\right).$$

We show that the probability that the process terminates in the next period is at least $1/2$, provided m large enough.

At period $t+1$, Player 1 compares the known value $\Psi(a^{t-1}, b^t)$ (which is the maximum of at least $m-1$ independent draws from F) to the maximum of the newly realised payoffs in R_1^t, which contains at most $m-1$ new samples from F. Meanwhile, Player 2 compares the known value $\Psi(a^t, b^{t-1}) = \Psi(a^{t-2}, b^{t-1})$ to no newly generated values (since a^{t-2} was just re-played), and so replays b^{t-1}.

Thus at period $t+1$, by the same symmetry argument as before, the probability that Player 1 repeats a^{t-1} again is at least:

$$\frac{m-1}{(m-1)+(m-1)} = \frac{1}{2}. \tag{21}$$

Hence, with probability at least $1/2$, the action profile (a^{t-1}, b^{t-1}) is repeated, and so the process terminates at period $t+1$.

Putting the two parts together: choose m_0 large enough that in each period $t \le T$ both

$$\mathbb{P}\big(E_1^t \cup E_2^t\big) \ge \frac{1}{2} \quad \text{and} \quad \mathbb{P}\big(\text{termination} \mid \text{exactly one of } E_1^{t-1}, E_2^{t-1}\big) \ge \frac{1}{2}. \tag{22}$$

Then the probability the process survives beyond T is bounded above by

$$\left(1 - \tfrac{1}{2}\cdot\tfrac{1}{2}\right)^T = \left(\tfrac{3}{4}\right)^T, \tag{23}$$

and for $T \ge \frac{\log \varepsilon}{\log(3/4)}$ this is at most ε.

□

A.4 PROOF OF LEMMA 3.5

Proof. Fix any horizon T. For $t = 3, \dots, T$, let E_t^1 be the event that

$$\max_{t' < t-2} \Psi(a^{t'}, b^{t-1}) > \Psi(a^{t-2}, b^{t-1}). \tag{24}$$

Using the same argument as in the previous proposition, $\Psi(a^{t-2}, b^{t-1})$ must be the maximum of $(t-1)m - \frac{t(t-1)}{2}$ i.i.d samples from F. Therefore, if E_t^1 occurs then one of the $t-2$ payoffs $\{\Psi(a^{t'}, b^{t-1}) : t' < t-2\}$ must exceed this maximum. By symmetry, for each fixed t

$$\mathbb{P}(E_t^1) \;\le\; \frac{t-2}{(t-1)m - \frac{t(t-1)}{2}} \;\le\; \frac{t-2}{m-1}. \tag{25}$$

Hence by the union bound,

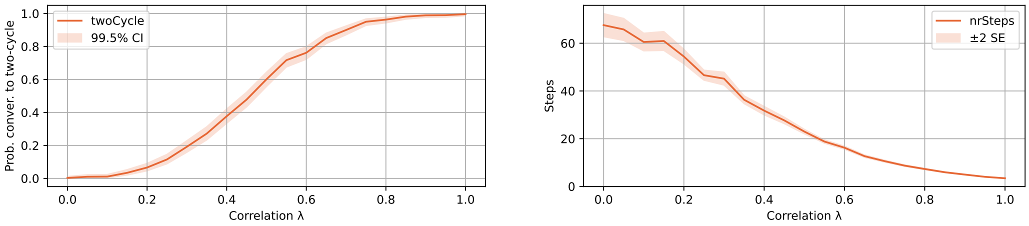
$$\mathbb{P}\Big(\bigcup_{t=3}^{T} E_t^1\Big) \;\le\; \sum_{t=3}^{T} \frac{t-2}{m-1} = \frac{(T-2)(T-1)/2}{m-1}, \tag{26}$$

which can be made below $\varepsilon/4$ by choosing m large. One can define E_t^2 to be the analogous event for player two, and achieve that $\mathbb{P}\Big(\bigcup_{t=3}^{T} E_t^2\Big) \le \varepsilon/4$ by the same argument. This bounds the probability that SBRD differs from INDD on account of any ‘old’ action–payoff comparison. □

A.5 PROOF OF LEMMA 3.6

Proof. Again fix horizon T. At each period $t = 1, \dots, T$, SBRD additionally compares the single value $\Psi(a^{t-1}, b^{t-1})$ against at least $m - t - 1$ fresh samples of the distribution F. By symmetry the chance it is the maximum is

$$\frac{1}{(m-t-1)+1} = \frac{1}{m-t}. \tag{27}$$



(a) Probability of converging to a two-cycle.

(b) Number of steps to convergence.

Figure 4: SBRD in a two-player 500-actions game. 1000 samples were drawn. Runtime: 127 seconds

Over T periods, a union-bound gives

$$\Pr\big(\exists\, t \leq T : \text{SBRD uses } \Psi(a^{t-1}, b^{t-1})\big) \;\leq\; \sum_{t=1}^{T} \frac{1}{m-T} = \frac{T}{m-T}\,, \tag{28}$$

which is below $\varepsilon/2$ for all $m \geq \frac{T(1+\varepsilon/2)}{\varepsilon/2}$. ☐

A.6 PROOF OF THEOREM 3.1

Proof. Fix $\varepsilon > 0$. We show that for sufficiently large m the following three events each occur with probability at least $1 - \frac{\varepsilon}{3}$, and hence by the union bound the SBRD process converges to a 2-cycle with probability at least $1 - \varepsilon$.

Firstly, by Lemma 3.4, there exist an integer m_0 such that whenever $m \geq m_0$ the INDD process terminates by period $T = \frac{\log(\varepsilon/3)}{\log(3/4)}$ with probability at least $1 - \frac{\varepsilon}{3}$.

As mentioned in A.1, INDD cannot terminate in a 1-cycle and cannot cycle with length > 2. Hence on termination it must enter a 2-cycle with probability one (and so at least $1 - \frac{\varepsilon}{3}$).

By Lemmas 3.5 and 3.6, there exists m_1 such that whenever $m \geq m_1$ the probability that INDD and SBRD differ at some period $t \leq T$ is at most $\frac{\varepsilon}{3}$. Equivalently, with probability at least $1 - \frac{\varepsilon}{3}$ they coincide up to time T.

Therefore, if $m \geq \max\{m_0, m_1\}$, then each of the three events has probability at least $1 - \frac{\varepsilon}{3}$, so by the union bound all three occur simultaneously with probability at least

$$1 - 3 \cdot \tfrac{\varepsilon}{3} = 1 - \varepsilon\,. \tag{29}$$

In that event, SBRD follows the same path as INDD up to period T, INDD terminates in a two-cycle by T, and hence SBRD too converges to that same two-cycle. Therefore

$$\mathbb{P}\big(\text{SBRD converges to a two-cycle by time } T\big) \;\geq\; 1 - \varepsilon\,, \tag{30}$$

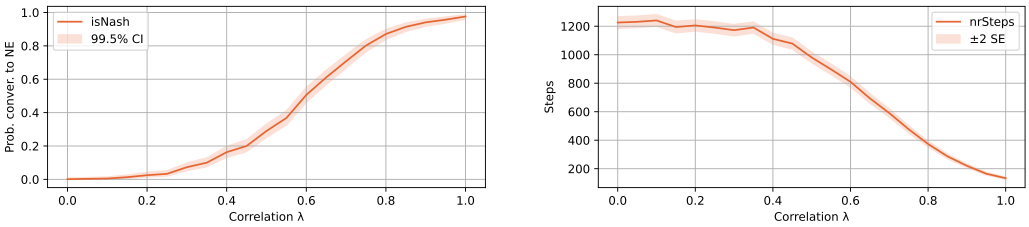
as required. ☐

B EXPERIMENTAL APPENDIX

B.1 ROBUSTNESS TO NUMBER OF ACTIONS FOR SECTION 4.2

We now complement the experimental findings of Section 4.2 by showing that the results are robust with respect to the number of actions. In particular, Figure 4 shows that two-player random games with 500 actions exhibit the same behaviour as in the 50-action case: the probability of convergence to a two-cycle varies similarly with λ, and the number of steps required to converge in highly correlated games remains of the same order of magnitude.

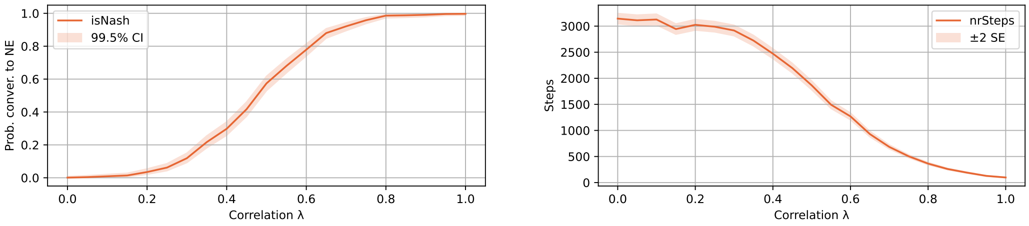
To achieve a convergence probability of at least 90%, values of $\lambda \geq 0.9$ were needed for $m = 50$, whereas for $m = 500$, values of $\lambda \geq 0.75$ were sufficient. This suggests that the behaviour predicted by Theorem 3.1 extends to larger games and can emerge even at lower levels of correlation.



(a) Probability of converging to a NE.

(b) Number of steps to convergence.

Figure 5: SBRD in a three-player 100-actions game. 1000 samples were drawn. Runtime: 14 minutes.



(a) Probability of converging to a NE.

(b) Number of steps to convergence.

Figure 6: SBRD in a four-player 50-actions game. 1000 samples were drawn. Runtime: 142 minutes.

These findings strongly support the claim made in Section 4.2 that in two-player highly correlated random games, SBRD quickly converges to a two-cycle.

The experiment shown in Figure 4 ran in 127 seconds.

B.2 ROBUSTNESS TO NUMBER OF ACTIONS FOR SECTION 4.3

We now show with Figure 5 that the number of actions does not affect the outcomes reported in Section 4.3. Specifically, we run experiments on three-player random games with 100 actions across various levels of correlation λ. The results closely mirror those observed in the 50-action case. For high values of λ, the probability that SBRD converges to a Nash equilibrium approaches one, and this behaviour appears smoothly as correlation increases. In other words, in highly correlated games, SBRD is very likely to converge to a Nash equilibrium.

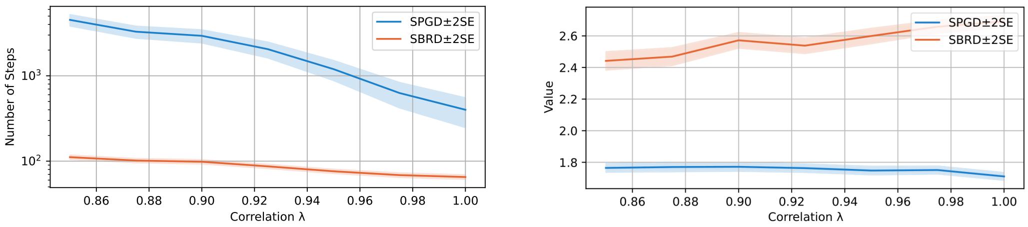
We also observe that the number of steps required for convergence remains of the same order of magnitude across both settings when λ is large.

This new evidence reinforces the claim made in Section 4.3 that in highly correlated three-player games, SBRD tends to quickly converge to a Nash equilibrium.

The experiment shown in Figure 5 ran in 14 minutes.

B.3 ROBUSTNESS TO NUMBER OF PLAYERS FOR SECTION 4.3

Having shown that the number of actions does not influence the behaviour of SBRD across different levels of λ, we now see if the behaviour is influenced by the number of players. As previously mentioned, we believe that the convergence to a two-cycle (and thus not to a NE) observed in the two-player setting is a special case, and that for games with three or more players and high payoff correlation, SBRD is likely to converge to a Nash equilibrium.



(a) Steps to convergence in SBRD and SPGD.

(b) SBRD equilibrium value vs SPGD average value.

Figure 7: Comparison of SPGD and SBRD in a three-player 50-actions game. 1000 samples were drawn. Runtime: 80 minutes.

Figure 6 confirms that SBRD behaves in the four-player case as it does in the three-player setting. Specifically, the probability of convergence to a NE is very high for large values of λ, and the number of steps required to converge decreases sharply as correlation increases.

While we do not experimentally test games with more than four players, nor provide a formal proof, ongoing research is aimed at establishing this behaviour theoretically.

The experiment shown in Figure 6 ran in 142 minutes.

B.4 COMPLEMENT TO SECTION 4.4

We now justify our claim that SBRD provides a viable alternative to SPGD when the correlation is high, especially in online settings. We do this by examining the trade-off between convergence speed and final payoff. As shown in Section 4.4, SBRD typically reaches slightly lower equilibrium payoffs than SPGD. However, Figure 7a demonstrates that SBRD converges in significantly fewer steps, allowing agents to begin benefitting from equilibrium payoffs much earlier.

When comparing the average payoff of SPGD along its learning trajectory with the final payoff obtained by SBRD (as shown in Figure 7b), we find that SPGD accumulates substantially lower rewards during training. This suggests that in online settings, or in environments where short to medium time horizons are critical, SBRD may be the preferable choice.

In the next section we show that these differences become even more pronounced when the number of actions increases.

The experiment shown in Figure 7 ran in 80 minutes.

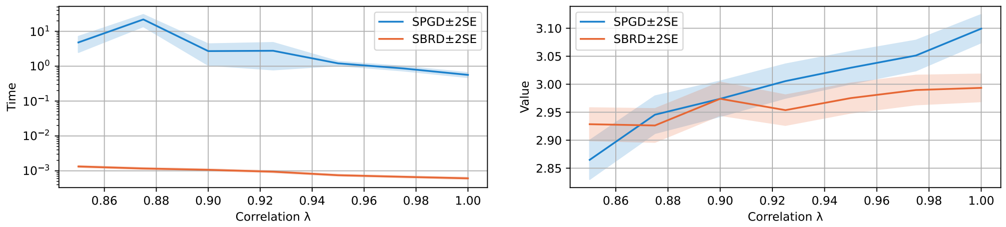
B.5 ROBUSTNESS TO NUMBER OF ACTIONS FOR SECTION 4.4

Finally, we present further evidence that highly correlated three-player random games with 100 actions exhibit behaviour consistent with the 50-action case discussed earlier. The findings of this section can all be found in Figure 8.

In particular, the findings show that SBRD converges to a Nash equilibrium significantly faster than SPGD (three to four orders of magnitude faster). This confirms the scalability of SBRD's performance as the size of the action space increases.

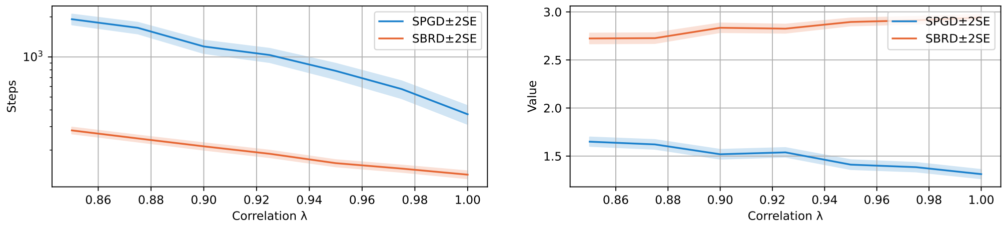
Moreover, we can see that the payoffs attained by both algorithms at equilibrium are closely comparable in magnitude. Notably, for relatively lower values of correlation, SBRD on average achieves better equilibrium values than SPGD. This highlights that SBRD's faster convergence does not come at a substantial cost in reward quality.

As with the 50-action experiments, we also find that SBRD requires far fewer steps to reach convergence. This reinforces our claim that SBRD is particularly well suited for online or time-sensitive environments. In such settings, agents often benefit more from earlier access to high-value strategies than from long-term optimality alone. Since the average payoff collected by SPGD along its learning trajectory is consistently lower than the payoff achieved at equilibrium by SBRD, the latter emerges as a competitive alternative in scenarios with limited time horizons.



(a) Time to convergence in SBRD and SPGD.

(b) Final value of equilibrium.



(c) Steps to convergence in SBRD and SPGD.

(d) SBRD equilibrium value vs SPGD average value.

Figure 8: Comparison of SPGD and SBRD in a three-player 100-actions game. 1000 samples were drawn. Runtime: 585 minutes.

The experiment shown in Figure 8 ran in 585 minutes.