Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 189 (2020) 105095

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Best-response dynamics in zero-sum stochastic games [☆]

David S. Leslie [a], Steven Perkins [b], Zibo Xu [c,*]

[a] Department of Mathematics and Statistics, Lancaster University, UK
[b] PwC, Bristol, UK
[c] Engineering Systems and Design, SUTD, Singapore

Received 14 April 2018; final version received 3 May 2020; accepted 2 July 2020
Available online 13 July 2020

Abstract

We define and analyse three learning dynamics for two-player zero-sum discounted-payoff stochastic games. A continuous-time best-response dynamic in mixed strategies is proved to converge to the set of Nash equilibrium stationary strategies. Extending this, we introduce a fictitious-play-like process in a continuous-time embedding of a stochastic zero-sum game, which is again shown to converge to the set of Nash equilibrium strategies. Finally, we present a modified δ-converging best-response dynamic, in which the discount rate converges to 1, and the learned value converges to the asymptotic value of the zero-sum stochastic game. The critical feature of all the dynamic processes is a separation of adaption rates: beliefs about the value of states adapt more slowly than the strategies adapt, and in the case of the δ-converging dynamic the discount rate adapts more slowly than everything else.

JEL classification: C73

[☆] The authors wish to thank Y. Babichenko, S. Hart, J. Hörner, X. Li, T.E.S. Raghavan, J. Renault, L. Samuelson, B. Sandholm, S. Sorin, M. Staudigl, S. Takahashi, O. Tercieux, F. Thuijsman, T. Tomala, X. Venel, N. Vieille, Y. Viossat, and J. Weibull for comments. Xu acknowledges financial support from Agence Nationale de la Recherche Grant ANR-10-BLAN-0112 JEUDY, the Start-up Research Grant 2014079 at SUTD, and the visitor program at Cowles Foundation for Research in Economics at Yale University. Perkins and Leslie's contributions were funded by grant numbers EP/D063485/1 and EP/I032622/1 respectively from the Engineering and Physical Sciences Research Council, UK.

[*] Corresponding author.
E-mail addresses: d.leslie@lancaster.ac.uk (D.S. Leslie), steven.perkins@pwc.com (S. Perkins), zibo.xu@alumni.lse.ac.uk (Z. Xu).

https://doi.org/10.1016/j.jet.2020.105095

Keywords: Stochastic games; Best-response dynamics; Zero-sum games; Convergence

1. Introduction

Evolutionary and learning approaches to game theory justify equilibrium play as the end point of a dynamic process resulting from adaptations made by boundedly rational players. However to date there has been only limited success in applying the evolutionary or adaptive learning approach to stochastic games. The continuous-time best-response dynamic, a staple of evolutionary game theory, has thus far only been studied in normal-form and extensive-form games. We therefore define and investigate best-response dynamics for two-player zero-sum stochastic games.

The standard best-response dynamic in a game is specified as a differential inclusion with a constant revision rate; see Matsui (1989), Gilboa and Matsui (1991), Hofbauer (1995), and Balkenborg et al. (2013). A state in the dynamic specifies the strategy profile of all players, and the frequency of a strategy increases only if it is a best response to the current state. It is worth noting that the continuous-time best-response dynamic is equivalent to a continuous-time fictitious play (Brown, 1949) after a time rescaling. The best-response dynamic has been analyzed in various classes of normal-form games (also called strategic-form games or one-shot games); see Hofbauer and Sigmund (1998) and Sandholm (2010). In particular, the convergence of a continuous-time best-response dynamic to the set of Nash equilibria has been shown in Harris (1998), Hofbauer (1995), and Hofbauer and Sorin (2006) for two-player zero-sum games, in Harris (1998) for weighted-potential games, and in Berger (2005) for $2 \times n$ games. For convergence in extensive-form games of perfect information, see Xu (2016).

In a stochastic game (Shapley, 1953) players are in some state each time a decision is to be made; the actions of players in the current state determine not only the instantaneous payoffs but also the transition probability to the state for the next decision making. Thus, each player has to balance between the two sometimes contradictory goals, namely the better instantaneous payoff today and the better state distribution tomorrow. Meanwhile, the other players are also maximizing their own goals, which makes the decision problem of each player even more complicated. The existence of Nash equilibrium in a stochastic game has been proved for several classes of stochastic games; see Solan (2009) for a survey.

The question addressed in this paper is whether boundedly rational players can reach an equilibrium in a stochastic game. In particular, if players are unable or unprepared to carry out equilibrium calculations or solve Bellman equations for future reward, could they learn the Nash equilibrium strategy in the end? In the present paper, we focus on zero-sum stochastic games with discounted payoff, as is introduced by Shapley (1953), and consider best-response dynamics.

We first point out that it is non-trivial to define a best-response dynamic in a stochastic game, and indeed no established notion is available in the literature yet. Some discrete-time algorithmic approaches that achieve convergence have been presented (e.g., Borkar, 2002; Szepesvári and Littman, 1999; Vrieze and Tijs, 1982), but their convergence has been proved using ad hoc methods instead of considering an underlying dynamic. Perkins (2013) studies a continuous-time best-response dynamic in a stochastic game in which an agent does not anticipate changes to future payoffs as a result of strategy evolution. In his model, a player can calculate the expected future discounted payoff starting at each state for any given stationary strategy profile. When a player is calculating the best response at a state, she assumes that her total payoff will consist of the instantaneous payoff for taking that action against the opponent's action in that state, followed by a future payoff that is determined by the current strategies of both players. Convergence is shown only when the players are sufficiently impatient.

In the present paper, we construct best-response dynamics in which the future payoffs are learned separately from the strategies, to circumvent the problems encountered by Perkins (2013). We suppose that players are myopic learners who cannot calculate the future expected discounted payoff in a zero-sum stochastic game. Instead, they assume an (initially) arbitrary set of continuation payoffs, one for each state. These continuation payoffs allow the definition of an auxiliary game for each state, in which the payoff to an action is given by the instantaneous payoff plus the expected continuation payoff at the subsequent state.

In all our learning dynamics, the continuation payoffs are updated more slowly as time goes on, at rate $1/t$. In this way, a continuation payoff is simply the time average of payoffs in the corresponding auxiliary game. As players do not have the ability to calculate the true continuation payoff for the current mixed strategies in the stochastic game, they view this time average as the current best estimate.

We first consider a best-response dynamic in which each player plays a mixed strategy in each auxiliary game, and continuously adjusts this auxiliary game strategy in the direction of the best response to the current mixed strategy of the opponent in that auxiliary game. Here, the speed of strategy adjustment in the best-response dynamic is independent of calendar time t. The key to the convergence of this best-response dynamic is simply the different adjustment speed between the best-response dynamic on players' strategies and the slow adaptation of the continuation payoffs. The slowly evolving auxiliary games allow the players to learn to play close to an equilibrium of the auxiliary game; this in turn allows the continuation payoffs to converge, so that the strategy profile being played approaches an equilibrium strategy profile in the stochastic game. We show in Section 3 that this dynamic converges at rate $1/t$ in payoff terms.

In the best-response dynamic so far proposed, both players update and play mixed strategies in all states at all times. To introduce a more natural learning model of play between two players, we also introduce a continuous-time state-dependent fictitious play process, in which actual play of the game takes place in real time. In this process, the game transitions through the states according to a controlled continuous-time Markov chain, where the controlling parameter is the action profile currently being played in the state. While the game is in a state, each player plays a best response to her belief about the opponent's action in that state as well as the current continuation payoffs. Specifically, each player observes the action taken by the opponent and updates her belief about the opponent's behaviour in the current state at constant rate in the direction of the currently-observed action. The continuation payoffs of all states are updated as in the best-response dynamic, tracking the empirical time average of auxiliary game payoffs. There is no need for these to be updated only in the current state, since these are unobserved hypothetical quantities anyway. Again, the separation of adjustment speeds ensures convergence of this state-dependent fictitious play process.

We finish by progressing further and propose a variant of the best-response dynamic such that the payoff in each auxiliary game converges to the corresponding asymptotic value of the zero-sum stochastic game when the discount factor increases to 1. This is achieved by once again evolving a parameter slowly in comparison to the others; in this case the discount factor adjusts towards 1 even more slowly than the continuation payoffs. So far as we can ascertain, this is the first adaptive dynamical procedure which converges to the asymptotic value of a zero-sum stochastic game.

We postpone the literature review of stochastic games, and the positioning of our work within that literature, to Section 6.

2. The game models

We begin by reviewing relevant results in two-player zero-sum normal-form games. These results will be used for the convergence within auxiliary games in the best-response dynamics for stochastic games. We then define zero-sum stochastic games and introduce the concepts that are central to the development of our learning dynamics in the rest of the paper.

2.1. Zero-sum normal-form games

In a two-player zero-sum game G where Player 1 and 2's finite pure strategy sets are A^1 are A^2, respectively, the (a^1, a^2) element $r(a^1, a^2)$ in the payoff matrix denotes the payoff to Player 1 when Player 1 plays a^1 and Player 2 plays a^2. We can then linearly extend the payoff function to mixed strategies, i.e. $r(x^1, x^2)$ is defined for any $x^1 \in \Delta(A^1)$ and $x^2 \in \Delta(A^2)$. For convenience, we may write $x = (x^i)_{i=1,2}$ as a strategy profile. Recall the *value* of the zero-sum game G is

$$v(G) := \max_{x^1\in\Delta(A^1)} \min_{x^2\in\Delta(A^2)} r(x^1, x^2) = \min_{x^2\in\Delta(A^2)} \max_{x^1\in\Delta(A^1)} r(x^1, x^2). \tag{2.1}$$

An optimal strategy of Player 1 guarantees the payoff no less than $v(G)$, regardless of the strategy of Player 2; similarly, an optimal strategy of Player 2 guarantees the payoff to Player 1 no more than $v(G)$. An optimal strategy profile is also a Nash equilibrium in G. (We use "optimal" here to mean a minimax strategy in a zero-sum game.)

The best-response dynamics have been well studied by authors including Brown (1949), Matsui (1989), Gilboa and Matsui (1991), Hofbauer (1995), Hofbauer and Sigmund (1998), Fudenberg and Levine (1998), Harris (1998), Hopkins (1999), Benaïm et al. (2005), Berger (2005), Hofbauer and Sorin (2006), Leslie and Collins (2006), Sandholm (2010), and Viossat and Zapechelnyuk (2013). They are motivated as a model of learning either by individuals constantly updating their mixed strategies towards a best response to opponent mixed strategies (e.g. Leslie and Collins, 2006), as a continuous-time fictitious play process in which beliefs are continuously adjusted towards observed opponent best responses (e.g. Harris, 1998), as a version of Bayesian updating process with a prior in a Dirichlet distribution (e.g. Fudenberg and Levine, 1998), or as a limiting process that can be used to study discrete time fictitious play (e.g. Benaïm et al., 2005). Others consider the best-response dynamics simply as a method for calculating equilibrium (Brown, 1949). Under these dynamics, strategies evolve at a constant rate in the direction of the current best response, defined for Player 1 and 2 respectively as

$$br^1(x^2) := \operatorname*{argmax}_{\rho^1\in\Delta(A^1)} r(\rho^1, x^2) \quad \text{and} \quad br^2(x^1) := \operatorname*{argmin}_{\rho^2\in\Delta(A^2)} r(x^1, \rho^2).$$

The best-response dynamic in a normal-form game is therefore defined by

$$\dot{x}^i \in br^i(x^{-i}) - x^i, \ \forall i = 1, 2, \tag{2.2}$$

where the dot represents derivative with respect to time, and we have suppressed the time argument t. Since best-response strategies are in general not unique, this is actually a differential inclusion. In normal-form games the set $br^i(x^{-i})$ is upper semi-continuous in x^{-i}, so a solution trajectory of (2.2) exists, though not necessarily unique; see Aubin and Cellina (1984) and Benaïm et al. (2005).

Given a strategy profile $x = (x^1, x^2)$, we define the *energy* to be

$$w(x) := \max_{\rho^1 \in \Delta(A^1)} r(\rho^1, x^2) - \min_{\rho^2 \in \Delta(A^2)} r(x^1, \rho^2). \tag{2.3}$$

It is straightforward to see that

$$|r(x) - v(G)| \le w(x), \ \forall x \in \Delta(A^1) \times \Delta(A^2), \tag{2.4}$$

and that $w(x) = 0$ if and only if x^1 and x^2 are optimal strategies of Player 1 and 2, respectively.

Harris (1998) and Hofbauer and Sorin (2006) show the following result:

Theorem 2.1. *Given a zero-sum normal-form game G, along every solution trajectory $(x(t))_{t\ge 0}$ of (2.2), $w(x(t))$ is a Lyapunov function with*

$$\tfrac{\mathrm{d}}{\mathrm{d}t} w(x(t)) = -w(x(t)) \quad \text{for almost all } t. \tag{2.5}$$

Hence

$$w(x(t)) = e^{-t} w(x(0)) \tag{2.6}$$

and every solution trajectory of (2.2) converges to the set of optimal strategy profiles. That is,

$$||x(t) - Z||_\infty := \inf_{z \in Z} ||x(t) - z||_\infty \to 0, \ \text{as } t \to \infty,$$

where Z denotes the set of optimal strategy profiles in G.

Sketch proof. In a solution trajectory $(x(t))_{t\ge 0}$ of the best-response dynamic (2.2), $\dot{x}(t)$ exists for almost all $t \ge 0$. Let us write $b(t) := x(t) + \dot{x}(t)$ whenever $x(t)$ is differentiable. Hofbauer and Sorin (2006) show, by a version of the envelope theorem, that

$$\tfrac{\mathrm{d}}{\mathrm{d}t} \max_{\rho^1 \in \Delta(A^1)} r(\rho^1, x^2(t)) = r(b^1(t), \dot{x}^2(t)) \quad \text{and}$$
$$\tfrac{\mathrm{d}}{\mathrm{d}t} \min_{\rho^2 \in \Delta(A^2)} r(x^1(t), \rho^2) = r(\dot{x}^1(t), b^2(t)).$$

Therefore, for almost all $t \ge 0$,

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t} w(x(t)) &= r(b^1(t), \dot{x}^2(t)) - r(\dot{x}^1(t), b^2(t)) \\ &= r(b^1(t), b^2(t) - x^2(t)) - r(b^1(t) - x^1(t), b^2(t)) \\ &= -r(b^1(t), x^2(t)) + r(x^1(t), b^2(t)) \\ &= -w(x(t)). \quad \square \end{aligned} \tag{2.7}$$

2.2. *Zero-sum stochastic games*

Our objective in this article is to develop similar results for a stochastic game, defined in this section. A two-player zero-sum discounted-payoff stochastic game is a tuple $\Gamma = \langle I, S, A, P, r, \delta \rangle$ constructed as follows.

- Let $I = \{1, 2\}$ be the set of players.
- Let S be a set of finitely many states.
- For each player i in state s, A_s^i denotes a set of finitely many actions. For each state s, we put the set of action pairs $A_s := A_s^1 \times A_s^2$.
- For each state pair (s, s') and each action pair $a \in A_s$, we define $P_{s,s'}(a)$ to be the *transition probability* from state s to state s' given the action pair a.
- We define $r_s(\cdot)$ to be the *stage payoff function* for Player 1. That is, when the process is in a state s, $r_s(a)$ is the instantaneous payoff to Player 1 for the action pair $a \in A_s$. Note that, in a zero-sum game, Player 2 always receives stage payoff $-r_s(a)$.
- δ is a discount factor that affects the importance of future stage payoffs relative to the current stage payoff.

In any state s, Player i plays an action $x_s^i \in \Delta(A_s^i) =: \Delta_s^i$. That is, $x_s^i(a^i)$ denotes the probability that when in state s, player i selects action $a^i \in A_s^i$. In this paper, we only consider stationary strategies for both players. A *stationary strategy* $x^i \in \Delta^i := \times_{s \in S} \Delta_s^i$ of player i specifies for each state s a mixed strategy x_s^i to be played whenever the state is s. We denote a *strategy profile* by $x = (x^1, x^2) = ((x_s^1)_{s\in S}, (x_s^2)_{s\in S})$, and the set of strategy profiles by $\Delta := \Delta^1 \times \Delta^2$. Given a strategy profile x in Δ, for any state s, we may write

$$r_s(x_s) = r_s(x_s^1, x_s^2) = \sum_{a \in A_s^1 \times A_s^2} x_s^1(a^1) x_s^2(a^2) r_s(a),$$

and similar treatment applies to a transition probability $P_{s,s'}(x_s)$. To ease the exposition, we denote a stochastic game Γ starting from a state s by Γ_s. We can then define the *expected discounted payoff* for Player 1 under the strategy profile x in Γ_s as

$$U_s(x) := \mathbb{E}\left[(1-\delta) \sum_{n=0}^{\infty} \delta^n r_{s_n}(x_{s_n}) \middle| s_0 = s \right], \tag{2.8}$$

where $\{s_n\}_{n\in\{0,1,2,\dots\}}$ is a stochastic process representing the state of the process at each iteration, and $(1-\delta)$ is to normalize the discounted payoff. Of course, Player 2 has an expected discounted payoff $-U_s(x)$. Define

$$b_1 := \min_{s\in S, a\in A_s} r_s(a), \; b_2 := \max_{s\in S, a\in A_s} r_s(a), \text{ and } B := [b_1, b_2]. \tag{2.9}$$

Then $U_s(x)$ is in B for any strategy profile x starting in any state s.

Shapley (1953) proves that for every two-player zero-sum discounted-payoff stochastic game Γ_s, there exists a unique value Val_s, called the *value of state s*, equal to the expected discounted payoff of Player 1 that she can guarantee by an optimal strategy. Shapley (1953) further shows the existence of a stationary optimal strategy profile, also called a Nash equilibrium; for any stationary optimal strategy profile $\tilde{x}$, Val_s satisfies equations

$$\mathrm{Val}_s = (1-\delta) r_s(\tilde{x}_s) + \delta \sum_{s'\in S} P_{s,s'}(\tilde{x}_s) \mathrm{Val}_{s'} \;\forall s \in S. \tag{2.10}$$

We can also study the asymptotic behaviour in a stochastic game $\Gamma_s(\delta)$ where δ increases to 1. Given a finite stochastic game, for each state $s \in S$, the asymptotic value $\lim_{\delta\to 1} \mathrm{Val}_s(\delta)$ exists; see Bewley and Kohlberg (1976) and Mertens and Neyman (1981).

2.3. An auxiliary game

A central concept in stochastic games is that of the *auxiliary game* formed by composing the stage game payoffs with the expected future discounted payoffs (Shapley, 1953). If Player 1 knows (or assumes) that the future discounted payoff achievable from every state s' is given by $u_{s'}$, then the expected future discounted payoff achievable by playing mixed strategy x_s in state s is given by

$$f_{s,\vec{u}}(x_s) := (1-\delta) r_s(x_s) + \delta \sum_{s' \in S} P_{s,s'}(x_s) u_{s'}, \ \forall x_s \in \Delta_s^1 \times \Delta_s^2, \tag{2.11}$$

where $\vec{u}$ is the vector of *continuation payoffs* u_s. The auxiliary game with payoff $f_{s,\vec{u}}(\cdot)$ is denoted as $G_{s,\vec{u}}$. Since the stage games are zero-sum, $G_{s,\vec{u}}$ is also zero-sum, and Player 2 receives payoff $-f_{s,\vec{u}}(\cdot)$.

To define a best-response dynamic in a stochastic game and to show the convergence, we will apply the continuous-time best-response dynamic in auxiliary games. It will therefore be convenient to consider best responses and energy in these auxiliary games. We denote the best responses in the auxiliary game in state s with the continuation payoff vector $\vec{u}$ by

$$br^1_{s,\vec{u}}(x_s^2) = \operatorname*{argmax}_{\rho_s^1 \in \Delta_s^1} f_{s,\vec{u}}(\rho_s^1, x_s^2) \quad \text{and} \quad br^2_{s,\vec{u}}(x_s^1) = \operatorname*{argmin}_{\rho_s^2 \in \Delta_s^2} f_{s,\vec{u}}(x_s^1, \rho_s^2).$$

Similarly, we denote the energy in the same auxiliary game by

$$w_{s,\vec{u}}(x_s) = \max_{\rho_s^1 \in \Delta_s^1} f_{s,\vec{u}}(\rho_s^1, x_s^2) - \min_{\rho_s^2 \in \Delta_s^2} f_{s,\vec{u}}(x_s^1, \rho_s^2). \tag{2.12}$$

3. The best-response dynamic in a stochastic game

Our first process is a continuous-time dynamical system in which continuation payoffs evolve slowly, while strategies follow a best-response dynamic defined in the auxiliary games. All strategies and continuation payoffs evolve at all times; we will consider a more plausible model of actual play in Section 4. Nevertheless, we can motivate the dynamic in this section as follows. At each time instant, each player knows $x(t)$, i.e., both her own and her opponent's mixed strategies, and estimates each continuation payoff $u_s(t)$ as the average auxiliary game payoff in state s up to time t. Each player i thus learns the current auxiliary games $G_{s,\vec{u}(t)}$ in all states s, and then calculates the auxiliary game payoffs $f_{s,\vec{u}(t)}(x(t))$ for the current mixed strategy as well as the best responses $br^i_{s,\vec{u}(t)}(x^{-i}(t))$. Meanwhile, the strategies are adapted, at constant rate, towards the best responses.

Formally, we pick an arbitrary initial vector $\vec{u}(1) = (u_s(1))_{s \in S}$ with $u_s(1) \in B$ for every $s \in S$, where B is the bounding interval defined in (2.9). Suppose that the initial stationary strategy profile $(x_s(1))_{s \in S}$ is given. We define the following dynamical system for every state $s \in S$ at every time $t \geq 1$

$$\begin{cases} \dot{u}_s(t) = \dfrac{f_{s,\vec{u}(t)}(x_s(t)) - u_s(t)}{t}, & \text{(a)} \\ \dot{x}_s^i(t) \in br^i_{s,\vec{u}(t)}(x_s^{-i}(t)) - x_s^i(t), \quad i = 1, 2, & \text{(b)} \end{cases} \tag{3.1}$$

and call such a dynamical system the *best-response dynamic* in stochastic game Γ. Note that (3.1)(a) is equivalent to $u_s(t) = \int_1^t f_{s,\vec{u}(\tau)}(x(\tau))\,\mathrm{d}\tau$, which is the average auxiliary game payoff up to time t, while (3.1)(b) indicates that $x_s(t)$ follows a best-response dynamic in the auxiliary game $G_{s,\vec{u}(t)}$. We start the dynamic at $t = 1$ simply for notational convenience in (3.1)(a).

Theorem 3.1. *Let Γ be a two-player zero-sum stochastic game, and let $x(t)$ and $\vec{u}(t)$ be any solution trajectory of the best-response dynamic (3.1).*

(i) For each state s, as $t \to \infty$, both $f_{s,\vec{u}(t)}(x_s(t))$ and $u_s(t)$ converge to Val_s, and $x^1(t)$ and $x^2(t)$ converge to the set of stationary optimal strategies of Player 1 and 2, respectively.
(ii) There exists a constant K such that, for all $s \in S$, $|\mathrm{Val}_s - u_s(t)| \le Kt^{-1}$, i.e. the continuation payoffs converge to Val_s at rate t^{-1}.

Sketch proof. The critical observation is that with $|f_{s,\vec{u}(t)}(x_s(t)) - u_s(t)|$ bounded, (3.1)(a) implies that $|\dot{u}_s(t)| \to 0$ as $t \to \infty$. This means that the continuation payoffs $\vec{u}(t)$ move very slowly, and the same energy-based arguments as used in Theorem 2.1 can be used to show that $w_{s,\vec{u}(t)}(x_s(t)) \to 0$. This in turn tells us that

$$|f_{s,\vec{u}(t)}(x_s(t)) - v(G_{s,\vec{u}(t)})| \to 0,$$

by (2.4).

If it were the case that $f_{s,\vec{u}(t)}(x_s(t)) = v(G_{s,\vec{u}(t)})$ then (3.1)(a) would become, essentially, a time rescaling of the scheme of Vigeral (2010); the remainder of our proof of part (i) of the theorem is simply a generalisation of that of Vigeral (2010).

Part (ii) of the theorem simply considers more carefully the bounds we place on the rates of convergence of each part of the dynamical system, and notes that the slowest rate is $1/t$.

The full proof is given in Appendices B and C. □

The dynamical system (3.1)(a)–(3.1)(b) can also be viewed as a feedback system in which $(f_{s,\vec{u}(t)}(a))_{a\in A_s, s\in S}$ transforms strategies to payoffs and the best-response dynamic (3.1)(b) transforms the payoffs back to strategies. Several recent works (e.g. Hofbauer and Sandholm, 2009; Sandholm, 2010; Fox and Shamma, 2013) consider evolutionary dynamics under this separation framework, with Zusai (2019) providing both a helpful summary of the concept, and using it to show the dynamic stability of general "economically reasonable" myopic dynamics in single-population games in which the equilibria are statically stable.

4. Continuous-time state-dependent fictitious play

In this section, we present a continuous-time embedding of actual play in a stochastic game, in which players transition through the state space, and always play an auxiliary-game best response to the current beliefs about opponent strategies. Each player plays an action in the current state at every time instant; the holding time in the current state and the distribution over successor states depends on the players' current actions. In this learning process, players update their actions, beliefs about opponent strategies, and continuation payoffs in continuous time, while playing the continuous-time embedding of the game.

We start by introducing our model of a continuous-time embedding of a stochastic game, which is closely related to the models in Guo and Hernández-Lerma (2005), Levy (2013), and Neyman (2017). In order to ensure that all states are visited at a comparable rate, we restrict to an *irreducible stochastic game* Γ, which requires

$$\min_{s,s'\in S}\left(\min_{a\in A_s} P_{s,s'}(a)\right) > 0.$$

In the continuous-time embedding of an irreducible game Γ, given any states s, s' and a pure action profile a_s, let $q_{s,s'}(a_s)$ be the transition rate from state s to s' when action a_s is being played. Thus if action a_s is played at time t then the probability of a transition from state s to $s' \neq s$ in time $[t, t+h]$ is simply $q_{s,s'}(a_s)h + o(h)$, and the probability of staying in state s during $[t, t+h]$ is $1 + q_{s,s}(a_s)h + o(h)$ (and thus $q_{s,s}(a_s) = -\sum_{s' \neq s} q_{s,s'}(a_s)$). We define a *regular embedding* of a game Γ to satisfy that $q_{s,s'}(a_s) \in (\lambda_{\min}, \lambda_{\max})$ for each tuple (s, s', a_s) with $s' \neq s$, for some $0 < \lambda_{\min} < \lambda_{\max}$. This condition ensures that the holding times are non-pathological. A *consistent regular embedding* of Γ further requires that the transitions in the continuous-time embedding follow the same distribution over successor states as the transitions in the original game: given any s, a_s, and any pair of states s' and s'' both different from s, $P_{s,s'}(a_s)/P_{s,s''}(a_s) = q_{s,s'}(a_s)/q_{s,s''}(a_s)$. The definition of the transition rate can be linearly generalized for a mixed strategy profile.

Consider now two boundedly-rational players playing this continuous-time embedding of the game. At time t they find themselves in state s, with beliefs $x_s^{-i}(t)$ about opponent play in this state. In the spirit of fictitious play, players will play an auxiliary-game best response to these beliefs, which requires the use of some continuation payoffs $\vec{u}$. We make the following modelling assumptions, along the lines of other models of boundedly-rational learning (see, e.g., Harris, 1998):

(i) Each player believes that in each state s the exponentially weighted average play of player $-i$ in state s up to time t is the best estimate of the stage-game mixed strategy in s.

(ii) Each player ignores strategic consideration in the dynamic adaptive process and believes that the realization of her plays in this process will not affect her opponent's predetermined strategy. The players will therefore play best responses to current beliefs; to do so requires some estimate of continuation payoffs.

(iii) Although a player has beliefs about opponent mixed strategy in all states, and could in theory calculate a solution to either (2.8) or Bellman's equation to find self-consistent continuation payoffs, she is unable or unwilling to do so. Hence the player takes the continuation payoffs for each state s' to be the historical time average of the believed auxiliary game payoffs $f_{s',\vec{u}(\cdot)}(x_{s'}(\cdot))$.

The consequences of these assumptions are that, when the players are in a state s at a time t in the learning process, each player i plays a best-response action, denoted by $b_s^i(t)$, to belief $x_s^{-i}(t)$ in auxiliary game $G_{s,\vec{u}(t)}$, as if the payoff against $x_s^{-i}(t)$ in this auxiliary game is the final payoff she will receive in Γ. At the same time, Player i updates her belief $x_s^{-i}(t)$ at a constant rate towards the observed best response, $b_s^{-i}(t)$, of the opponent $-i$, and updates the continuation payoffs $\vec{u}(t)$ in all states, to ensure they are always the average of $f_{s,\vec{u}(t)}(x_s(t))$.

We formalise the dynamics as follows. Define an indicator function $\mathbb{1}_s(t)$ such that $\mathbb{1}_s(t) = 1$ if the players are in state s at time t, otherwise $\mathbb{1}_s(t) = 0$. We pick an arbitrary initial vector $\vec{u}(1) = (u_s(1))_{s \in S}$ with $u_s(1) \in B$ for every $s \in S$, where B is the bounding interval defined in (2.9). Suppose that the initial stationary strategy profile $(x_s(1))_{s \in S}$ is given. The continuation payoffs and beliefs evolve according to

$$\forall s \in S\ \forall t \geq 1, \qquad \begin{cases} \dot{u}_s(t) = \dfrac{f_{s,\vec{u}(t)}(x_s(t)) - u_s(t)}{t} & \text{(a)} \\ \dot{x}_s^i(t) \in \mathbb{1}_s(t)\left(br^i_{s,\vec{u}(t)}(x_s^{-i}(t)) - x_s^i(t)\right). & \text{(b)} \end{cases} \tag{4.1}$$

Equation (4.1)(b) is simply the best-response dynamic (3.1)(b) activated whenever players are in state s; (4.1)(a) ensures that the continuation payoffs are the time average of $f_{s,\vec{u}(t)}(x_s(t))$.

Once again, the continuation payoffs $u_s(t)$ are updated more slowly than the belief $x_s^{-i}(t)$. The continuation vector $\vec{u}(t)$ may be viewed as a preference parameter in auxiliary game $G_{s,\vec{u}(t)}$. In the literature of evolutionary game theory, preference update is often more slowly than behaviour update; see, e.g., Ely and Yilankaya (2001) and Sandholm (2001). The model is therefore consistent with this theory.

A natural question is why we don't assume that players use $u_s(t) = f_{s,\vec{u}(t)}(x_s(t))$ to calculate the best responses in the dynamic instead of $u_s(t)$ evolving towards $f_{s,\vec{u}(t)}(x_s(t))$. Firstly, note that we make a bounded-rationality assumption that players would like to maximise the discounted payoff in the stochastic game but do not know how to calculate (2.8) or solve Bellman's equation. Hence a belief is needed on the continuation payoff in order to calculate a best-response action $b_s^i(t)$ based on this belief and the behaviour of the other player. However, a player knows that neither belief $u_s(t)$ nor payoff $f_{s,\vec{u}(t)}(x_s(t))$ is likely the true discounted payoff $U_s(x(t))$. In this case, she understands that if she forces $u_s(t) = f_{s,\vec{u}(t)}(x_s(t))$, then the resultant new payoff in (2.11) is not the old $f_{s,\vec{u}(t)}(x_s(t))$, which means the guess $u_s(t)$ is not internally consistent until the full Bellman equations are solved; evolving towards reasonable values is a sensible boundedly-rational approach. A secondary consideration is that removing this boundedly-rational assumption, and allowing players to use correct continuation payoffs given current strategy beliefs, is only known to converge when the players are sufficiently impatient (Perkins, 2013).

Under the assumption that players adjust their continuation payoff estimates towards $f_{s,\vec{u}(t)}(x_s(t))$, a second obvious question is why in (4.1)(a) the target payoff is $f_{s,\vec{u}(t)}(x_s(t))$ instead of $f_{s,\vec{u}(t)}(b_s(t))$. After all, action profile $b_s(t)$ is played and the perceived instantaneous payoff should be the latter one. However, we would like to emphasise that the belief about players' actions is $x_s(t)$, and so the current best estimate for the continuation payoff in state s is $f_{s,\vec{u}(t)}(x_s(t))$; the action $b_s^{-i}(t)$ is simply the new information at time t that will be used by player i to update her belief $x_s^{-i}(t)$.

Denote the set of stationary optimal strategy profiles by Z, and define the distance in the space of stationary strategy profiles by the infinity norm.

Theorem 4.1. *Let Γ be a two-player irreducible zero-sum stochastic game, and let $\vec{u}(t)$ and $x(t)$ evolve according to the learning dynamic (4.1) in a regular embedding of Γ. Then, given any $\mu > 0$, there exists a time $\bar{t}$ such that for each $\hat{t} > \bar{t}$,*

$$P\left(|u_s(t) - \mathrm{Val}_s| < \mu \text{ and } ||x(t) - Z||_\infty < \mu \quad \forall t \in [\mu\hat{t}, \hat{t}]\right) > 1 - \mu.$$

The proof is given in Appendix D. We first show in Lemma D.1 and Corollary D.2 that with high probability, for a sufficiently long period, players stay in each state for at least a fixed proportion of that period of time, irrespective of what actions they play in the embedding process. We then build on the proof of Theorem 3.1 to give convergence of first the $x_s(t)$ to a neighbourhood of the auxiliary game equilibria, then the convergence of the continuation payoffs, conditional on the event that all states are updated in a sufficient proportion of the time.

5. The δ-converging best-response dynamic

In addition to its own interest, the study of the value of a zero-sum stochastic games is essential to the study of related non-zero-sum stochastic games; see, e.g., Dutta (1995) and Hörner et al.

(2011). In particular, for Folk Theorem, it is often assumed that players are patient in a non-zero-sum stochastic game; the asymptotic value with the discount factor converging to 1 in the corresponding zero-sum stochastic game gives the limit of individually rational payoff in the non-zero-sum stochastic game.

The asymptotic value exists in a finite zero-sum stochastic game: Bewley and Kohlberg (1976) prove the existence by a semi-algebraic approach, and Oliu-Barton (2014) prove it by an approach of asymptotically optimal strategies. Based on the existence result, we present below a δ-converging best-response dynamic as an adaptive approach to compute the asymptotic value. Note that neither of the previous approaches (Bewley and Kohlberg, 1976; Oliu-Barton, 2014) are readily accessible in computation: the former uses the Tarski-Seidenberg elimination theorem from real algebraic geometry, while the latter needs stationary optimal strategies in an infinite sequence of zero-sum stochastic games.

It is also worth noting that the asymptotic value in discounted payoff is equal to the value in limit average payoff for any finite zero-sum stochastic game. For the formulation of value in limit average payoff, let us first observe that the value exists in a stochastic game where the interaction lasts only for a natural number T stages and the final payoff is the average of these T stage payoffs. If T increases to ∞, then the payoff at each given stage is insignificant as compared to the payoffs in all other stages. Mertens and Neyman (1981) prove that a value exists under the condition that the limsup average stage payoff is applied as T increases to ∞. Moreover, this value is the same as the asymptotic value in discounted payoff when δ increases to 1. So far no direct computational method to reach the value in limit average payoff is available in the literature.

As δ is not a constant in the following model, let us rewrite (2.11):

$$f_{s,\vec{u},\delta}(x_s) = (1-\delta)r_s(x_s) + \delta \sum_{s'\in S} P_{s,s'}(x_s)u_{s'}, \ \forall x_s \in \Delta_s^1 \times \Delta_s^2. \tag{5.1}$$

Similarly to the best-response dynamic in Section 3, pick an arbitrary $\delta(2) \in (0,1)$, and $\vec{u}(2) = (u_s(2))_{s\in S}$ with $u_s(2) \in B$ for each $s \in S$, where B is the bounding interval defined in (2.9) (starting the process at $t = 2$ is once again solely for notational convenience). We show here that given any state s in a zero-sum stochastic game, $u_s(t)$ of any solution trajectory to the following system with initial time $t = 2$ converges to the asymptotic value of Γ_s:

$$\begin{cases} \dot{\delta}(t) = \dfrac{1-\delta(t)}{t\log t} & \text{(a)} \\ \dot{u}_s(t) = \dfrac{f_{s,\vec{u}(t),\delta(t)}(x_s(t)) - u_s(t)}{t} & \text{(b)} \\ \dot{x}_s^i \in br_{s,\vec{u}(t)}^i(x_s^{-i}) - x_s^i, \quad i = 1,2. & \text{(c)} \end{cases} \tag{5.2}$$

We call such a dynamic a *δ-converging best-response dynamic*. Again, one can show the existence of a solution trajectory to the dynamical system from any initial condition $((x_s(2))_{s\in S}, \vec{u}(2), \delta(2))$, by the results in Aubin and Cellina (1984).

Theorem 5.1. *Let Γ be a two-player zero-sum stochastic game, and let $x(t)$, $\vec{u}(t)$ and $\delta(t)$ evolve according to the δ-converging best response dynamic (5.2). Then for each state s, as $t \to \infty$, both $f_{s,\vec{u}(t)}(x(t))$ and $u_s(t)$ converge to the asymptotic value of Γ_s.*

The only difference between this δ-converging best-response dynamic and the best-response dynamic in Section 3 is the evolution of the discount factor $\delta(t)$ given by (5.2)(a). Note that this discount factor adapts even more slowly than both continuation payoffs and players' actions, and is independent of players' actions and continuation payoffs, taking values $\delta(t) = 1 - c(\log t)^{-1}$ for a constant c determined by the initial condition $\delta(2)$. The specific formulation (5.2)(a) is just one example of a sufficiently slow δ-increasing process, satisfying the important condition that $\delta(t) \to 1$ and $\dot{\delta}(t) = o(1/t)$. To see why we need (5.2)(a), first note that the speed difference between (5.2)(b) and (5.2)(c) allows each player to learn an approximately optimal action in each auxiliary game equipped with the current continuation payoff vector, as we have discussed before. The slowness of discount factor adaption allows the continuation payoff vector defined in (5.2)(b) to eventually converge to a small set of vectors in which each one is approximately valid as the continuation payoff vector for all the time when $\delta(t)$ is sufficiently close to 1. The proof is given in Appendix E.

6. Discussion

We note that several alternative approaches to learning in stochastic games might also be considered appropriate. We could translate the stochastic game into a normal-form game with actions equal to the stationary pure strategies of the stochastic game, and payoffs given by the corresponding discounted payoffs $U_s^i(\cdot)$ in the stochastic game, perhaps aggregated over s. Standard learning dynamics can be deployed in the normal-form representation, and will converge since the game is zero-sum. However, a mixed strategy in the normal form does not correspond to a stationary mixed strategy in the stochastic game. To illustrate this point, consider a one-player stochastic game with two states, α and β. β is an absorbing state with stage payoff -4. There are two actions, a and b in state α. If the player selects a then she receives payoff $r_\alpha(a) = 0$ and the state in the next stage is still α with probability 1; if the player selects b then she receives $r_\alpha(b) = 1$ and $P_{\alpha,\alpha}(b) = P_{\alpha,\beta}(b) = \frac{1}{2}$. A mixed strategy in the normal-form representation corresponds to using pure strategy a for all time with probability $1 - \rho$, and pure strategy b for all time with probability ρ, for some $\rho \in [0, 1]$. A stationary mixed strategy in the stochastic game will correspond to selecting a with probability $1 - \rho$ (and b with probability ρ) independently each time state α is encountered. Thus convergence of the dynamics in the normal-form representation does not necessarily result in convergence to a stationary Nash equilibrium in the stochastic game, as the normal-form representation and the original stochastic game are related but different games.

Another natural approach is to note that the stationary strategy space Δ is a compact and convex space. Results of Hofbauer and Sorin (2006) on dynamics in compact and convex strategy spaces might then be applied. Note however that the state transition formulation makes the payoff structure more complex than those studied by Hofbauer and Sorin (2006). In particular, they consider only those games with payoff concave in Player 1's strategy space and convex in Player 2's strategy space. Consider again the game introduced in the previous paragraph. We abuse the notation and denote by ρ the strategy that assigns probability ρ to playing b in state α. The expected discounted payoff in state α satisfies

$$U_\alpha(\rho) = (1-\delta)\rho + \delta\left(\left(1 - \rho + \frac{\rho}{2}\right)U_\alpha(\rho) + \frac{\rho}{2}\cdot(-4)\right).$$

It follows that

$$U_\alpha(\rho) = \frac{1-3\delta}{\frac{1-\delta}{\rho} + \frac{\delta}{2}} \quad \text{and} \quad \frac{d^2 U_\alpha(\rho)}{d\rho^2} = -\frac{(1-3\delta)(1-\delta)\delta}{(1-\delta+\frac{\delta\rho}{2})^3}.$$

If $\delta > 1/3$, then the second derivative is positive, and hence $U_\alpha(\rho)$ is convex in ρ, taking us outside of the framework of Hofbauer and Sorin (2006).

One may also be tempted to apply the convergence result of the best-response dynamic defined on convex/concave envelopes of the payoff function in a continuous quasiconcave-quasiconvex zero-sum game, proved by Barron et al. (2010). However, they also show that the envelopes are necessary by a counterexample that the dynamic may not converge with respect to the payoff function itself. The construction of convex/concave envelopes makes the learning procedure much more complicated than the implementation of best-response strategies only. Relying on these earlier results in normal-form games is thus not appropriate.

There exist other learning methods explicitly designed for stochastic games, such as Szepesvári and Littman (1999), Vrieze and Tijs (1982) and Borkar (2002). Note however that Szepesvári and Littman (1999) requires the solution of a linear program on every iteration of learning, Vrieze and Tijs (1982) presents a somewhat unnatural dynamic relying on very specific starting beliefs, and Borkar (2002)'s results are weaker than ours, albeit using players that require less information about the game. These results can be viewed as computational techniques to find the value.

The most well-known algorithm to compute the value of a zero-sum stochastic game with discounted payoff is still the value iteration process in Shapley (1953). However, this algorithm needs to compute the values of all zero-sum auxiliary games in each round. A continuous-time extension of this value iteration process is presented in Vigeral (2010) as follows. In a zero-sum stochastic game with discounted payoff, the so-called Shapley operator $v_{(\cdot,\cdot)}$ is *nonexpansive*. That is, for each pair of continuation payoff vectors $(\vec{u}, \vec{u}')$,

$$\max_{s\in S} |v_{s,\vec{u}} - v_{s,\vec{u}'}| \leq \delta \max_{s\in S} |u_s - u'_s|.$$

By this property, Vigeral (2010) proves that the dynamic system

$$\dot{u}_s(t) = v_{s,\vec{u}(t)} - u_s(t), \qquad \forall s \in S \tag{6.1}$$

converges to the value of the zero-sum stochastic game. The basic idea of the proof is derived from the property that in the state with the maximum distance of $|v_{s,u\vec{(t)}} - u_s(t)|$, this distance is always decreasing, which follows an intermediate result (B.13) in our proof of Theorem 3.1(i). Vigeral (2010) also shows the convergence of a variation of dynamic (6.1) with discount factor increasing to 1, analogous to our δ-converging result in Section 5. Our results can therefore be considered as a boundedly-rational extension of Vigeral (2010) in which players do not calculate values of games, and simply play best responses to current beliefs; the end product of this myopic adjustment process is an optimal stationary strategy profile, and associated values, in the zero-sum stochastic game.

We would like to emphasize again that our work focuses on stochastic games with discounted payoff. In addition to expected discounted payoffs defined in (2.8), one can also apply limit average payoffs, in which the players only care about the long-run average payoffs, and the payoff at any given stage is insignificant as compared with all the other stage payoffs. Schoenmakers et al. (2007) provide a counterexample demonstrating that a natural fictitious play dynamic need not converge in the case of limit average payoffs, and we leave it an open question as to whether a dynamic such as those present in this article may converge.

Finally we note that our results, like the vast majority of those in learning in games, consider the setting where all players use the same algorithm. Stronger results would provide consistency results for a learner that deploys the algorithm without knowing what algorithm the other players would use, along the lines of Fudenberg and Levine (2014). However we are aware of no results along these lines that apply to stochastic games.

Appendix A. Properties of zero-sum normal-form games

We present two standard preliminary results for a zero-sum normal-form game G with payoff function u.

Lemma A.1. *Given a positive finite number c, if we modify the payoff function u to u' with the property $|u'(a^1,a^2)-u(a^1,a^2)|\le c$ for all $(a^1,a^2)\in A^1\times A^2$, then for any (mixed) strategy profile (x^1,x^2), $|u'(x^1,x^2)-u(x^1,x^2)|\le c$.*

Proof. This follows from the linear property of u. □

Lemma A.2. *Given a positive finite number c, if we modify the payoff function u to $\bar{u}$ with the property $|\bar{u}(a^1,a^2)-u(a^1,a^2)|\le c$ for all $(a^1,a^2)\in A^1\times A^2$, then $|v(\bar{G})-v(G)|\le c$, where $\bar{G}$ is the game with the modified payoff function $\bar{u}$.*

Proof. For any optimal strategy profile (x^1,x^2) in G and any optimal strategy profile $(\bar{x}^1,\bar{x}^2)$ in $\bar{G}$, we have

$$u(x^1,\bar{x}^2)\ge u(x^1,x^2) \text{ and } \bar{u}(\bar{x}^1,\bar{x}^2)\ge \bar{u}(x^1,\bar{x}^2).$$

Thus

$$u(x^1,x^2)-\bar{u}(\bar{x}^1,\bar{x}^2)\le u(x^1,\bar{x}^2)-\bar{u}(x^1,\bar{x}^2)\le \max_{a^1,a^2}|\bar{u}(a^1,a^2)-u(a^1,a^2)|\le c,$$

by Lemma A.1. Similarly, we can show that

$$\bar{u}(\bar{x}^1,\bar{x}^2)-u(x^1,x^2)\le \max_{a^1,a^2}|\bar{u}(a^1,a^2)-u(a^1,a^2)|\le c. \quad \square$$

Appendix B. Proof of Theorem 3.1(i)

With similar argument to the standard best response differential inclusion (2.2), from any initial condition $(x_s(1),u_s(1))_{s\in S}$, there exists a solution trajectory $(x_s(t),u_s(t))_{s\in S,t\ge 1}$ for the best-response dynamic (3.1), where $x_s(t)$ and $f_{s,\vec{u}(t)}(x_s(t))$ are differentiable for almost all $t\ge 1$ in all states s; see Aubin and Cellina (1984). It then follows that the derivatives of $v_{s,\vec{u}(t)}$ exist for almost all $t\ge 1$ at all s. Fix a solution trajectory $(u_s(t),x_s(t))_{s\in S,t\ge 1}$ throughout the proof. For each state $s\in S$, at each time $t\ge 0$, we denote the value of the auxiliary game $G_{s,\vec{u}(t)}$ by $v_{s,\vec{u}(t)}:=v(G_{s,\vec{u}(t)})$, which is defined in (2.1), and recall from (2.12) that the energy in $G_{s,\vec{u}(t)}$ under $x_s(t)$ is denoted by $w_{s,\vec{u}(t)}(x_s(t))$. We study this energy before proving the convergence of the auxiliary game play $x_s(t)$.

Lemma B.1. *For any state s, $w_{s,\vec{u}(t)}(x_s(t))$ is Lipschitz continuous with respect to t.*

Proof. It is clear from the definition that $f_{s,\vec{u}}(x_s)$ is Lipschitz with respect to both $\vec{u}$ and x_s. Both $\vec{u}(t)$ and $x_s(t)$ are Lipschitz with respect to t, by the definition of a trajectory. Hence $f_{s,\vec{u}(t)}(x_s(t))$ is Lipschitz with respect to t. From Theorem A.4 in Hofbauer and Sandholm (2009), it follows that both $\max_{\rho^1 \in \Delta_s^1} f_{s,\vec{u}(t)}(\rho^1, x_s^2(t))$ and $\min_{\rho^2 \in \Delta_s^2} f_{s,\vec{u}(t)}(x_s^1(t), \rho^2)$ are Lipschitz continuous with respect to t. Therefore,

$$w_{s,\vec{u}(t)}(x_s(t)) = \max_{\rho^1 \in \Delta_s^1} f_{s,\vec{u}(t)}(\rho^1, x_s^2(t)) - \min_{\rho^2 \in \Delta_s^2} f_{s,\vec{u}(t)}(x_s^1(t), \rho^2)$$

is Lipschitz continuous with respect to t. □

From the definition (3.1)(a) of the dynamical system, $\dot{u}_s(t)$ exists everywhere for all states s. From definitions (2.11) and (2.12) of the energy for auxiliary game $G_{s,\vec{u}(t)}$, we observe that $D_{\vec{u}} w_{s,\vec{u}(t)}(x_s(t))$ always exists. Finally, from (2.7) in the proof of Theorem 2.1, we may infer that $\dot{x}_s D_{x_s} w_{s,\vec{u}(t)}(x_s(t))$ exists for almost all t. We can then conclude by the chain rule that

$$\tfrac{\mathrm{d}}{\mathrm{d}t} w_{s,\vec{u}(t)}(x_s(t)) = \dot{\vec{u}} \cdot D_{\vec{u}} w_{s,\vec{u}(t)}(x_s(t)) + \dot{x}_s D_{x_s} w_{s,\vec{u}(t)}(x_s(t)) \tag{B.1}$$

holds for almost all t. Throughout the proofs in the present paper, all statements about derivatives are to be taken to hold where the derivatives exist, which is everywhere except on a set of time of measure 0.

Lemma B.2. *For each state s in S, $|f_{s,\vec{u}(t)}(x_s(t)) - v_{s,\vec{u}(t)}| \to 0$ as t increases to infinity.*

Proof. First note that, by (2.4), $|f_{s,\vec{u}(t)}(x_s(t)) - v_{s,\vec{u}(t)}| \to 0$ is an immediate consequence of $w_{s,\vec{u}(t)}(x_s(t)) \to 0$, which we prove below by extending Theorem 2.1.

Suppose that an arbitrarily small $\epsilon > 0$ is given. The definitions of the bounding constants b_1 and b_2 in (2.9) imply that in any state s,

$$|f_{s,\vec{u}(t)}(x_s(t)) - u_s(t)| \le b_2 - b_1, \ \forall t \ge 1. \tag{B.2}$$

Therefore, it follows from the definition of the dynamic (3.1)(a) that there exists $t_\epsilon > 1$ such that

$$|\dot{u}_s(t)| = \frac{|f_{s,\vec{u}(t)}(x_s(t)) - u_s(t)|}{t} \le \epsilon \ \forall t \ge t_\epsilon, \ \forall s \in S. \tag{B.3}$$

Note, from (2.11) and (2.12), that a change in continuation payoffs $\vec{u}$ with maximal change ϵ corresponds to a change in $w_{s,\vec{u}}(x)$ of at most $2\delta\epsilon$. Hence $\dot{\vec{u}} \cdot D_{\vec{u}} w_{s,\vec{u}(t)}(x_s(t)) \le 2\delta \max_{s' \in S} \dot{u}_{s'}$. Furthermore, Harris (1998) and Hofbauer and Sorin (2006) show that

$$\dot{x}_s D_{x_s} w_{s,\vec{u}(t)}(x_s(t)) \le -w_{s,\vec{u}(t)}(x_s(t)). \tag{B.4}$$

Therefore, (B.1) implies that

$$\dot{w}_{s,\vec{u}(t)} \le -w_{s,\vec{u}(t)} + 2\delta\epsilon \tag{B.5}$$

for all time $t \ge t_\epsilon$ and for all $s \in S$. This in turn implies that, for a sufficiently large t, $w_{s,\vec{u}(t)}(x_s(t)) < 2\epsilon$ for all states $s \in S$.[1] Since $\epsilon > 0$ is arbitrarily small, $w_{s,\vec{u}(t)}(x_s(t))$ converges to 0, and the result follows. □

Lemma B.2 shows that for large t the auxiliary game play will be close to the equilibrium determined by current continuation payoffs. Note that (B.4) is the only line in the proof of Theorem 3.1 where we use a property of the best-response dynamic (3.1)(b), and other revision

[1] If $w_{s,\vec{u}(t)}(x_s(t)) \ge (1+\delta)\epsilon$ then, by (B.5), $\dot{w}_{s,\vec{u}(t)} \le -(1-\delta)\epsilon$. So, eventually, $w_{s,\vec{u}(t)}(x_s(t)) \le (1+\delta)\epsilon$. Once $w_{s,\vec{u}(t)}(x_s(t))$ is less than or equal to $(1+\delta)\epsilon$ it will never increase above this level again. The result follows.

protocols that give rise to the conclusion of Lemma B.2 would also result in an equivalent of Theorem 3.1(i). For the rest of the proof, we only need the formulation of continuation payoff adjustment (3.1)(a) and the auxiliary game structure (2.11).

Let $\epsilon > 0$ be arbitrary, and let $t_1(\epsilon)$ be such that for all $t \geq t_1(\epsilon)$ and all states s in S,

$$|f_{s,\vec{u}(t)}(x_s(t)) - v_{s,\vec{u}(t)}| \leq (1-\delta)\epsilon/16. \tag{B.6}$$

Such a $t_1(\epsilon)$ exists by Lemma B.2. For the rest of the proof we will assume that $t \geq t_1(\epsilon)$ and hence that (B.6) holds.

It remains to show that the continuation payoffs will converge to the correct values, i.e. those of a Nash equilibrium. This part of the proof extends the approach of Vigeral (2010), who proves that continuation payoffs converge to equilibrium values if the payoff adjustment dynamics (3.1)(a) are modified to (6.1) so that $u_s(t)$ moves in the direction of the value of the auxiliary game instead of in the direction of the current payoff in the auxiliary game. We start with some preliminary definitions:

- For any time $t \geq 1$, we mark a state

  $$s_f(t) \in \operatorname*{argmax}_{s \in S} |f_{s,\vec{u}(t)}(x_s(t)) - u_s(t)|, \tag{B.7}$$

  which, by (3.1)(a), implies that

  $$s_f(t) \in \operatorname*{argmax}_{s \in S} |\dot{u}_s(t)|. \tag{B.8}$$

- We also, for any time $t \geq 1$, mark a state

  $$s_v(t) \in \operatorname*{argmax}_{s \in S} |v_{s,\vec{u}(t)} - u_s(t)|. \tag{B.9}$$

Recall that Lemma B.2 shows that $f_{s,\vec{u}(t)}(x(t))$ becomes close to $v_{s,\vec{u}(t)}(t)$ for all s. By showing that $|v_{s_v(t),\vec{u}(t)} - u_{s_v(t)}(t)| \to 0$ and $|f_{s_f(t),\vec{u}(t)}(x_{s_f(t)}(t)) - u_{s_f(t)}(t)| \to 0$ we will show that, in the limit, for each s, all of $f_{s,\vec{u}(t)}(x_s(t))$, $u_s(t)$ and $v_{s,\vec{u}(t)}$ are equal. This is sufficient to prove the theorem. Below is a technical lemma.

Lemma B.3. *At any time $t \geq t_1(\epsilon)$, if*

$$|u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)| \geq \epsilon, \tag{B.10}$$

then for any state s with the property

$$\left||u_{s_f(t)}(t) - v_{s_f(t),\vec{u}(t)}| - |u_s(t) - v_{s,\vec{u}(t)}|\right| \leq \frac{(1-\delta)\epsilon}{8}, \tag{B.11}$$

we have

$$\tfrac{d}{dt}|u_s(t) - v_{s,\vec{u}(t)}| \leq -\frac{3(1-\delta)\epsilon}{4t}. \tag{B.12}$$

This lemma says that if the maximal distance between $u_s(t)$ and $f_{s,\vec{u}(t)}x_s(t)$ is big enough, then the absolute value between some $u_s(t)$ and $v_{s,\vec{u}(t)}$ is decreasing at a rate at least linear in $1/t$. (Since this rate would result in the absolute value becoming negative, condition (B.10) cannot always hold, as we will see in Lemma B.4.)

Proof. From Lemma A.2 and the definition (B.8) of $s_f(t)$ as the maximiser of $|\dot{u}_s(t)|$, it follows that

$$\forall s \in S, \ \left|\dot{v}_{s,\vec{u}(t)}\right| \leq \delta \max_{s\in S} |\dot{u}_s(t)| = \delta \left|\dot{u}_{s_f(t)}(t)\right|. \tag{B.13}$$

Now fix a state s with the property (B.11) at time $t \geq t_1(\epsilon)$. We may infer from (B.11) and the fact that $|v_{s,\vec{u}(t)} - f_{s,\vec{u}(t)}(x_s(t))| \leq (1-\delta)\epsilon/16$ for all s, by (B.6), that

$$\begin{aligned}
&|u_s(t) - f_{s,\vec{u}(t)}(x_s(t))| - |u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)| \\
\geq &(|u_s(t) - v_{s,\vec{u}(t)}| - |v_{s,\vec{u}(t)} - f_{s,\vec{u}(t)}(x_s(t))|) \\
&- (|u_{s_f(t)}(t) - v_{s_f(t),\vec{u}(t)}| + |v_{s_f(t),\vec{u}(t)} - f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)|) \\
\geq &|u_s(t) - v_{s,\vec{u}(t)}| - |u_{s_f(t)}(t) - v_{s_f(t),\vec{u}(t)}| - \frac{(1-\delta)\epsilon}{8} \\
\geq &-\frac{(1-\delta)\epsilon}{4}.
\end{aligned} \tag{B.14}$$

Thus, from the dynamic (3.1)(a) for $u_s(t)$, it follows that for this s,

$$\begin{aligned}
|\dot{u}_s(t)| &= \frac{\left|f_{s,\vec{u}(t)}(x_s(t)) - u_s(t)\right|}{t} \\
&\geq \frac{\left|f_{s_f(t),\vec{u}(t)}(x_{s_f(t)}(t)) - u_{s_f(t)}(t)\right|}{t} - \frac{(1-\delta)\epsilon}{4t} \\
&\geq \left|\dot{u}_{s_f(t)}(t)\right| \left(1 - \frac{1-\delta}{4}\right) && \text{(B.15)} \\
&= \left|\dot{u}_{s_f(t)}(t)\right| \frac{3+\delta}{4}, && \text{(B.16)}
\end{aligned}$$

where the last inequality holds since $\left|\dot{u}_{s_f(t)}(t)\right| \geq \epsilon/t$ by (B.10) and (3.1)(a). Combining our inequalities (B.13) and (B.16) we see that

$$\left|\dot{v}_{s,\vec{u}(t)}\right| \leq \frac{4\delta}{3+\delta} |\dot{u}_s(t)| < |\dot{u}_s(t)|. \tag{B.17}$$

Suppose now that $u_s(t) > v_{s,\vec{u}(t)}$. The closeness of $v_{s,\vec{u}(t)}$ and $f_{s,\vec{u}(t)}(x_s(t))$ given by (B.6), along with the conditions (B.10) and (B.11) of the lemma, give that

$$\begin{aligned}
&u_s(t) - v_{s,\vec{u}(t)} \\
\geq &|u_{s_f(t)}(t) - v_{s_f(t),\vec{u}(t)}| - \frac{(1-\delta)\epsilon}{8} \\
\geq &\left(|u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)| - |f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right) - v_{s_f(t),\vec{u}(t)}|\right) - \frac{(1-\delta)\epsilon}{8} \\
\geq &\left(1 - \frac{3(1-\delta)}{16}\right)\epsilon.
\end{aligned}$$

Invoking (B.6) once more we see that

$$u_s(t) \geq f_{s,\vec{u}(t)}(x_s(t)) + \left(1 - \frac{(1-\delta)}{4}\right)\epsilon > f_{s,\vec{u}(t)}(x_s(t))$$

and so, by the definition of the dynamic (3.1)(a), $\dot{u}_s(t) = (f_{s,\vec{u}(t)}(x(t)) - u_s(t))/t < 0$. Combined with (B.17), this implies that

$$\tfrac{d}{dt}\left(u_s(t) - v_{s,\vec{u}(t)}\right) < 0.$$

Recalling the lower bound of $|\dot{u}_s(t)|$ in (B.16) and the upper bound of $\left|\dot{v}_{s,\vec{u}(t)}\right|$ in (B.13), we may then infer that

$$\tfrac{d}{dt}\left(u_s(t) - v_{s,\vec{u}(t)}\right) \leq -\frac{3+\delta}{4}\left|\dot{u}_{s_f(t)}(t)\right| + \delta\left|\dot{u}_{s_f(t)}(t)\right| = -\frac{3}{4}(1-\delta)\left|\dot{u}_{s_f(t)}(t)\right|.$$

A near-identical calculation shows the same conclusion if $u_s(t) < v_{s,\vec{u}(t)}$. Thus, we have

$$\tfrac{d}{dt}\left|u_s(t) - v_{s,\vec{u}(t)}\right| \leq -\frac{3}{4}(1-\delta)\left|\dot{u}_{s_f(t)}(t)\right|. \tag{B.18}$$

The result then follows on noting, once again, that $\left|\dot{u}_{s_f(t)}(t)\right| = |f_{s_f(t),\vec{u}(t)}(x_{s_f(t)}(t)) - u_{s_f(t)}|/t \geq \epsilon/t$ using (B.10) to bound $|f_{s_f(t),\vec{u}(t)}(x_{s_f(t)}(t)) - u_{s_f(t)}|$ below by ϵ. □

This now puts us in a position to prove the important final lemma.

Lemma B.4. *There exists time $t_2(\epsilon)$ such that for all $t \geq t_2(\epsilon)$,*

$$\max_{s\in S}|u_s(t) - v_{s,\vec{u}(t)}| = |u_{s_v(t)}(t) - v_{s_v(t),\vec{u}(t)}| \leq \left(1 + \frac{3(1-\delta)}{16}\right)\epsilon \tag{B.19}$$

and

$$\max_{s\in S}|u_s(t) - f_{s,\vec{u}(t)}(x_s(t))| = |u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)| \leq \left(1 + \frac{1-\delta}{2}\right)\epsilon. \tag{B.20}$$

Proof. Fix $\epsilon > 0$, let $t_1(\epsilon)$ be defined as in (B.6), so that $|f_{s,\vec{u}(t)}(x_s(t)) - v_{s,\vec{u}(t)}| \leq (1-\delta)\epsilon/16$ for all s for $t \geq t_1(\epsilon)$. We start by showing that condition (B.11) of Lemma B.3 always holds for state $s_v(t)$ when $t \geq t_1(\epsilon)$:

$$\begin{aligned}
&\left||u_{s_f(t)}(t) - v_{s_f(t),\vec{u}(t)}| - |u_{s_v(t)}(t) - v_{s_v(t),\vec{u}(t)}|\right| \\
&\quad = |u_{s_v(t)}(t) - v_{s_v(t),\vec{u}(t)}| - |u_{s_f(t)}(t) - v_{s_f(t),\vec{u}(t)}| \\
&\quad \leq (|u_{s_v(t)}(t) - f_{s_v(t),\vec{u}(t)}(x_{s_v(t)}(t))| + |f_{s_v(t),\vec{u}(t)}(x_{s_v(t)}(t)) - v_{s_v(t),\vec{u}(t)}|) \\
&\qquad - (|u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}(x_{s_f(t)}(t))| - |f_{s_f(t),\vec{u}(t)}(x_{s_f(t)}(t)) - v_{s_f(t),\vec{u}(t)}|) \\
&\quad \leq (|u_{s_v(t)}(t) - f_{s_v(t),\vec{u}(t)}(x_{s_v(t)}(t))| - |u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}(x_{s_f(t)}(t))|) + \frac{(1-\delta)\epsilon}{8} \\
&\quad \leq \frac{(1-\delta)\epsilon}{8},
\end{aligned} \tag{B.21}$$

where the penultimate inequality is since $t \geq t_1(\epsilon)$, so that $|f_{s,\vec{u}(t)}(x_s(t)) - v_{s,\vec{u}(t)}| \leq (1-\delta)\epsilon/16$, and the final inequality is because $s_f(t)$ maximizes $|u_s(t) - f_{s,\vec{u}(t)}(x_s(t))|$.

Since condition (B.11) holds for $s_v(t)$, it follows that, for any $t \geq t_1(\epsilon)$ such that property (B.10) holds (i.e. $|u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)| \geq \epsilon$), the conclusion of Lemma B.3 holds for $s_v(t)$, i.e.

$$\frac{d|u_{s_v(t)}(t) - v_{s_v(t),\vec{u}(t)}|}{dt} \leq -\frac{3(1-\delta)\epsilon}{4t}. \tag{B.22}$$

Hence, since $|u_{s_v(t)}(t) - v_{s_v(t),\vec{u}(t)}|$ is bounded below, there must exist a time $t_2(\epsilon) \geq t_1(\epsilon)$ at which property (B.10) ceases to hold, i.e. such that

$$|u_{s_f(t_2(\epsilon))}(t_2(\epsilon)) - f_{s_f(t_2(\epsilon)),\vec{u}(t_2(\epsilon))}\left(x_{s_f(t_2(\epsilon))}(t_2(\epsilon))\right)| < \epsilon.$$

Since $|v_{s,\vec{u}(t)} - f_{s,\vec{u}(t)}(x(t))| \leq (1-\delta)\epsilon/16$ when $t \geq t_1(\epsilon)$ by (B.6), and by (B.21), at time $t_2(\epsilon)$ we have that

$$|u_{s_v(t_2(\epsilon))}(t_2(\epsilon)) - v_{s_v(t_2(\epsilon)),\vec{u}(t_2(\epsilon))}| \leq \left(1 + \frac{3(1-\delta)}{16}\right)\epsilon.$$

So far, we have shown that there exists a time $t_2(\epsilon)$ when the desired result holds. We now show that the desired result holds for arbitrary $t > t_2(\epsilon)$, by checking two cases.

Case 1: (B.10) does not hold at t, so that $|u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)| < \epsilon$, and (B.20) follows immediately. Furthermore, by (B.6) and (B.21),

$$|u_{s_v(t)}(t) - v_{s_v(t),\vec{u}(t)}| \leq \left(1 + \frac{3(1-\delta)}{16}\right)\epsilon.$$

Case 2: (B.10) holds at t. Then, the existence of time $t_2(\epsilon)$ implies that there exists a time $t_3(\epsilon)$ with $t_2(\epsilon) < t_3(\epsilon) \leq t$ such that

$$|u_{s_f(t_3^-(\epsilon))}(t_3^-(\epsilon)) - f_{s_f(t_3^-(\epsilon)),\vec{u}(t_3^-(\epsilon))}\left(x_{s_f(t_3^-(\epsilon))}(t_3^-(\epsilon))\right)| < \epsilon,$$

where $t_3^-(\epsilon)$ denotes the left limit of $t_3(\epsilon)$, and

$$|u_{s_f(t_3(\epsilon))}(t_3(\epsilon)) - f_{s_f(t_3(\epsilon)),\vec{u}(t_3(\epsilon))}\left(x_{s_f(t_3(\epsilon))}(t_3(\epsilon))\right)| = \epsilon.$$

Without loss of generality, we assume that (B.10) holds throughout the time period $[t_3(\epsilon), t]$. By the continuity of u and v, we may infer from Case 1 that

$$|u_{s_v(t_3(\epsilon))}(t_3(\epsilon)) - v_{s_v(t_3(\epsilon)),\vec{u}(t_3(\epsilon))}| \leq \left(1 + \frac{3(1-\delta)}{16}\right)\epsilon.$$

From (B.22), it further implies that

$$\forall t' \in [t_3, t],\ |u_{s_v(t')}(t') - v_{s_v(t'),\vec{u}(t')}| \leq \left(1 + \frac{3(1-\delta)}{16}\right)\epsilon. \tag{B.23}$$

To show (B.20), from (B.21) and (B.6), we may infer that for all $t' \in [t_3(\epsilon), t]$,

$$\left||u_{s_f(t')}(t') - f_{s_f(t'),\vec{u}(t')}\left(x_{s_f(t')}(t')\right)| - |u_{s_v(t')}(t') - v_{s_v(t'),\vec{u}(t')}|\right| \leq \frac{3(1-\delta)\epsilon}{16}. \tag{B.24}$$

From (B.23) and (B.24), it follows that

$$\forall t' \in [t_3, t],\ |u_{s_f(t')}(t') - f_{s_f(t'),\vec{u}(t')}\left(x_{s_f(t')}(t')\right)| \leq \left(1 + \frac{3(1-\delta)}{8}\right)\epsilon. \quad \square$$

Proof of Theorem 3.1(i). From Lemma B.4, we see that for each state s,

$$|f_{s,\vec{u}(t)}(x_s(t)) - u_s(t)| \to 0 \text{ and } |u_s(t) - v_{s,\vec{u}(t)}| \to 0, \text{ as } t \to \infty. \tag{B.25}$$

Let Z denote the set of optimal strategy profiles for zero-sum auxiliary games $G_{s,(\mathrm{Val}_s)_{s\in S}}$, where vector $(\mathrm{Val}_s)_{s\in S}$ is a solution to equation (2.10). It follows that $x(t)$ converges to the set Z as $t \to \infty$. We now only need to know that Val_s is unique for each s, and each $z \in Z$ is an optimal strategy profile in stochastic game Γ_s regardless of the initial state s. This is proved in Theorem 2 of Shapley (1953). □

Appendix C. Proof of Theorem 3.1(ii)

Proof of Theorem 3.1(ii): We consider any solution trajectory $(x_s(t), u_s(t))_{s\in S, t\geq 1}$ of the best-response dynamic in a zero-sum stochastic game with a discount factor $\delta < 1$. From Lemma B.2, it follows that $|f_{s,\vec{u}(t)}(x_s(t)) - v_{s,\vec{u}(t)}|$ decreases to 0 in all states s. We can adopt a similar approach to Harris (1998) to find the convergence rate.

Tightening up the analysis in Lemma B.2, from (3.1)(a), (B.2), and (B.1), we may infer that

$$\forall s,\ \dot{w}_s \leq -w_s + \frac{2\delta(b_2 - b_1)}{t}.$$

Note that when $w_s(t) > \frac{4\delta(b_2-b_1)}{t}$, $\dot{w}_s < -\frac{w_s}{2} < -\frac{2\delta(b_2-b_1)}{t}$. Thus, as before, $w_s(t)$ converges to 0 at rate $1/t$, and hence $f_{s,\vec{u}(t)}(x_s(t))$ converges to $v_{s,\vec{u}(t)}$ at rate $1/t$ in all states s. We now consider two cases.

Case 1: There exists time $\bar{t}$ such that for all $t > \bar{t}$,

$$\begin{aligned}
&\max\{|f_{s_v(t),\vec{u}(t)}\left(x_{s_v(t)}(t)\right) - v_{s_v(t),\vec{u}(t)}|, |f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right) - v_{s_f(t),\vec{u}(t)}|\} \\
&> \frac{1-\delta}{4}|u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)|. \qquad \text{(C.1)}
\end{aligned}$$

Then for all $s \in S$, $u_s(t)$, $f_{s,\vec{u}(t)}(x_s(t))$, and $v_{s,\vec{u}(t)}$ all converge at the same rate as of $f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)$ converging to $v_{s_f(t),\vec{u}(t)}$, i.e., $1/t$. Together with the argument in Theorem 2 in Shapley (1953), as we have used in the proof of Theorem 3.1.i, they all converge to Val_s at rate $1/t$.

Case 2: For any $\bar{t}$, there exists some time $t > \bar{t}$ at which (C.1) does not hold. Then, at this t, by the definition of $s_v(t)$ in (B.9),

$$\begin{aligned}
&|u_{s_v(t)}(t) - f_{s_v(t),\vec{u}(t)}\left(x_{s_v(t)}(t)\right)| - |u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)| \\
\geq&(|u_{s_v(t)}(t) - v_{s_v(t),\vec{u}(t)}| - |f_{s_v(t),\vec{u}(t)}\left(x_{s_v(t)}(t)\right) - v_{s_v(t),\vec{u}(t)}|) \\
&-(|u_{s_f(t)}(t) - v_{s_f(t),\vec{u}(t)}| + |f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right) - v_{s_f(t),\vec{u}(t)}|) \\
\geq& -\frac{1-\delta}{2}|u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)|.
\end{aligned}$$

Thus,

$$|u_{s_v(t)}(t) - f_{s_v(t),\vec{u}(t)}\left(x_{s_v(t)}(t)\right)| \geq \left(1 - \frac{1-\delta}{2}\right)|u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)|.$$

Similarly to (B.15), we may further infer that

$$\begin{aligned}
\left|\frac{du_{s_v(t)}(t)}{dt}\right| &= \frac{|f_{s_v(t),\vec{u}(t)}\left(x_{s_v(t)}(t)\right) - u_{s_v(t)}(t)|}{t} \\
&\geq \frac{\left(1 - \frac{1-\delta}{2}\right)|u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)|}{t} \\
&= \left(1 - \frac{1-\delta}{2}\right)\left|\frac{du_{s_f(t)}(t)}{dt}\right|.
\end{aligned}$$

Along the argument in the proof of Lemma B.3, we have

$$\left|\frac{dv_{s_v(t),\vec{u}(t)}}{dt}\right| \leq \delta\left|\frac{du_{s_f(t)}(t)}{dt}\right|$$

by (B.13), and we can further deduce that

$$\begin{aligned}\frac{d|u_{s_v(t)}(t)-v_{s_v(t),\vec{u}(t)}|}{dt} &\leq -\left|\frac{du_{s_v(t)}(t)}{dt}\right| + \left|\frac{dv_{s_v(t),\vec{u}(t)}}{dt}\right| \\ &\leq \left(-\left(1-\frac{1-\delta}{2}\right)+\delta\right)\left|\frac{du_{s_f(t)}(t)}{dt}\right| \\ &= -\frac{1-\delta}{2}\left|\frac{du_{s_f(t)}(t)}{dt}\right| \\ &= -\frac{1-\delta}{2t}|f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)-u_{s_f(t)}(t)| \qquad \text{(C.2)}\end{aligned}$$

From the assumption that (C.1) does not hold, it follows that

$$\begin{aligned}&|f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)-u_{s_f(t)}(t)| \\ \geq&|f_{s_v(t),\vec{u}(t)}\left(x_{s_v(t)}(t)\right)-u_{s_v(t)}(t)| \\ \geq&|u_{s_v(t)}(t)-v_{s_v(t),\vec{u}(t)}|-|f_{s_v(t),\vec{u}(t)}\left(x_{s_v(t)}(t)\right)-v_{s_v(t),\vec{u}(t)}| \\ \geq&|u_{s_v(t)}(t)-v_{s_v(t),\vec{u}(t)}|-\frac{1-\delta}{4}|f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)-u_{s_f(t)}(t)|,\end{aligned}$$

and hence

$$|f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)-u_{s_f(t)}(t)| \geq \frac{|u_{s_v(t)}(t)-v_{s_v(t),\vec{u}(t)}|}{1+\frac{1-\delta}{4}}.$$

Combined with (C.2), we observe that

$$\frac{d|u_{s_v(t)}(t)-v_{s_v(t),\vec{u}(t)}|}{dt} \leq -\frac{1-\delta}{2t}\frac{|u_{s_v(t)}(t)-v_{s_v(t),\vec{u}(t)}|}{1+\frac{1-\delta}{4}}.$$

Thus, $u_{s_v(t)}(t)$ converges to $v_{s_v(t),\vec{u}(t)}$ at rate $1/t$. Recall that $f_{s,\vec{u}(t)}(x_s(t))$ converges to $v_{s,\vec{u}(t)}$ at rate $1/t$ in all states s. Therefore, for all $s \in S$, $u_s(t)$, $f_{s,\vec{u}(t)}(x_s(t))$, and $v_{s,\vec{u}(t)}$ all converge at rate $1/t$. Together with the argument in Theorem 2 in Shapley (1953), as we have used in the proof of Theorem 3.1.i, they all converge to Val_s at rate $1/t$.

Appendix D. Proof of Theorem 4.1

We start by proving a seemingly-obvious result about the occupation times of states in a controlled Markov chain, which is needed to ensure that the action is updated sufficiently frequently in every state despite only the action at the current state being updated at any particular time. The reason we cannot apply standard ergodicity results directly to the controlled Markov chain is because the transition rates of the chain are probably continually evolving under the control parameter $x(t)$; the result is likely to already exist elsewhere, but we have not managed to find it and hence include the proof here for completeness.

Lemma D.1. *Consider a continuous-time controlled Markov chain on a finite state space S. Let the transition rates between states s and s' be given by $q_{s,s'}(x(t))$ where $x(t)$ is an arbitrary control parameter, and define*

$$q_s(x(t)) := -q_{s,s}(x(t)) = \sum_{s'\neq s} q_{s,s'}(x(t)).$$

Assume that:

- *there exists $\eta > 0$ such that $q_{s,s'}(x(t))/q_s(x(t)) \geq \eta$ for all s, s', and t, so that when a jump occurs the probability of jumping to any state is bounded below by η, and*
- *there exist $\lambda_{\min}$ and $\lambda_{\max}$ such that $0 < \lambda_{\min} < q_s(x(t)) < \lambda_{\max}$ for all s and t, so that the holding times in states are well-behaved.*

Let $Q > 0$ and $\epsilon > 0$. Then, there exists a $\Delta T > 0$ such that for all $T \geq 0$, all $s \in S$, and irrespective of $x(t)_{t\geq 0}$,

$$P\left(\int_T^{T+\Delta T} \mathbb{1}_s(t)\,\mathrm{d}t \geq Q\right) \geq 1-\epsilon. \tag{D.1}$$

Proof. We construct a proof using a coupling argument, linking our original process to one in which simple renewal-reward arguments (e.g., Grimmett and Stirzaker, 2001) show the probability of the event we care about is sufficiently high. Throughout, we assume nothing about the control parameter $x(t)$, and we show that our result holds irrespective of $x(t)$.

First note that our Markov model can be implemented using a sequence of independent uniform random variables as follows. If the kth state is s_k and the process arrives here at time t_k, a uniform random variable $U_k \sim \text{Unif}(0, 1)$ is sampled; the state remains at s_k until t_{k+1} which satisfies

$$\int_{t_k}^{t_{k+1}} q_{s_k}(x(\tau))\,\mathrm{d}\tau = -\log(1-U_k); \tag{D.2}$$

a further uniform random variable $V_{k+1} \sim \text{Unif}(0, 1)$ is then sampled to determine the state transition, with the next state s_{k+1} being selected using the inverse cumulative distribution function method on the probability mass function $\left(q_{s_k,s}(x(t_{k+1}))/q_{s_k}(x(t_{k+1}))\right)_{s\neq s_k}$. If transition rates did not depend on $x(t)$, (D.2) would result in the standard exponential holding times

$$t_{k+1} - t_k = -\frac{\log(1-U_k)}{q_{s_k}},$$

with jump chain transition probabilities $q_{s_k,s_{k+1}}/q_{s_k}$, as in Grimmett and Stirzaker (2001, Section 6.9). When we have non-constant transition rates, it is an easy calculation to see that the instantaneous transition rates in the above construction, if the state is s at time t, are given by $q_{s,s'}(x(t))$; thus the construction is a valid implementation of the state sequence.

Without loss of generality, we will show (D.1) for a state $s^* \in S$, for $T = 0$.

We start modifying our process by introducing a new state $s^\dagger$. Suppose that at time t_k a state transition occurs from a state $s_{k-1} \neq s^*$, and we have sampled a V_k to determine the state s_k. If in the original process we would have transitioned to s^* (i.e. $V_k < q_{s_{k-1},s^*}(x(t))/q_{s_{k-1}}(x(t))$) then in our modified process we transition to $s_k = s^*$ only if $V_k < \eta \leq q_{s_{k-1},s^*}(x(t))/q_{s_{k-1}}(x(t))$; otherwise we transition to $s^\dagger$. We stay at either of s^* or $s^\dagger$ until t_{k+1} satisfying (D.2) for $s_k = s^*$ then transition to a successor state s_{k+1} determined by using V_{k+1} in the inverse cdf method on $\left(q_{s^*,s}(x(t_{k+1}))/q_{s^*}(x(t_{k+1}))\right)_{s\notin\{s^*,s^\dagger\}}$ (i.e. we use the transition rates for state s^* irrespective of whether we are in s^* or $s^\dagger$). When the original process is in a state other than s^*, the modified process is in the same state; when the original process is in s^*, the modified process is in either

s^* or $s^\dagger$. The modified process therefore spends no more time in s^* than the original process, in any interval $[T, T+\Delta T]$.

Our next modification homogenises the holding times, and amalgamates all states other than s^*. We introduce a new state sequence $\tilde{s}_k$ such that if $V_k < \eta$ and $\tilde{s}_{k-1} \neq s^*$ then $\tilde{s}_k = s^*$; otherwise $\tilde{s}_k = s^-$, where s^- is a new state amalgamating all states other than s^*. This means that if $s_k = s^*$ in the first modification then $\tilde{s}_k = s^*$, whereas if $s_k \neq s^*$ in the first modification then $\tilde{s}_k = s^-$. We also define new holding times, such that the holding time in state $\tilde{s}_k$ is given by $-\log(1-U_k)/\lambda_{\tilde{s}_k}$ with $\lambda_{s^*} = \lambda_{\max}$ and $\lambda_{s^-} = \lambda_{\min}$. This means that the kth holding time when $\tilde{s}_k = s^*$ is bounded above by the kth holding time in the original process, whereas the kth holding time when $\tilde{s}_k = s^-$ is bounded below by the kth holding time in the original process. Once again, the $\tilde{s}_k$ process spends no more time in s^* than the original process, in any interval $[T, T+\Delta T]$.

Finally note that the $\tilde{s}_k$ process has a very simple transition structure: when in state s^*, wait for an $\text{Exp}(\lambda_{\max})$ holding time then transition to s^-; when in state s^- wait for an $\text{Exp}(\lambda_{\min})$ holding time then transition to s^* with probability η, otherwise return to s^- and restart the clock. Simple renewal-reward theory (e.g. Grimmett and Stirzaker, 2001) easily gives that there exists a ΔT such that (D.1) holds for the $\tilde{s}_k$ process.

However note that any U_k, V_k sequence for which the $\tilde{s}_k$ process occupies state s^* for time at least Q in $[T, T+\Delta T]$ also ensures that the same holds for the original state transition process. It follows immediately that (D.1) holds for the original process. □

We make use of this result in the context of a regular embedding of an irreducible stochastic game as follows.

Corollary D.2. *Consider a regular embedding of an irreducible stochastic game* Γ. *For any* $\epsilon > 0$ *and* $k > 0$, *there exists a time* $\Delta T > Q$, *depending on* Q, ϵ *and* k, *such that for any time* $T \geq 1$, *any initial state in* S, *and any measurable strategy process* $x(t)$,

$$P\left(\forall s \in S, \int\limits_{T}^{T+\Delta T} \mathbb{1}_s(t)\,\mathrm{d}t \geq Q\right) \geq 1 - \frac{\epsilon}{k}. \tag{D.3}$$

Proof. The definition of a regular embedding of an irreducible game, given in Section 4, ensures that the rates $q_{s,s'}(x(t))$ meet the conditions of Lemma D.1. Hence there exists $\Delta T > 0$ such that

$$\forall T \geq 1,\ \forall s \in S,\ P\left(\int\limits_{T}^{T+\Delta T} \mathbb{1}_s(t)\,\mathrm{d}t \geq Q\right) \geq 1 - \frac{\epsilon}{|S|k}.$$

Therefore,

$$P\left(\exists s \in S \text{ s.t. } \int\limits_{\tau}^{\tau+\Delta T} \mathbb{1}_s(t) dt < Q\right) \leq \sum_{s \in S} P\left(\int\limits_{\tau}^{\tau+\Delta T} \mathbb{1}_s(t) dt < Q\right) \leq \frac{\epsilon}{k}. \quad \square$$

Hence, in any time interval of length ΔT, the probability that each $x_s(t)$ is updated for at least Q time units is high. Now fix $\epsilon > 0$ for the remainder of the proof, define

$$Q := \frac{64(b_2 - b_1)}{(1-\delta)\epsilon}$$

where b_1 and b_2 are the bounds on rewards defined in (2.9), and let ΔT be appropriate for this choice of ϵ, Q and an as yet unspecified k. Define $A(T)$ to be the event

$$A(T) = \left\{ \forall s \in S, \quad \int_{T}^{T+\Delta T} \mathbb{1}_s(t)\,\mathrm{d}t \geq Q \right\}.$$

Each of the subsequent lemmas will be conditioned on some $A(T)$, and hence (by Corollary D.2) will be true with a controlled probability.

Lemma D.3. *There exists an integer $k_0 > 1$ depending only on b_1, b_2, δ and ϵ such that, for any time $\bar{T} \geq (k_0 - 1)\Delta T$, if $A(\bar{T})$ holds, then*

$$\forall s \in S, \; w_{s,\vec{u}(\bar{T}+\Delta T)}(x_s(\bar{T} + \Delta T)) \leq \frac{(1-\delta)\epsilon}{32}. \tag{D.4}$$

Proof. Firstly, Lemma B.1 implies that $w_{s,\vec{u}(t)}(x_s(t))$ is differentiable for almost all time t. In any state s and at any time t, from (B.1) it follows that

$$\frac{dw_{s,\vec{u}(t)}(x_s(t))}{dt} \leq \dot{\vec{u}} \cdot D_{\vec{u}} w_{s,\vec{u}(t)}(x_s(t)) - w_{s,\vec{u}(t)}(x_s(t))\mathbb{1}_s(t) \tag{D.5}$$

Recall from (2.9) and (4.1)(a) that we can assume $|\dot{u}_s(t)| \leq (b_2 - b_1)/t$ for all s and all t. Hence, as in the proof of Lemma B.2, we can choose k_0 sufficiently large (depending only on b_1, b_2, δ and ϵ) such that $|\dot{\vec{u}} \cdot D_{\vec{u}} w_{s,\vec{u}(t)}(x_s(t))| \leq (1-\delta)\epsilon/(64\Delta T)$ for all $t \geq (k_0 - 1)\Delta T$. Thus, for any $t \geq (k_0 - 1)\Delta T$,

$$\frac{dw_{s,\vec{u}(t)}(x_s(t))}{dt} \leq \frac{(1-\delta)\epsilon}{64\Delta T} - w_{s,\vec{u}(t)}(x_s(t))\mathbb{1}_s(t) \leq \frac{(1-\delta)\epsilon}{64\Delta T}. \tag{D.6}$$

Now let $\bar{T} \geq k_0 \Delta T$. If, for our state s, there exists some time $T' \in [\bar{T}, \bar{T} + \Delta T]$ such that

$$w_{s,\vec{u}(T')}(x_s(T')) \leq \frac{(1-\delta)\epsilon}{64},$$

then it follows from (D.6) that

$$w_{s,\vec{u}(\bar{T}+\Delta T)}(x_s(\bar{T} + \Delta T)) \leq \frac{(1-\delta)\epsilon}{64} + \int_{T'}^{\bar{T}+\Delta T} \frac{(1-\delta)\epsilon}{64\Delta T}\,\mathrm{d}t \leq \frac{(1-\delta)\epsilon}{32}. \tag{D.7}$$

Now suppose that, contrary to the conclusion of the lemma,

$$\exists \tilde{s} \in S \text{ s.t. } w_{\tilde{s},\vec{u}(\bar{T}+\Delta T)}(x_{\tilde{s}}(\bar{T} + \Delta T)) > \frac{(1-\delta)\epsilon}{32}. \tag{D.8}$$

By the previous calculation, it follows that $w_{\tilde{s},\vec{u}(t)}(x_{\tilde{s}}(t)) > \frac{(1-\delta)\epsilon}{64}$ for all $t \in [\bar{T}, \bar{T} + \Delta T]$. Since $w_{\cdot,\cdot}(\cdot) \leq b_2 - b_1$ by (2.9), it follows from (D.6) and the definition of Q that

$$w_{\tilde{s},\vec{u}(\bar{T}+\Delta T)}(x_{\tilde{s}}(\bar{T} + \Delta T)) \leq b_2 - b_1 + \int_{\bar{T}}^{\bar{T}+\Delta T} \frac{(1-\delta)\epsilon}{64\Delta T}\,\mathrm{d}t - \int_{\bar{T}}^{\bar{T}+\Delta T} \mathbb{1}_{\tilde{s}}(t)\frac{(1-\delta)\epsilon}{64}\,\mathrm{d}t$$

$$\leq b_2 - b_1 + \frac{(1-\delta)\epsilon}{64} - \frac{Q(1-\delta)\epsilon}{64}$$
$$= \frac{(1-\delta)\epsilon}{64},$$

contradicting (D.8). □

Comment. As in the proof of the best response dynamic in Appendix B, (D.5) is the only line in the proof of Theorem 4.1 where we use a property of the best-response dynamic (4.1)(b). For the rest of the proof, we only need the formulation of payoff adjustment (4.1)(a) and the stochastic game structure.

Lemma D.4. *Fix an integer $K > k_0$ and suppose that event $A(k\Delta T)$ holds for each integer $k \in \{k_0 - 1, k_0, \dots, K-1\}$. Then*

$$w_{s,\vec{u}(t)}(x_s(t)) < \frac{(1-\delta)\epsilon}{16} \tag{D.9}$$

for all $s \in S$ and all $t \in [k_0\Delta T, K\Delta T]$.

Proof. Fix $s \in S$ and $k \in \{k_0, k_0+1, \dots, K-1\}$. By Lemma D.3,

$$w_{s,\vec{u}(k\Delta T)}(x_s(k\Delta T)) \leq \frac{(1-\delta)\epsilon}{32}.$$

For $k\Delta T < t \leq (k+1)\Delta T$ we therefore have, by (D.6),

$$w_{s,\vec{u}(t)}(x(t)) \leq w_{s,\vec{u}(k\Delta T)}(x_s(k\Delta T)) + \int_{k\Delta T}^{t} \frac{(1-\delta)\epsilon}{64\Delta T}\,\mathrm{d}t$$
$$\leq \frac{(1-\delta)\epsilon}{32} + \frac{(1-\delta)\epsilon}{64}$$
$$< \frac{(1-\delta)\epsilon}{16}. \quad \square$$

This result is the analogue of Lemma B.2, and allows us to bound the difference between $f_{s,\vec{u}(t)}(x_s(t))$ and $v_{s,\vec{u}(t)}$. We will now proceed along similar lines as for Lemma B.4.

Lemma D.5. *Fix $K > k_0$, and suppose that $A(k\Delta T)$ holds for each integer $k \in \{k_0 - 1, k_0, \dots, K-1\}$. There exists a $k_1 > k_0$, depending only on b_1, b_2, δ and ϵ, such that, if $K > k_1$, for all $t \in [k_1\Delta T, K\Delta T]$,*

$$\max_{s\in S} |u_s(t) - v_{s,\vec{u}(t)}| \leq \left(1 + \frac{3(1-\delta)}{16}\right)\epsilon \tag{D.10}$$

and

$$\max_{s\in S} |u_s(t) - f_{s,\vec{u}(t)}(x_s(t))| \leq \left(1 + \frac{1-\delta}{2}\right)\epsilon. \tag{D.11}$$

Proof. Lemma D.4 shows that

$$\forall s \in S,\ \forall t \in [k_0\Delta T, K\Delta T],\ w_{s,\vec{u}(t)}(x_s(t)) \leq \frac{(1-\delta)\epsilon}{16}.$$

Recall the definition of $s_f(\cdot)$ in (B.7) and $s_v(\cdot)$ in (B.9):

$$s_f(t) \in \underset{s\in S}{\operatorname{argmax}} |u_s(t) - f_{s,\vec{u}(t)}(x_s(t))|$$
$$s_v(t) \in \underset{s\in S}{\operatorname{argmax}} |u_s(t) - v_{s,\vec{u}(t)}|.$$

As in Lemma B.4, we start by showing that there exists a $\tau \in [k_0\Delta T, K\Delta T]$ such that

$$|u_{s_v(\tau)}(\tau) - v_{s_v(\tau),\vec{u}(\tau)}| < \left(1 + \frac{3(1-\delta)}{16}\right)\epsilon, \text{ and} \tag{D.12}$$
$$|u_{s_f(\tau)}(\tau) - f_{s_f(\tau),\vec{u}(\tau)}\left(x_{s_f(\tau)}(\tau)\right)| < \epsilon. \tag{D.13}$$

Our basic facts about w (see (2.4)) give that, $\forall s \in S$, $\forall t \in [k_0\Delta T, K\Delta T]$,

$$|f_{s,\vec{u}(t)}(x_s(t)) - v_{s,\vec{u}(t)}| \le w_{s,\vec{u}(t)}(x_s(t)) \le \frac{(1-\delta)\epsilon}{16}.$$

This is precisely condition (B.6). As in the proof of Lemma B.4, (B.21) follows, so that (D.12) is an immediate consequence of (D.13).

Suppose now, for a contradiction, that (D.13) does not hold for any $t \in [k_0\Delta T, k_1\Delta T]$. As in the proofs of Lemmas B.3 and B.4, we may infer that (B.22) also holds:

$$\frac{d|u_{s_v(t)}(t) - v_{s_v(t),\vec{u}(t)}|}{dt} \le -\frac{3(1-\delta)\epsilon}{4t}.$$

By (B.22), and noting that $b_1 \le u_{k_0\Delta T}(k_0\Delta T), v_{s_v(k_0\Delta T),\vec{u}(k_0\Delta T)} \le b_2$, we observe

$$\begin{aligned}
&|u_{s_v(k_1\Delta T)}(k_1\Delta T) - v_{s_v(k_1\Delta T),\vec{u}(k_1\Delta T)}|\\
\le& b_2 - b_1 - \int_{k_0\Delta T}^{k_1\Delta T} \frac{3(1-\delta)\epsilon}{4t} dt\\
=& b_2 - b_1 - \frac{3(1-\delta)\epsilon}{4}\left(\log(k_1\Delta T) - \log(k_0\Delta T)\right)\\
\le& b_2 - b_1 - \frac{3(1-\delta)\epsilon}{4}\log(k_1/k_0).
\end{aligned}$$

Hence for sufficiently large k_1, depending only on b_1, b_2, δ and ϵ, we have that $|u_{s_v(k_1\Delta T)}(k_1\Delta T) - v_{s_v(k_1\Delta T),\vec{u}(k_1\Delta T)}| < (1 - 3(1-\delta)/16)\epsilon$. By (B.21) and (B.6),

$$|u_{s_f(k_1\Delta T)}(k_1\Delta T) - f_{s_f(k_1\Delta T),\vec{u}(k_1\Delta T)}\left(x_{s_f(k_1\Delta T)}(k_1\Delta T)\right)| < \epsilon,$$

contradicting our assumption that (D.13) never holds. Thus there exists a $\tau \in [k_0\Delta T, k_1\Delta T]$ such that (D.12) and (D.13) hold.

Finally, note that (D.10) and (D.11) are identical to (B.19) and (B.20) in Lemma B.4. We have already seen that the conditions of this lemma imply that (B.6) holds for all $t \in [k_0\Delta T, K\Delta T]$. The argument to extend from (D.12) and (D.13) to the conclusion of the lemma is identical to that in the proof of Lemma B.4. □

Proof of Theorem 4.1. Fix $\epsilon > 0$, and hence k_0 and k_1. Let $K = \lceil k_1/\epsilon \rceil$, recall that $Q = \frac{64(b_2-b_1)}{(1-\delta)\epsilon}$, and let ΔT be chosen such that

$$\forall \tau \geq 1,\ P\left(\forall s \in S,\ \int_{\tau}^{\tau+\Delta T} \mathbb{1}_s(t)dt \geq Q\right) \geq 1 - \frac{\epsilon}{K},$$

which is possible by Corollary D.2. It follows that

$$P\left(\forall s \in S,\ \forall k \in \{k_0, \ldots, K-1\},\ \int_{k\Delta T}^{(k+1)\Delta T} \mathbb{1}_s(t)dt \geq Q\right) \geq 1 - \epsilon.$$

From Lemma D.5, noting that $k_1 \leq \epsilon K$, it follows that

$$P\left(\forall t \in [\epsilon K \Delta T, K\Delta T], E_s(t) \text{ occurs}\right) \geq 1 - \epsilon$$

where the event

$$E_s(t) := \begin{cases} |u_{s_v(t)}(t) - v_{s_v(t),\vec{u}(t)}| \leq \left(1 + \frac{3(1-\delta)}{16}\right)\epsilon \\ |u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}\left(x_{s_f(t)}(t)\right)| \leq \left(1 + \frac{1-\delta}{2}\right)\epsilon. \end{cases}$$

Given any $\hat{t} \geq K\Delta T$, we can replace the ΔT by $\frac{\hat{t}}{K}$, and the above result still holds, i.e.,

$$P\left(\forall t \in [\epsilon \hat{t}, \hat{t}],\ E_s(t) \text{ occurs}\right) \geq 1 - \epsilon.$$

The proof concludes in an identical manner to the proof of Theorem 3.1(i). □

Appendix E. Proof of Theorem 5.1

To emphasize that $\delta(t)$ is a variable, we denote the auxiliary game by $G_{s,\vec{u}(t),\delta(t)}$, its value by $v_{s,\vec{u}(t),\delta(t)}$, and its energy by $w_{s,\vec{u}(t),\delta(t)}$, for each state $s \in S$ at each time $t \geq 0$.

Begin by noting that it is immediate from (5.2)(a) that

$$\forall t \geq 2,\ \delta(t) = 1 - \frac{e^c}{\log t} \tag{E.1}$$

where

$$1 - \frac{e^c}{\log 2} = \delta(2). \tag{E.2}$$

Lemma E.1. *In any δ-converging best-response dynamic, for all $\epsilon > 0$, there exists a time $t_1 \geq 2$ such that for all $t \geq t_1$,*

$$\forall s \in S,\ |f_{s,\vec{u}(t),\delta(t)}\left(x_s(t)\right) - v_{s,\vec{u}(t),\delta(t)}| \leq \frac{(1-\delta(t))\epsilon}{16}. \tag{E.3}$$

Proof. We put

$$\zeta := \max\{\max_{s \in S, a \in A_s} |r_s(a)|, \max_{s \in S, a \in A_s} r_s(a) - \min_{s \in S, a' \in A_s} r_s(a')\}. \tag{E.4}$$

Note that

$$f_{s,\vec{u}(t),\delta(t)}(x_s(t)) - u_s(t) \leq \zeta \tag{E.5}$$

always holds, as in the initial condition

$$\forall s \in S, \quad \min_{s' \in S, a' \in A_s} r_{s'}(a') \le u_s(0) \le \max_{s' \in S, a' \in A_s} r_{s'}(a').$$

By a similar proof to that of Lemma B.1, we can show that $w_{s,\vec{u}(t),\delta(t)}(x_s(t))$ is differentiable for almost all time t. Observe the following results of partial derivative of $w_{s,\vec{u}(t),\delta(t)}(x_s(t))$:

(i) by (2.5),

$$\dot{x}_s \cdot D_{x_s} w_{s,\vec{u},\delta} = -w_{s,\vec{u},\delta};$$

(ii) by (2.11) and (2.12),

$$\dot{\vec{u}} \cdot D_{\vec{u}} w_{s,\vec{u},\delta} \le 2\delta \max_{s' \in S} |\dot{u}_{s'}|;$$

(iii) by (2.11) and (2.12),

$$\frac{\partial w_{s,\vec{u},\delta}}{\partial \delta} \frac{d\delta}{dt} \le 2\zeta \frac{d\delta}{dt}.$$

Thus, by (5.2)(a), (5.2)(b), and (E.5),

$$\begin{aligned} \frac{dw_{s,\vec{u}(t),\delta(t)}(x_s(t))}{dt} &\le -w_{s,\vec{u}(t),\delta(t)}(x_s(t)) + 2\delta(t) \max_{s' \in S} |\dot{u}_{s'}| + 2\zeta \frac{d\delta}{dt} \\ &= -w_{s,\vec{u}(t),\delta(t)}(x_s(t)) + \frac{2\delta(t)\zeta}{t} + \frac{2\zeta(1-\delta(t))}{t \ln t}. \end{aligned} \tag{E.6}$$

Claim:

$$\exists t_1 \text{ s.t. } \forall t \ge t_1, \ \forall s \in S, \ w_{s,\vec{u}(t),\delta(t)}(x_s(t)) \le \frac{(1-\delta(t))\epsilon}{16}.$$

To see this, we first infer from (E.1) that

$$\frac{(1-\delta(t))\epsilon}{16} = \frac{\epsilon e^c}{16 \ln t},$$

where c is defined in (E.2). Thus, there exists a time T_1 such that for all $t \ge T_1$,

$$\max\left\{\frac{2\delta(t)\zeta}{t}, \frac{2\zeta(1-\delta(t))}{t \ln t}\right\} < \frac{(1-\delta(t))\epsilon}{64}.$$

At any $t \ge T_1$, if $w_{s,\vec{u}(t),\delta(t)}(x_s(t)) > \frac{(1-\delta(t))\epsilon}{16}$, then it follows from (E.6) that

$$\frac{dw_{s,\vec{u}(t),\delta(t)}(x_s(t))}{dt} < -\frac{w_{s,\vec{u}(t),\delta(t)}(x_s(t))}{2}.$$

We have completed the proof of the claim, and (E.3) follows from the definition of $w_{s,\vec{u}(t),\delta(t)}(x_s(t))$. □

Comment. Similarly to the other dynamical systems we consider, the partial derivative (i) is the only line in the proof of Theorem 5.1 where we use a property of the best-response dynamic (5.2)(c), i.e., an implication of the revision protocol. For the rest of the proof, we only need the formulations (5.2)(a) and (5.2)(b) as well as the auxiliary game structure (2.11).

Lemma E.2. *In any δ-converging best-response dynamic, for all $\epsilon > 0$, there exists $\bar{t}$ such that for all $t > \bar{t}$,*

$$|u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t),\delta(t)}\left(x_{s_f(t)}(t)\right)| < 2\epsilon$$

and

$$|u_{s_v(t)}(t) - v_{s_v(t),\vec{u}(t),\delta(t)}| < 2\epsilon.$$

The notations of $s_f(t)$ and $s_v(t)$ are defined in (B.7) and (B.9), respectively.

Proof. Recall ζ defined in (E.4). We can then take a time $t_2 \geq t_1$ such that

$$\forall t \geq t_2,\ \frac{\zeta}{\ln t} \leq \frac{\epsilon}{8}. \tag{E.7}$$

Suppose that at a time $t \geq t_2$

$$|u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t),\delta(t)}\left(x_{s_f(t)}(t)\right)| \geq \epsilon. \tag{E.8}$$

Then, from (5.2)(b), it follows that at this t,

$$\left|\frac{du_{s_f(t)}(t)}{dt}\right| \geq \frac{\epsilon}{t}. \tag{E.9}$$

For game $G_{s_v(t),\vec{u}(t),\delta(t)}$, it follows from (5.1), Lemma A.2, and (E.7) that at this t

$$\frac{\left|\frac{\partial v_{s_v(t),\vec{u}(t),\delta(t)}}{\partial\delta} \cdot \frac{d\delta}{dt}\right|}{\left|\frac{du_{s_f(t)}(t)}{dt}\right|} \leq \frac{\frac{\zeta(1-\delta(t))}{t\ln t}}{\frac{\epsilon}{t}} = \frac{\zeta(1-\delta(t))}{\epsilon \ln t} \leq \frac{1-\delta(t)}{8}. \tag{E.10}$$

On the other hand, after applying (E.3) to (B.21), we find that

$$\left||u_{s_f(t)}(t) - v_{s_f(t),\vec{u}(t),\delta(t)}| - |u_{s_v(t)}(t) - v_{s_v(t),\vec{u}(t),\delta(t)}|\right| \leq \frac{(1-\delta(t))\epsilon}{8},$$

and thus condition (B.11) holds for state $s_v(t)$. We have the following property by the argument for (B.18):

- if $v_{s_v(t),\vec{u}(t),\delta(t)} > u_{s_v(t)}(t)$, then
$$\dot{\vec{u}}(t) D_{\vec{u}(t)} v_{s_v(t),\vec{u}(t),\delta(t)} \leq \frac{du_{s_v(t)}(t)}{dt} - \frac{3(1-\delta(t))}{4}\left|\frac{du_{s_f(t)}(t)}{dt}\right|;$$
- if $v_{s_v(t),\vec{u}(t),\delta(t)} < u_{s_v(t)}(t)$, then
$$\dot{\vec{u}}(t) D_{\vec{u}(t)} v_{s_v(t),\vec{u}(t),\delta(t)} \geq \frac{du_{s_v(t)}(t)}{dt} + \frac{3(1-\delta(t))}{4}\left|\frac{du_{s_f(t)}(t)}{dt}\right|.$$

Together with (E.10) and (E.9), we have

$$\frac{d|v_{s_v(t),\vec{u}(t),\delta(t)} - u_{s_v(t)}(t)|}{dt} \leq -\frac{1-\delta(t)}{2}\left|\frac{du_{s_f(t)}(t)}{dt}\right| \leq -\frac{\epsilon(1-\delta(t))}{2t}. \tag{E.11}$$

We may further deduce from (E.1) that

$$\frac{d|v_{s_v(t),\vec{u}(t),\delta(t)} - u_{s_v(t)}(t)|}{dt} \le -\frac{\epsilon e^c}{2t\ln t}, \tag{E.12}$$

where c is defined in (E.2).

Thus, by the similar argument to the one after (B.22) in the proof of Lemma B.4, we can deduce that there exists time $\bar{t} \ge t_2$ such that

$$|f_{s_f(t),\vec{u}(t),\delta(t)}\left(x_{s_f(t)}(t)\right) - u_{s_f(t)}(t)| < 2\epsilon$$

and

$$|u_{s_v(t)}(t) - v_{s_v(t),\vec{u}(t),\delta(t)}| < 2\epsilon,$$

for all $t > \bar{t}$. □

Proof of Theorem 5.1. Recall the convergence of $\mathrm{Val}_s(\delta)$ as δ increases to 1, shown in Bewley and Kohlberg (1976). The desired conclusion follows from Lemmata E.1 and E.2. □

References

Aubin, J.P., Cellina, A., 1984. Differential Inclusion. Springer, Berlin.

Balkenborg, D., Kuzmics, C., Hofbauer, J., 2013. Refined best-response correspondence and dynamics. Theor. Econ. 8 (1), 165–192.

Barron, E.N., Goebel, R., Jensen, R.R., 2010. Best response dynamics for continuous games. Proc. Am. Math. Soc. 138 (3), 1069–1083.

Benaïm, M., Hofbauer, J., Sorin, S., 2005. Stochastic approximations and differential inclusions. SIAM J. Control Optim. 2005 (44), 328–348.

Berger, U., 2005. Fictitious play in 2 × n games. J. Econ. Theory 120, 139–154.

Bewley, T., Kohlberg, E., 1976. The asymptotic theory of stochastic games. Math. Oper. Res. 1, 197–208.

Borkar, V., 2002. Reinforcement learning in Markovian evolutionary games. Adv. Complex Syst. 5, 55–72.

Brown, G.W., 1949. Some Notes on Computation of Games Solutions. Report P-78. The Rand Corporation.

Dutta, P.K., 1995. A folk theorem for stochastic games. J. Econ. Theory 66, 1–32.

Ely, J.C., Yilankaya, O., 2001. Nash equilibrium and the evolution of preferences. J. Econ. Theory 97, 255–272.

Fox, M.J., Shamma, J.S., 2013. Population games, stable games, and passivity. Games 4, 561–583.

Fudenberg, D., Levine, D.K., 1998. The Theory of Learning in Games. MIT Press.

Fudenberg, D., Levine, D.K., 2014. Recency, consistent learning, and Nash equilibrium. Proc. Natl. Acad. Sci. 111 (Supplement 3), 10826–10829.

Guo, X., Hernández-Lerma, O., 2005. Zero-sum continuous-time Markov games with unbounded transition and discounted payoff rates. Bernoulli 11 (6), 1009–1029.

Grimmett, G.R., Stirzaker, D.R., 2001. Probability and Random Processes. Oxford University Press.

Harris, C., 1998. On the rate of convergence of continuous-time fictitious play. Games Econ. Behav. 22, 238–259.

Hofbauer, J., 1995. Stability for the Best Response Dynamics. University of Vienna. Mimeo.

Hofbauer, J., Sandholm, W., 2009. Stable games and their dynamics. J. Econ. Theory 144, 1665–1693.

Hofbauer, J., Sigmund, K., 1998. Evolutionary Games and Population Dynamics. Cambridge University Press.

Hofbauer, J., Sorin, S., 2006. Best response dynamics for continuous zero-sum games. Discrete Contin. Dyn. Syst., Ser. B 6 (1), 215–224.

Hopkins, E., 1999. A note on best response dynamics. Games Econ. Behav. 29, 138–150.

Hörner, J., Sugaya, T., Takahashi, S., Vieille, N., 2011. Recursive methods in discounted stochastic games: an algorithm for $\delta \to 1$ and a Folk theorem. Econometrica 79 (4), 1277–1318.

Gilboa, Y., Matsui, A., 1991. Social stability and equilibrium. Econometrica 59, 859–867.

Leslie, D.S., Collins, E.J., 2006. Generalised weakened fictitious play. Games Econ. Behav. 56, 285–298.

Levy, Y., 2013. Continuous-time stochastic games of fixed duration. Dyn. Games Appl. 3, 279–312.

Matsui, A., 1989. Social Stability and Equilibrium. CMS-DMS No. 819. Northwestern University.

Mertens, J.-F., Neyman, A., 1981. Stochastic games. Int. J. Game Theory 10, 53–66.

Neyman, A., 2017. Continuous-time stochastic games. Games Econ. Behav. 104, 92–130.

Oliu-Barton, M., 2014. The asymptotic value in finite stochastic games. Math. Oper. Res. 39 (3), 712–721.

Perkins, S., 2013. Advanced Stochastic Approximation Frameworks and their Applications. PhD thesis. University of Bristol.

Sandholm, W.H., 2001. Preference evolution, two-speed dynamics, and rapid social change. Rev. Econ. Dyn. 4, 637–679.

Sandholm, W.H., 2010. Population Games and Evolutionary Dynamics. MIT Press.

Schoenmakers, G., Flesch, J., Thuijsman, F., 2007. Fictitious play in stochastic games. Math. Methods Oper. Res. 66, 315–325.

Shapley, L., 1953. Stochastic games. Proc. Natl. Acad. Sci. USA 39, 1095–1100.

Solan, E., 2009. Stochastic games. In: Encyclopedia of Database Systems. Springer.

Szepesvári, C., Littman, M., 1999. A unified analysis of value-function-based reinforcement-learning algorithms. Neural Comput. 11, 2017–2060.

Vigeral, G., 2010. Evolution equations in discrete and continuous time for nonexpansive operators in Banach spaces. ESAIM Control Optim. Calc. Var. 16, 809–832.

Viossat, Y., Zapechelnyuk, A., 2013. No-regret dynamics and fictitious play. J. Econ. Theory 148, 825–842.

Vrieze, O., Tijs, S., 1982. Fictitious play applied to sequences of games and discounted stochastic games. Int. J. Game Theory 11, 71–85.

Xu, Z., 2016. Convergence of best-response dynamics in extensive-form games. J. Econ. Theory 162, 21–54.

Zusai, D., 2019. Gains in evolutionary dynamics: a unifying and intuitive approach to linking static and dynamic stability. arXiv:1805.04898v5.