A Convergence of smooth best-response dynamics in identical-interest and zero-sum stochastic games

In this appendix, we provide detailed proofs of results of Section 4. Since solutions of differential inclusion SBRD are special solutions of differential inclusion MBRD, it is sufficient to study MBRD. In what follows, unless otherwise specified, $t \mapsto (x_s, u_s^i, \hat{P}_s, \hat{r}_s^i)_{s,i}$ is a solution of MBRD.

We start with general properties of the smooth best response, convergence of estimates and regularity of solutions and then proceed with the study of two special cases: identical-interest stochastic games (subsection A.4) and zero-sum stochastic games (subsection A.5).

A.1 Properties of the smooth best-response

Lemma A.1. *Function $(u, x_s) \mapsto \mathrm{sbr}_{s,u}^i(x_s)$ is continuous in u and x_s.*

Proof. It follows from a simple application of the maximum theorem that the smooth best-response is upper hemicontinuous as a set-valued map. The strict concavity of h^i implies that there is a single profile that maximizes the smooth auxiliary payoff. Therefore h^i is continuous as a single-valued application. □

Lemma A.2. *For $B > 0$, there exists $\eta > 0$ such that for all $x \in \Pi_{j \in I} \Delta\left(A^j\right)$, $u \in \mathbb{R}^S$ such that $\|u\|_\infty \leq B$ and $a_s^i \in A^i$:*

$$\mathrm{sbr}_{s,u}^i(x_s)(a_s^i) \geq \eta$$

Proof. This is a classical property of the smooth best-response under assumptions (H1). Point $\mathrm{sbr}_{s,u}^i(x_s)$ is a maximum of the function $y_s^i \mapsto f_{s,u}^i\left(y_s^i, x_s^{-i}\right) + \epsilon h^i\left(y_s^i, x_s^{-i}\right)$. However, let x_s be an interior point of the simplex and (y_s^i, x_s^{-i}) be on the boundary for player i. Then, the composition of the linear interpolation between (y_s^i, x_s^{-i}) and x_s and $x_s \mapsto f_{s,u}^i(x_s) + \epsilon h^i(x_s)$ is concave because both functions are. However, the slope of this composed function is infinite (because the norm of the gradient goes to ∞ and x_s is interior), therefore it goes to $-\infty$ (otherwise it cannot be concave), which implies that (y_s^i, x_s^{-i}) cannot be a maximum. Then, by compacity, smooth best-response are away from the boundary of the simplex. □

A.2 Convergence of estimates $\hat{P}$ and $\hat{r}$

Lemma A.3. *There exists $C > 0$ such that for all states s action profiles $b_s \in A$ and $t \geq 0$,*

$$|\hat{r}_s^i(b_s)(t) - r_s^i(b_s)| \leq C\exp(-\eta\alpha_- t)$$
$$|\hat{P}_{ss'}(b_s)(t) - P_{ss'}(b_s)| \leq C\exp(-\eta\alpha_- t)$$

Proof. Let $s \in S$ and $b_s \in A$.
We write $r(t) = \left|\hat{r}_s^i(b_s)(t) - r_s^i(b_s)\right|$, so as $\dot{r} = -\alpha_s(t) a(t)(b_s) r(t)$.

Then we can deduce from Lemma A.2 and the fact that $\alpha_s(t) \geq \alpha_-$ that $\dot{r} \leq -\eta\alpha_- r(t)$.

Therefore, Grönwall Lemma implies that $r(t) \leq r(0)\exp(-\eta\alpha_- t)$. The same proof is valid for $\hat{P}_s(t)$. □

A.3 Regularity of solutions

Lemma A.4. *Functions u_s^i, x_s, $\hat{f}_{s,u^i}^i$ are bounded.*

Proof. Because of the definition of the derivative of x_s, it stays in the simplex and as such it is bounded.

It is clear that $\hat{f}_{s,u^i(t)}^i(x_s(t)) \leq (1-\delta)\|r_s^i\|_\infty + \delta\|u_i(t)\|_\infty$. Therefore, as long as $\|u^i(t)\|_\infty$ is lower than $\|r_s^i\|_\infty$, then $\hat{f}_{s,u^i(t)}^i(x_s(t)) \leq \|r_s^i\|_\infty$. By definition of the derivative of u_s^i, this implies that u_s^i is always smaller than $\|r_s^i\|_\infty$ assuming it is true for the initial value. □

Lemma A.5. *Functions u_s^i, x_s, $\hat{f}^i_{s,u^i}$ are Lipsichitz.*

Proof. All functions are differentiable almost everywhere. The derivatives are bounded by composition since $\beta(t)$ and $\alpha_s(t)$ are bounded and because of Lemma A.4. □

A.4 Convergence in identical-interest games

In identical-interest games, all payoff functions are equal. Therefore, assuming the same initial conditions for all players i for u_s^i, we have $u_s^i(t) = u_s^j(t)$ for all players i and j. So, we can omit the superscript, and the same is true for the payoff functions and the auxiliary payoff functions.

The proof proceeds as follows: we show that the differential of u_s^i becomes a good approximation of the target auxiliary payoff Γ_s, and furthermore that their difference is lower bounded by something integrable. Then we can deduce that there is a limit, for this difference and that it is necessarily 0 and the convergence of actions follows.

A.4.1 Convergence of payoffs

We define $\Gamma_s(t) := \hat{f}_{s,u(t)}(x_s(t)) + \epsilon h(x_s(t))$ and $s_-(t) \in \arg\min_{s\in S} \Gamma_s(t) - u_s(t)$. For a given solution of MBRD, there may be several possible choices of $s_-(t)$, results below are valid for any such choice.

Lemma A.6. *The differential of $\Gamma_{s_-} - u_{s_-}$ is lower bounded for almost every t:*

$$\frac{d\Gamma_{s_-} - u_{s_-}}{dt} \geq -2C\exp(-\eta\alpha_- t) + \beta(t)(\delta - 1)(\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t))$$

where C is defined in Lemma A.3 and η in Lemma A.2.

Proof. First, notice that since $\Gamma_{s_-} - u_{s_-}$ is a minimum of continuous, differentiable almost everywhere functions, it is continuous and differentiable almost everywhere. Let $t \in \mathbb{R}^+, \tau > 0$.

$$\begin{aligned}
&\Gamma_{s_-(t+\tau)}(t+\tau) - u_{s_-(t+\tau)}(t+\tau) - \Gamma_{s_-(t)}(t) + u_{s_-(t)}(t) \\
&\quad = \Gamma_{s_-(t+\tau)}(t) + \tau\frac{d\Gamma_{s_-(t+\tau)}}{dt}(t) - u_{s_-(t+\tau)}(t) \\
&\quad -\tau\frac{du_{s_-(t+\tau)}}{dt}(t) + o(\tau) - \Gamma_{s_-(t)}(t) + u_{s_-(t)}(t)
\end{aligned} \tag{5}$$

Then, for any τ:

$$\begin{aligned}
\frac{du_{s_-(t+\tau)}}{dt}(t) &= \beta(t)\left(\Gamma_{s_-(t+\tau)}(t) - u_{s_-(t+\tau)}(t)\right) \\
&= \beta(t)\left(\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)\right) + o(1)
\end{aligned} \tag{6}$$

Moreover, for any τ:

$$\Gamma_{s_-(t+\tau)}(t) - u_{s_-(t+\tau)}(t) \geq \Gamma_{s_-(t)}(t) - u_{s_-(t)}(t) \tag{7}$$

Now, we need to lower bound $\frac{d\Gamma_{s_-(t+\tau)}}{dt}(t)$. We use the chain rule and Lemma A.3 to get (with $s = s_-(t+\tau)$ to ease reading–since this differentation does not depend on the state):

$$\begin{aligned}
\frac{d\Gamma_s}{dt}(t) \geq \underbrace{-2C\exp(-\eta\alpha_- t)}_{\text{changes in } \hat{r} \text{ and } \hat{P}} + \underbrace{\sum_{j\in I}\alpha_s(t)\langle \mathrm{sbr}^j_{s,u_s(t)}(x_s(j)) - x_s^j(t), \nabla_j \hat{f}_{s,u(t)}(x_s(t)) + \epsilon h(x_s(t))\rangle}_{\text{changes in } x_s,\ \nabla_j \text{ is the gradient with respect to } x_s^j} \\
\underbrace{+\delta\sum_{s'\in S}\hat{P}_{ss'}(x_s(t))\,\dot{u}_{s'}(t)}_{\text{changes in } u_s}
\end{aligned} \tag{8}$$

The second term is positive because $\hat{f}_{s,u} + \epsilon h\left(x_s(t)\right)$ is concave: $\mathrm{sbr}^j_{s,u_s(t)}\left(x_s(j)\right)$ is the maximum of the fonction, therefore its gradient is null and the opposite of the gradient of a concave function is monotone.
The last one is greater than $\delta\beta(t)\left(\Gamma_{s_-(t)}\left(t\right) - u_{s_-(t)}(t)\right)$, leading to:

$$\frac{d\Gamma_{s_-(t+\tau)}}{dt}(t) \geq -2C\exp(-\eta\alpha_- t) + \delta\beta(t)\left(\Gamma_{s_-(t)}\left(t\right) - u_{s_-(t)}(t)\right) \tag{9}$$

Using (6), (7) and (9) in (5) leads to:

$$\begin{aligned}&\Gamma_{s_-(t+\tau)}\left(t+\tau\right) - u_{s_-(t+\tau)}(t+\tau) - \Gamma_{s_-(t)}\left(t\right) + u_{s_-(t)}(t)\\ &\geq \tau\left(-2C\exp(-\eta\alpha_- t) + \beta(t)(\delta-1)\left(\Gamma_{s_-(t)}\left(t\right) - u_{s_-(t)}(t)\right)\right) + o(\tau)\end{aligned}$$

And the result follows.

□

Lemma A.7. *There exists $u_{\infty,s} \in \mathbb{R}^S$ such that for all $s \in S$, $u_s(t) \to u_{\infty,s}$ and $\Gamma_s\left(t\right) \to u_{\infty,s}$.*

Proof. We consider the following differential equation:

$$y' = -2C\exp(-\eta\alpha_- t) + \beta(t)(\delta-1)y \tag{10}$$

Lemma A.6 implies that solutions of (10) with initial condition $y(0) = \Gamma_{s_-(0)}\left(0\right) - u_{s_-(0)}(0)$ lower bound $\Gamma_{s_-} - u_{s_-}$.

Moreover, solutions of (10) are of the form:

$$\left(y(0) - \int_0^t 2C\exp(-\eta\alpha_- v)\exp\left(\int_0^v \beta(u)(1-\delta)du\right)dv\right)\exp\left(-\int_0^t \beta(u)(1-\delta)du\right)$$

which is equal to:

$$y(0)\exp\left(-\int_0^t \beta(u)(1-\delta)du\right) - \int_0^t 2C\exp(-\eta\alpha_- v)\exp\left(-\int_v^t \beta(u)(1-\delta)du\right)dv$$

Which is greater than:

$$-|y(0)|\exp\left(-\int_0^t \beta(u)(1-\delta)du\right) - \int_0^t 2C\exp(-\eta\alpha_- v)\exp\left(-\int_v^t \beta(u)(1-\delta)du\right)dv \tag{11}$$

So for every s, $\Gamma_s\left(t\right) - u_s(t)$ is greater than (11). We study the differential of $u_s(t)$, that is $\beta(t)(\Gamma_s\left(t\right) - u_s(t))$ and show that it is lower bounded by an integrable quantity:

$$\begin{aligned}\dot{u}_s \geq -\beta(t)|y(0)|\exp\left(-\int_0^t \beta(u)(1-\delta)du\right)&\\ - \beta(t)\int_0^t 2C\exp(-\eta\alpha_- v)\exp\left(-\int_v^t \beta(u)(1-\delta)du\right)dv&\end{aligned} \tag{12}$$

The integral of the first term is:

$$\begin{aligned}\int_0^t -\beta(v)|y(0)|\exp\left(-\int_0^v \beta(u)(1-\delta)du\right)&\\ = -|y(0)|\frac{1}{1-\delta}\left(1-\exp\left(-\int_0^t \beta(u)(1-\delta)du\right)\right) > -\infty&\end{aligned}$$

And the integral of the second term is:

$$\begin{aligned}
&\int_0^t -\beta(v) \int_0^v 2C \exp(-\eta\alpha_- w) \exp\left(-\int_w^v \beta(u)(1-\delta)du\right) dwdv \\
&= \int_0^t \int_0^t -1_{w\leq v}\beta(v) 2C \exp(-\eta\alpha_- w) \exp\left(-\int_w^v \beta(u)(1-\delta)du\right) dwdv \\
&= \int_0^t \int_w^t -\beta(v) 2C \exp(-\eta\alpha_- w) \exp\left(-\int_w^v \beta(u)(1-\delta)du\right) dvdw \\
&= \int_0^t 2C \exp(-\eta\alpha_- w) \int_w^t -\beta(v) \exp\left(-\int_w^v \beta(u)(1-\delta)du\right) dvdw \\
&= \int_0^t 2C \exp(-\eta\alpha_- w) \left[\frac{1}{1-\delta} \exp\left(-\int_w^v \beta(u)(1-\delta)du\right)\right]_w^t dw \\
&= \int_0^t 2C \exp(-\eta\alpha_- w) \frac{1}{1-\delta} \left(\exp\left(-\int_w^t \beta(u)(1-\delta)du\right) - 1\right) dw \\
&\geq -\frac{1}{1-\delta}\frac{2C}{\eta\alpha_-}(1-\exp(-\eta\alpha_- t)) > -\infty
\end{aligned}$$

Therefore, both term of the integral of (12) are greater than $-\infty$. Since $u_s(t)$ is bounded (Lemma A.4) and its derivative is $\beta(t)(\Gamma_s(t) - u_s(t))$, then $u_s(t)$ converges to its $\limsup$. Moreover, the same reasoning can be made with Γ_s (see its differentiation in (8)), so it has a limit which is necessarily the same as u_s: otherwise, the derivative of $u_s(t)$ would converge towards $\beta(t)$ times the difference of the limits and $u_s(t)$ would be unbounded because of hypothesis H2. □

A.4.2 Convergence of actions

Lemma A.8. *Action profile $x_s(t)$ converges to a fixpoint of $\mathrm{sbr}_{s,u_{\infty,s}}$. Therefore, $x(t)$ converges to a regularized Nash equilibria of the stochastic game.*

Proof. We use equation (8) to bound above the scalar product:

$$\langle \mathrm{sbr}^j_{s,u_s(t)}(x_s(j)) - x^j_s(t), \nabla_j \hat{f}_{s,u(t)}(x_s(t)) + \epsilon h(x_s(t)) \rangle \tag{13}$$

Since Γ_s and u_s have limits and are Lipsichitz (Lemma A.5), then their derivative goes to 0, and the first term of (8) also goes to 0. As a consequence, the limsup of (13) is bounded above by 0, and this is positive, so it goes to 0. Therefore, the action profile converges to a fixpoint of $\mathrm{sbr}_{s,u_{\infty,s}}$. □

A.5 Convergence in zero-sum games

In this subsection we suppose that the game is zero-sum, that is there are only two players and for all states, $r^1_s(a_s) = -r^2_s(a_s)$ for all action profiles a_s. Therefore, we can omit the superscript with the convention that $r_s = r^1_s$.

The proof is inspired from Leslie et al. [2020], later extended in Baudin and Laraki [2022].

Moreover, we suppose that both players use the same regularizer:

$$h\left(x^1_s, x^2_s\right) = h^1\left(x^1_s, x^2_s\right) = -h^2\left(x^1_s, x^2_s\right) \tag{14}$$

where both h^1 and h^2 are concave in respectively x^1_s and x^2_s.

Therefore, with the same initial conditions, continuation payoffs u_s are opposite for both players, and we also omit the superscript.

A.5.1 Convergence of payoffs

We define the energy of the system, also known as the duality gap (Benaïm et al. [2005]):

$$\begin{aligned}
w_s(t) = &\max_{y^1_s \in \Delta(A^1)} \hat{f}_{s,u(t)}\left(y^1_s, x^2_s(t)\right) + \epsilon h\left(y^1_s, x^2_s(t)\right) \\
&- \min_{y^2_s \in \Delta(A^2)} \hat{f}_{s,u(t)}\left(x^1_s(t), y^2_s\right) + \epsilon h\left(x^1_s(t), y^2_s\right)
\end{aligned} \tag{15}$$

This is a positive quantity because the first (second) term is greater (lower) than $\hat{f}_{s,u(t)}\left(x_s(t)\right) + \epsilon h\left(x_s(t)\right)$. We are going to show that it is mostly decreasing.

In the rest of the section, we use a $\beta^\star$ such that:

$$\beta^\star \geq \limsup \beta(t) \tag{H3}$$

A special case is when $\beta(t)$ goes to 0, in this case $\beta^\star$ can be taken arbitrarily small.

Lemma A.9. *The differential of w_s is bounded:*

$$\frac{dw_s}{dt} \leq -\alpha_- w_s(t) + D\beta^\star + D\exp\left(-\eta\alpha_- t\right)$$

Proof. We write $y_s^{i\star}(t) = \mathrm{sbr}^i_{s,u_s(t)}\left(x_s(t)\right)$ and:

$$g(y_s^{1\star}(t)) := \hat{f}_{s,u(t)}\left(y_s^{1\star}(t), x_s^2(t)\right) + \epsilon h\left(y_s^{1\star}(t), x_s^2(t)\right)$$

where the dependency on $u_s, \hat{P}_s, \hat{r}_s, x_s$ is left implicit. This implies:

$$g(y_s^{1\star}(t)) = \max_{y_s^1 \in \Delta(A^1)} \hat{f}_{s,u(t)}\left(y_s^1, x_s^2(t)\right) + \epsilon h\left(y_s^1, x_s^2(t)\right)$$

Therefore, using the envelope theorem, the derivative of g is written using only derivatives on all variables but $y_s^{1\star}(t)$:

$$\frac{dg \circ y_s^{1\star}}{dt} = D_u g \cdot \dot{u} + D_{x_s^2} g \cdot \dot{x}_s^2 + D_{\hat{r}_s} g \cdot \dot{\hat{r}}_s + D_{\hat{P}_s} g \cdot \dot{\hat{P}}_s$$

With Lemma A.3, Lemma A.4 and Hypothesis H3,

$$\frac{dg \circ y_s^{1\star}}{dt} \leq \beta^\star \|D_u g\| \|r_s\| + D_{x_s^2} g \cdot \dot{x}_s^2 + (\|D_{\hat{r}_s} g\| + \|D_{\hat{P}_s} g\|)\alpha_s(t)\eta C \exp(-\eta\alpha_- t) \tag{16}$$

$\hat{f}_{s,u(t)}$ is linear, so:

$$D_{x_s^2} g \cdot \dot{x}_s^2 = \hat{f}_{s,u(t)}\left(y_s^{1\star}(t), \dot{x}_s^2\right) + \epsilon \nabla_{x_s^2} h\left(y_s^{1\star}(t), x_s^2(t)\right) \cdot \dot{x}_s^2 \tag{17}$$

And since $h = h^1 = -h^2$ (with (14)), $\nabla_{x_s^2} h = -\nabla_{x_s^2} h^2$ by definition, and h^2 is concave in x_s^2, so:

$$\nabla_{x_s^2} h^2\left(y_s^{1\star}(t), x_s^2(t)\right) \cdot \left(y_s^{2\star}(t) - x_s^2(t)\right) \geq h^2\left(y_s^{1\star}(t), y_s^{2\star}(t)\right) - h^2\left(y_s^{1\star}(t), x_s^2(t)\right) \tag{18}$$

It follows from (17), (18), (14) and $\dot{x}_s^2 = \alpha_s(t)(y_s^{2\star}(t) - x_s^2(t))$ that

$$\begin{aligned} D_{x_s^2} g \cdot \dot{x}_s^2 &\leq \alpha_s(t)\hat{f}_{s,u(t)}\left(y_s^{1\star}(t), y_s^{2\star}(t) - x_s^2(t)\right) + \epsilon\alpha_s(t) h\left(y_s^{1\star}(t), y_s^{2\star}(t)\right) - \alpha_s(t)\epsilon h\left(y_s^{1\star}(t), x_s^2(t)\right) \\ &\leq -\alpha_s(t) g(y_s^{1\star}(t)) + \alpha_s(t)\hat{f}_{s,u(t)}\left(y_s^{1\star}(t), y_s^{2\star}(t)\right) + \epsilon\alpha_s(t) h\left(y_s^{1\star}(t), y_s^{2\star}(t)\right) \end{aligned} \tag{19}$$

Since g as a function of $u_s, \hat{P}_s, \hat{r}_s, x_s$ is Lipschitz, and using (16) and (19), there exists D such that:

$$\begin{aligned} \frac{dg \circ y_s^{1\star}}{dt} \leq \frac{D}{2}\beta^\star + \frac{D}{2}\alpha_s(t)\exp(-\eta\alpha_- t) - \alpha_s(t) g(y_s^{1\star}(t)) \\ + \alpha_s(t)\hat{f}_{s,u(t)}\left(y_s^{1\star}(t), y_s^{2\star}(t)\right) + \epsilon\alpha_s(t) h\left(y_s^{1\star}(t), y_s^{2\star}(t)\right) \end{aligned} \tag{20}$$

The same reasoning apply for the second term of w_s with the opposite payoff function (notice that in this case, the second line of (20) is exactly the opposite, therefore when summed it cancels out), leading to, when summed:

$$\begin{aligned} \frac{dw_s}{dt} &\leq D\beta^\star + D\alpha_s(t)\exp(-\eta\alpha_- t) - \alpha_s(t) w_s(t) \\ &\leq D\beta^\star + D\exp(-\eta\alpha_- t) - \alpha_- w_s(t) \end{aligned}$$

because $\alpha_- \leq \alpha_s(t) \leq 1$ and $w_s \geq 0$ (its first term is greater than $\hat{f}_{s,u(t)}\left(x_s(t)\right) + \epsilon h\left(x_s(t)\right)$ and its second one is lower). □

The following lemma implies that the auxiliary payoff is close to the value of the auxiliary game because the duality gap of the auxiliary game is small enough.

Lemma A.10. *For all states $s \in S$,*

$$\limsup w_s(t) \leq 2D\beta^\star\alpha_-^{-1}$$

Furthermore, $\max\{w_s - 2D, 0\}$ *is a Lyapunov function of SBRD (i.e., when payoff and transitions estimate are exact) in the autonomous case.*

Proof. Since w_s is positive (the first term is greater than $\hat{f}_{s,u(t)}\left(x_s(t)\right) + \epsilon h\left(x_s(t)\right)$ and the second one is lower), Lemma A.9 makes it possible to use Grönwall's Lemma on $w_s - 2D\beta^\star\alpha_-^{-1}$. as soon as $\exp(-\eta\alpha_- t) \leq \beta^\star$.

□

Define ξ such as $\frac{(1-\delta)\xi}{16} = 4D\beta^\star\alpha_-^{-1}$.

Estimates of transitions and payoffs are close to real values for t large enough, so Lemma A.10 implies that there exists $t_1(\xi)$ such that for $t \geq t_1(\xi)$:

$$\begin{aligned} |\hat{f}_{s,u} - f_{s,u}| &\leq 4D\beta^\star\alpha_-^{-1} = \frac{(1-\delta)\xi}{16} \\ |\hat{f}_{s,u} + \epsilon h - v_{s,u(t)}| &\leq 2D\beta^\star\alpha_-^{-1} = \frac{(1-\delta)\xi}{32} \\ |f_{s,u} + \epsilon h - v_{s,u(t)}| &\leq 2D\beta^\star\alpha_-^{-1} = \frac{(1-\delta)\xi}{32} \end{aligned} \tag{A1}$$

where $v_{s,u(t)}$ is the value of the auxiliary game parameterized by $u(t)$ (and functions h, $f_{s,u}\left(x_s\right)$ are $\hat{f}_{s,u}\left(x_s\right)$ are valued at $x_s(t)$, omitted for readability).

We define two distinguished states (notice that we use $f_{s,u}$ and not $\hat{f}_{s,u}$):

- $s_f(t) \in \arg\max_{s\in S} |f_{s,u(t)}\left(x_s(t)\right) + \epsilon h\left(x_s(t)\right) - u_s(t)|$
- $s_v(t) \in \arg\max_{s\in S} |v_{s,u(t)} - u_s(t)|$

Lemma A.11. *If (A1) is satisfied (for instance if $t \geq t_1(\xi)$) and*

$$|u_{s_f}(t) - f_{s_f(t),u(t)}\left(x_{s_f(t)}(t)\right) - \epsilon h\left(x_{s_f(t)}(t)\right)| \geq \xi$$

and for an $s \in S$,

$$\left||u_{s_f}(t) - v_{s_f,u(t)}| - |u_s(s) - v_{s,u(t)}|\right| \leq \frac{(1-\delta)\xi}{8}$$

then:

$$\frac{d|u_s(s) - v_{s,u(t)}|}{dt} \leq -\frac{(1-\delta)\beta(t)\xi}{2}$$

Proof. First, using Lemma A.2 of Leslie et al. [2020] on the regularity of the value of a zero-sum (static) game, it follows:

$$\begin{aligned} \left|\frac{dv_{s,u(t)}}{dt}\right| &\leq \delta \max_{s\in S} |\dot{u}_s| \\ &= \delta\beta(t)|f_{s_f(t),u(t)}\left(x_{s_f(t)}(t)\right) + \epsilon h\left(x_{s_f(t)}(t)\right) - u_{s_f}(t)| + \delta\beta(t)\frac{(1-\delta)\xi}{16} \\ &\leq \delta\beta(t)\xi(1 + \frac{1-\delta}{16}) \end{aligned} \tag{21}$$

We now prove that $u_s(t)$ moves towards $v_{s,u(t)}$ at a constant speed relatively to $\beta(t)$:

- If $u_s(t) \geq v_{s,u(t)}$, then $|u_{s_f}(t) - v_{s_f,u(t)}| - u_s(t) + v_{s,u(t)} \leq \frac{(1-\delta)\xi}{8}$.

$$\begin{aligned}
\dot{u}_s &= \beta(t)\left(\hat{f}_{s,u(t)}\left(x_s(t)\right) + \epsilon h\left(x_s(t)\right) - u_s(t)\right) \\
&\leq \beta(t)\left(f_{s,u(t)}\left(x_s(t)\right) + \epsilon h\left(x_s(t)\right) - u_s(t)\right) + \beta(t)\frac{(1-\delta)\xi}{16} \\
&\leq \beta(t)\left(f_{s,u(t)}\left(x_s(t)\right) + \epsilon h\left(x_s(t)\right) + \frac{(1-\delta)\xi}{8} - v_{s,u(t)} - |u_{s_f}(t) - v_{s_f,u(t)}|\right) \\
&\quad + \beta(t)\frac{(1-\delta)\xi}{16} \\
&\leq \beta(t)\left(\frac{4(1-\delta)\xi}{16} - |u_{s_f}(t) - v_{s_f,u(t)}|\right) \\
&\leq \beta(t)\left(\frac{(1-\delta)\xi}{4} - |u_{s_f}(t) - f_{s_f(t),u(t)}\left(x_{s_f(t)}(t)\right) - \epsilon h\left(x_{s_f(t)}(t)\right)|\right) \\
&\leq \beta(t)\left(\frac{(1-\delta)\xi}{4} - \xi\right)
\end{aligned}$$

Summing with $v_{s,u(t)}$ and using (21):

$$\begin{aligned}
\frac{du_s(t) - v_{s,u(t)}}{dt} &\leq \beta(t)\left(\frac{(1-\delta)\xi}{4} - \xi + \delta\xi + \delta\xi\frac{1-\delta}{16}\right) \\
&\leq \beta(t)\xi\left(\frac{1-\delta}{2} - 1 + \delta\right) \\
&\leq -\beta(t)\left(\frac{2(1-\delta)\xi}{4}\right)
\end{aligned}$$

- If $u_s(t) \leq v_{s,u(t)}$, similar calculations yield the same result.

□

Lemma A.12. *For all $s \in S$, $\limsup_{t\to\infty} |u_s(t) - f_{s,u(t)}\left(x_s(t)\right) - \epsilon h\left(x_s(t)\right)| \leq 4\xi$.*

Proof. We define $g(t) = \max\{|u_{s_f}(t) - v_{s_f,u(t)}|, 3\xi\}$.

Now, if $|u_{s_f}(t) - v_{s_f,u(t)}| \leq 2\xi$, then $\frac{dg}{dt} = 0$. If $|u_{s_f}(t) - v_{s_f,u(t)}| \geq 2\xi$ and if t is greater than $t^1(\xi)$, then $|u_{s_f}(t) - f_{s_f(t),u(t)}\left(x_{s_f(t)}(t)\right) - \epsilon h\left(x_{s_f(t)}(t)\right)| \geq \xi$: indeed, $|f_{s_f(t),u(t)}\left(x_{s_f(t)}(t)\right) + \epsilon h\left(x_{s_f(t)}(t)\right) - v_{s_f,u(t)}| \leq \xi$ because of Lemma A.10 and its corollary A1. Similarly, on a neighbourhood of t, every s that maximizes $|f_{s,u(t)}\left(x_s(t)\right) + \epsilon h\left(x_s(t)\right) - u_s(t)|$ satisfies the condition of Lemma A.11, because $|f_{s,u(t)}\left(x_s(t)\right) + \epsilon h\left(x_s(t)\right) - v_{s,u(t)}| \leq \xi$ according to the same Lemma A.10. Therefore, Lemma A.11 can be used and:

$$\frac{dg}{dt} \leq -\frac{3(1-\delta)\beta(t)\xi}{4}$$

This holds as soon as $g(t) > 2\xi$ and $t > t^1(\xi)$. The integral of β is infinite (hypothesis H2), so there is a $t^2(\xi)$ such that for $t \geq t^2(\xi)$, $g(t) = 2\xi$.

Then, using A1, we have $|u_{s_f}(t) - f_{s_f(t),u(t)}\left(x_{s_f(t)}(t)\right) - \epsilon h\left(x_{s_f(t)}(t)\right)| \leq 3\xi$ and by definition of s_f, the inequality of the lemma.

□

A.5.2 Convergence of actions

Lemma A.13. *For all $s \in S$, $x_s(t)$ converge to the set of 3ξ-Regularized Nash equilibria of the auxiliary game.*

Proof. The previous proof gives that $f_{s,u(t)}\left(x_s(t)\right)$ is 3ξ close to $v_{s,u(t)}$, hence the result. □

B Convergence of smooth fictitious-play in identical-interest and zero-sum stochastic games

In order to prove the convergence of the discrete-time procedures described in this paper, we use the theory of stochastic approximations to relate continuous-time and discrete-time systems. The proof is in two steps: first we characterize the internally chain transitive sets of the continuous-time systems (definition below), then we show with stochastic approximations theorem that limit sets of the discrete time systems are included in internally chain transitive sets.

Definition B.1 (Internally chain transitive)**.** A set A is internally chain transitive for a differential inclusion $\frac{dy}{dt} \in F(y)$ if it is compact and if for all $x, x' \in A$, $\epsilon > 0$ and $T > 0$ there exists an integer $n \in \mathbb{N}$, solutions $y_1, \dots y_n$ to the differential inclusion and real numbers $t_1, t_2, \dots, t_n$ greater than T such that:

- $y_i(s) \in A$ for $0 \leq s \leq t_i$
- $\|y_i(t_i) - y_{i+1}(0)\| \leq \epsilon$
- $\|y_1(0) - x| \leq \epsilon$ and $\|y_n(t_n) - x'\| \leq \epsilon$

The previous definition means that if a set A is internally chain transitive, two points can be linked with solutions of at least length T in at most n steps, where n depends on T. These sets contain the limit sets of the discrete-time counterparts of the differential inclusions (see (Benaïm et al. [2005]) for an introduction to the theory of stochastic approximations and formal statements), that is systems of the form:

$$y_{n+1} - y_n \in \beta_n(F(y_n) + U_n)$$

where U_n is typically a zero-mean noise with bounded variance.

In the following, we characterize internally chain transitive sets and use an asynchronous extension of the theory of stochastic approximations initially presented in Perkins [2013] and later extended in Baudin and Laraki [2022].

B.1 Payoff perturbation

In MBRD, we supposed that players did not observe the actual stage reward $r^i_{s_n}(a_n)$ but a perturbation of this value. Formally, the expectancy of R^i_n must be $r^i_{s_n}(a_n)$ conditionally on the history:

$$\mathbb{E}\left[R^i_n | \mathcal{F}_{n-1}\right] = r^i_{s_n}(a_n) \tag{22}$$

where $\mathcal{F}_{n-1}$ is the σ-algebra that contains all information up to step $n-1$. Furthermore, we require the variance to be bounded, i.e. $\mathrm{var}(R^i_n | \mathcal{F}_{n-1})$ is bounded.

B.2 Convergence of estimates

Lemma B.2. *If L is an internally chain transitive set of MBRD, then it is included in A, where:*

$$A := \left\{ (x^i_s, u^i_s, \hat{r}^i_s, \hat{P}_s)_{s,i} \;\middle|\; \forall s, \hat{r}^i_s = r^i_s \wedge \hat{P}_s = P_s \right\}$$

Proof. We notice that functions $\hat{r}^i_s \mapsto \|\hat{r}^i_s(t) - r^i_s\|_\infty$ and $\hat{P}_s \mapsto \|\hat{P}_s(t) - P_s\|_\infty$ are Lyapunov functions with calculations similar to Lemma A.3: their derivative is smaller than $-\alpha_-$ multiplied by their value. This implies that L is contained in the inverse of zero of such functions, that is A. □

B.3 Convergence in identical-interest stochastic games

Lemma B.3. *Suppose that the system is autonomous, that is $\beta(t)$ is constant and we suppose $\beta(t) = 1$ (it can be generalized to any constant). Let L be an internally chain transitive set of MBRD, then:*

$$L \subseteq \left\{ (x_s, u_s, \hat{r}_s, \hat{P}_s)_s \;\middle|\; \forall s, f_{s,u}(x_s) + \epsilon h(x_s) = u_s \wedge x_s = \mathrm{sbr}^i_{s,u_s}(x_s) \wedge \hat{r}_s = r_s \wedge \hat{P}_s = P_s \right\} \tag{23}$$

which means that L contains only regularized equilibria.

Proof. The proof proceeds with a sequence of inclusion. We show that any internally chain transitive set is contained in:

$$A := \left\{ (x_s, u_s, \hat{r}_s, \hat{P}_s)_s \;\middle|\; \forall s, \hat{r}_s = r_s \wedge \hat{P}_s = P_s \right\}$$

$$B := \left\{ (x_s, u_s, \hat{r}_s, \hat{P}_s)_s \mid \forall s, f_{s,u}(x_s) + \epsilon h(x_s) \geq u_s \right\}$$

and then in the set of Eq. (23).

Regarding $L \subseteq A$, this is exactly Lemma B.2 when payoffs functions are equal.

In (9), the term $-2C\exp(-\eta\alpha_- t)$ comes from the fact that $\hat{r}_s(t)$ and $\hat{P}_s(t)$ have not converged yet. Therefore, if we are in A, these terms disappears and the new differential inequality resulting from computations of Lemma A.6 is:

$$\frac{d\Gamma_{s_-} - u_{s_-}}{dt} \geq \beta(t)(\delta - 1)(\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t))$$

which, using Grönwall lemma and $\beta(t) = 1$, implies that:

$$\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t) \geq \left(\Gamma_{s_-(0)}(0) - u_{s_-(0)}(0)\right) \exp\left(-(1-\delta)t\right)$$

Therefore, for all states s:

$$\Gamma_s(t) - u_s(t) \geq \left(\Gamma_{s_-(0)}(0) - u_{s_-(0)}(0)\right) \exp\left(-(1-\delta)t\right) \tag{24}$$

Then, let a be a point of L with for all s, $\Gamma_s - u_s \geq -\epsilon$ with equality for a state (and $\epsilon > 0$). We suppose that the lipschitz constant of Γ_s is 1, other cases are analoguous. By definition, it is linked to itself by solutions of MBRD which are collated with an arbitrary gap, we can take $\epsilon/2$ and of length at least an arbitrary T, take $\log(4)/(1-\delta)$. Now, after the first solution, by (24), the difference between the auxiliary payoff Γ_s and the auxiliary value u_s is greater than $-3\epsilon/2\exp(-\log(4))$ (because the initial value of $\Gamma_s - u_s$ is greater than $-\epsilon - \epsilon/2$ and then using (24)). This implies, by recurrence, that the difference between $\Gamma_s - u_s$ is strictly greater than $-\epsilon$, which is absurd since the end of the chain is a.

Therefore, $L \subseteq B$.

Then, relatively to $A \cap B$, Γ_s is a Lyapunov function because $\frac{d\Gamma_s}{dt}$ is greater than 0 based on (8), and strictly greater than 0 for points outside of the set of (23), which concludes the lemma.

□

Proof of Theorem 3.1. With Lemma B.3, we known that ICT sets of MBRD are contained in the set of regularized equilibria. The rest of the proof uses stochastic approximations results to show that ICT sets contains the limit sets of MFP and SFP.

In order to do this, we use Theorem D.5 of Baudin and Laraki [2022], an extension of Theorems of Perkins and Leslie [2012] and Benaïm et al. [2005].

Let us recall the two systems that we want to relate:

$$\begin{cases} u^i_{s,n+1} - u^i_{s,t} = \dfrac{\beta}{t+1}\left(\hat{f}^i_{s,u^i_n}\left(x_{s,n}\right) + \epsilon h^i\left(x_{s,n}\right) - u^i_{s,t}\right) \\ x_{s,n+1} - x_{s,t} = \dfrac{1_{s=s_t}}{s^\sharp_t}\left(a_n - x_{s,n}\right) \\ \hat{P}_{ss',n+1}\left(a\right) - \hat{P}_{ss',n}\left(a\right) = \dfrac{1_{s_n=s\wedge a(n)=a}}{\sum_{k=0}^n 1_{s_k=s\wedge a(k)=a}}\left(1_{s_{n+1}=s'} - \hat{P}_{ss',n}\left(a\right)\right) \\ \hat{r}_{s,n+1}\left(a\right) - \hat{r}_{s,n}\left(a\right) = \dfrac{1_{s_n=s\wedge a(n)=a}}{\sum_{k=0}^n 1_{s_k=s\wedge a(k)=a}}\left(R^n_i - \hat{r}_{s,n}\left(a\right)\right) \\ a^i_n \sim \mathrm{sbr}^i_{s,u_n}\left(x_{s,n}\right) \end{cases} \tag{MFP}$$

which is MFP rewritten with incremental updates for estimators, and

$$\begin{cases} \dot{u}_s = \beta\left(\hat{f}_{s,u(t)}\left(x_s(t)\right) + \epsilon h\left(x_s(t)\right) - u_s(t)\right) \\ \dot{x}_s = \alpha_s(t)\left(a\left(t\right) - x_s(t)\right) \\ \dot{\hat{P}}_{ss'}\left(b\right) = \alpha_s(t) a\left(t\right)\left(b\right)\left(P_{ss'}(b) - \hat{P}_{ss'}\left(b\right)\left(t\right)\right) \\ \dot{\hat{r}}^i_s\left(b\right) = \alpha_s(t) a\left(t\right)\left(b\right)\left(r_s\left(b\right) - \hat{r}_s\left(b\right)\left(t\right)\right) \\ a^i\left(t\right) = \mathrm{sbr}^i_{s,u(t)}\left(x_s(t)\right) \\ \alpha_s(t) \in [\alpha_-, 1] \end{cases} \tag{MBRD}$$

The first system is of the form:

$$y_{n+1} - y_n \in S_n \cdot \left(F(y_n) + U_n\right) \tag{25}$$

where $\cdot$ is the pairwise multiplication of two vectors, and the second one is:

$$\frac{dy}{dt} \in S_t \cdot F(y) \tag{26}$$

Note that this is the same F between two systems and that the expectancy of U_n is zero. Vectors S_n and S_t are the update rates of every variable. For every discrete variable that are updated with an indicator function, the steps are decreasing in $\frac{1}{n}$ in the number of times that the indicator was equal to 1. For $u^i_{s,t}$, it is also updated in $\frac{1}{n}$. Therefore, our update step sequence is $\gamma_n = \frac{1}{n}$. Furthermore, update steps S_n and S_t are correlated vectors, meaning that the update rate of every variable is not independent (for instance, for a fixed state and a fixed action, $\hat{r}_{s,t}\left(a\right)$ and $\hat{P}_{ss',t}\left(a\right)$ are updated at the same rate. So this is a correlated asynchronous system, as described in Baudin and Laraki [2022].

Now, we must verify every hypothesis of the theorem:

(i) All values of MFP are bounded, so they belong to a compact.

(ii) The right hand side F of MBRD is a Marchaud map because it has bounded, closed and convex images, and its graph is itself closed.

(iii) We use steps $\gamma_n = \frac{1}{n}$ which satisfy the usual properties: it is decreasing, it goes to 0 and the sum is equal to ∞.

(iv) The game is ergodic, and the next states are randomly chosen based uniquely on the current state and history, so properties on the transitions are verified.

(v) There is a correlation between U_n and S_n, therefore it is necessary to prove the Kushner-Clark noise condition separately (see Perkins and Leslie [2012]), one variable at a time. Regarding $\hat{P}_{s,n}$ and $\hat{r}^i_{s,n}$, this is true because the noise is uniformly bounded in L^2 (bounded variance) and the update step is $1/t$, as noted in Proposition 1.4 and Remark 1.5 of Benaïm et al. [2005]. Regarding variables $x_{s,n}$ and $u^i_{s,t}$ there is no correlation between the update steps and the noise, so standard theorems apply (for instance Lemma 3.3 of Perkins and Leslie [2012]).

(vi) The squared sum of γ_n converges and the noise is bounded.

(vii) There is no additional drift, so $d_n = 0$ in our case.

Therefore, we can use the theorem, which implies that the limit set of MFP is an internally chain transitive set of MBRD. Combined with Lemma B.3, this gives the desired result.

□

B.4 Convergence in zero-sum stochastic games

The scheme of the proof is similar to that of identical-interest stochastic games: first, we characterize internally chain transitive sets of MBRD, then we show that systems MBRD and MFP can be related using stochastic approximations theory, thus the characterization of limit sets of MFP. Throughout this part, we use notations and hypothesis from subsection A.5.

Lemma B.4 (Internally chain transitive sets in ZS case)**.** *Let L be an internally chain transitive sets of system MBRD in the ZS case with the same initial values for all players. Then there exists E such that L is contained in the following set:*

$$B^{E\beta^\star} := \left\{ (x_s, u_s^i, \hat{r}_s, \hat{P}_s)_s \;\middle|\; \forall s, |f_{s,u^i}^i(x_s) + \epsilon h^i(x_s) - u_s^i| \leq E\beta^\star \right\}$$

Proof. With notations of subsection A.5, we use the same function $g(u_s^i) = \max\{|u_{s_f}(t) - v_{s_f,u(t)}|, 2\epsilon\}$ as in Lemma A.12. Furthermore, assumption (A1) is satisfied for points of L because L is contained in A (Lemma B.2) and the duality gap $w_s(t)$ is guaranteed to be small enough. Indeed, $w_s(t)$ defined in subsection A.5 is almost a Lyapunov function: relatively to set A, $w_s - 2D\beta^\star\eta^{-1}$ is a Lyapunov function (Lemma A.10). Therefore, Lemma A.11 implies that g is Lyapunov as well (Lemma A.12) and this implies that $L \subseteq B^{E\beta^\star}$. □

Proof of Theorem 3.2. Theorem of stochastic approximations apply identically to the identical-interest stochastic games case in the previous subsection. Therefore, with the characterization of internally chain transitive sets (Lemma B.4), we have an inclusion for the limit sets of MFP in the zero-sum case as well. □

C Examples and Simulations

See the notebook in the supplementary material for an example of stochastic game with an implementation of the proposed algorithm and simulations.