Lancaster University

PROWLER.io
the decision company

Best-response Dynamics in Zero-sum Stochastic Games

March 13, 2018

David S. Leslie, with Steve Perkins and Zibo Xu

Introduction

BR dynamics in normal form games

Lancaster University

PROWLER.io the decision company

Given a game with

- fintely many players i,
- finite action sets A^i,
- reward functions $r^i : \times_i A^i \to \mathbb{R}$,
- mixed strategies $x^i \in \Delta(A^i)$, and
- best responses $b^i(x^{-i}) = \underset{y^i \in \Delta(A^i)}{\operatorname{argmax}}\, r^i(y^i, x^{-i})$,

the best response dynamics are

$$\dot{x}^i \in b^i(x^{-i}) - x^i$$

Introduction

Stochastic games (Shapley 1953)

Lancaster University

PROWLER.io the decision company

- Players $i \in I$
- Finite set of states $s \in S$
- Finite action sets A_s^i
- Transition probabilities $P_{s,s'}(a)$ for $s, s' \in S$, $a \in \prod_{i \in I} A_s^i$
- Stage payoff functions $r_s^i(\cdot) : \prod_{i \in I} A_s^i \to \mathbb{R}$
- Discount factor $\delta \in (0, 1)$

Introduction

Stochastic games (Shapley 1953)

Lancaster University

PROWLER.io the decision company

- Players $i \in I$
- Finite set of states $s \in S$
- Finite action sets A_s^i
- Transition probabilities $P_{s,s'}(a)$ for $s, s' \in S$, $a \in \prod_{i \in I} A_s^i$
- Stage payoff functions $r_s^i(\cdot) : \prod_{i \in I} A_s^i \to \mathbb{R}$
- Discount factor $\delta \in (0, 1)$

Focus on two-player zero-sum stochastic games, in which:

- $I = \{1, 2\}$
- $r_s^1(a) = -r_s^2(a) =: r_s(a)$

Introduction

BR dynamics in stochastic games

Lancaster University

PROWLER.io the decision company

BR dynamics

Let x_s^i be the mixed strategy of Player i in state s. Then the dynamics are

$$\dot{x}_s^i = b_s^i(x^{-i}) - x_s^i$$

But what is $b_s^i(x^{-i})$?

- The action for state s from the overall best response to x^{-i}?
- The best response in the auxiliary game formed at s with continuation payoffs given by (x^i, x^{-i})?
- ...

Plan

Lancaster University

PROWLER.io
the decision company

- Recap relevant results from normal form games
- Open-loop BR dynamics
- Closed-loop BR dynamics

Normal form games

Value and energy

Lancaster University

PROWLER.io the decision company

Value

Player 1 can guarantee she receives the game's value

$$v(G) = \max_{x^1} \min_{x^2} r(x^1, x^2)$$

Energy

Given a joint mixed strategy $x := (x^1, x^2)$, we define the energy

$$w(x) = \max_{y^1} r(y^1, x^2) - \min_{y^2} r(x^1, y^2)$$

Normal form games

Properties of energy

Lancaster University

PROWLER.io the decision company

Lemma

$$|r(x) - v(G)| \leq w(x)$$

Theorem (Harris 1998, Hofbauer and Sorin 2006)

If $\dot{x}^i \in b^i(x^{-i}) - x^i$ then

$$\dot{w} = -w.$$

Therefore $w(x(t)) = \mathrm{e}^{-t} w(x(0))$

Stochastic games

Preliminaries

Lancaster University

PROWLER.io the decision company

- A stationary strategy is an $x^i \in \times_{s \in S} \Delta(A^i_s)$
- A strategy profile is an $x = (x^1, x^2) = ((x^1_s)_{s \in S}, (x^2_x)_{s \in S})$
- The expected stage payoff is $r_s(x) = \mathbb{E}_{a^i \sim x^i} r_s(a^1, a^2)$
- The transition function is $P_{s,s'}(x) = \mathbb{E}_{a^i \sim x^i} P_{s,s'}(a^1, a^2)$

The expected discounted payoff for Player 1 starting in s is

$$\begin{aligned} U_s(x) &= \mathbb{E}\left[(1-\delta)\sum_{n=0}^{\infty} \delta^n r_{s_n}(x) \,|\, s_0 = s\right] \\ &= (1-\delta) r_s(x) + \delta \sum_{s'} P_{s,s'}(x) U_{s'}(x) \end{aligned}$$

Stochastic games

Value

Lancaster University

PROWLER.io the decision company

- Shapley (1953) proves the existence of equilibrium stationary strategy profiles.
- As in normal form games, Player 1 can guarantee herself a value, now dependent on s.
- Let x^* be a stationary optimal strategy. Then

$$\mathrm{Val}_s = (1-\delta) r_s(x^*) + \delta \sum_{s'} P_{s,s'}(x^*)\mathrm{Val}_{s'}$$

Stochastic games

Auxiliary games - definition

Lancaster University

PROWLER.io the decision company

Suppose we know the expected discounted future reward $u_{s'}$ from each state.
Then from state s the joint action a will give Player 1

$$f_{s,\vec{u}}(a) := (1-\delta)r_s(a) + \delta \sum_{s'} P_{s,s'}(a)u_{s'}$$

Denote by $G_{s,\vec{u}}$ the auxiliary game with payoffs $f_{s,\vec{u}}$

Stochastic games

Auxiliary games - definition

Lancaster University

PROWLER.io the decision company

Suppose we know the expected discounted future reward $u_{s'}$ from each state.
Then from state s the joint action a will give Player 1

$$f_{s,\vec{u}}(a) := (1-\delta)r_s(a) + \delta \sum_{s'} P_{s,s'}(a)u_{s'}$$

Denote by $G_{s,\vec{u}}$ the auxiliary game with payoffs $f_{s,\vec{u}}$

Note that

$$|f_{s,\vec{u}}(x) - v(G_{s,\vec{u}})| = |f_{s,\vec{u}}(x) - u_s| = 0$$

implies that $f_{x,\vec{u}}(x) = u_s = \mathrm{Val}_s$ for all s.

Stochastic games

Auxiliary games - best responses and energy

Lancaster University

PROWLER.io the decision company

Best responses defined in the auxiliary games:

$$b^i_{s,\vec{u}} := \underset{y \in \Delta(A^i_s)}{\operatorname{argmax}} f_{s,\vec{u}}(y, x_s^{-i}).$$

Auxiliary game energy also depends on $\vec{u}$:

$$w_{s,\vec{u}}(x) = \max_{y^1} f_{s,\vec{u}}(y^1, x_s^2) - \min_{y^2} f_{s,\vec{u}}(x^1, y^2)$$

Open-loop dynamics

Perkins (2013)

Lancaster University

PROWLER.io the decision company

$$\dot{x}_s^i \in b^i_{s,\vec{U}(x)}(x^{-i}) - x_s^i$$

Strategy at s adjusts towards the best action if it's assumed players revert to x^i after the next transition

Theorem

In a two-player zero-sum stochastic game with

$$\delta \leq \left(1 + \max_s \sum_{s'} \max_a P_{s,s'}(a)\right)^{-1}$$

any open-loop best-response trajectory converges to the set of stationary equilibria.

Fixed continuation payoffs

Give me a moment. . .

Lancaster University

PROWLER.io the decision company

Suppose we fix continuation payoffs $\vec{u}$

Each auxiliary game is now an independent and fixed two-player zero-sum game with payoff function $f_{s,\vec{u}}(a)$

Under BR dynamics using $b_{s,\vec{u}}$, the stage game energy $w_{s,\vec{u}}$ converges to 0 at exponential rate

Closed-loop dynamics

Lancaster University

PROWLER.io the decision company

Instead of $\vec{u}$ being fixed, let it change continuously but slowly:

$$\begin{aligned} \dot{x}_s^i(t) &= b^i_{s,\vec{u}(t)}(x(t)) - x_s^i(t) \\ \dot{u}_s(t) &= \frac{1}{t+1}\left\{ f_{s,\vec{u}(t)}(x(t)) - u_s(t) \right\} \end{aligned}$$

There is constant feedback from 'perceived' continuation payoffs $f_{s,\vec{u}(t)}(x(t))$ into 'believed' continuation payoffs $u_s(t)$.

Closed-loop dynamics

Proposition: $w_{s,\vec{u}(t)}(x(t)) \to 0$

Lancaster University

PROWLER.io the decision company

Recall that $w_{s,\vec{u}(t)}(x(t)) \to 0$ implies

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

so that $x_s(t)$ is close to equilibrium play in the auxiliary game $G_{s,\vec{u}(t)}$.

Closed-loop dynamics

Proposition: $w_{s,\vec{u}(t)}(x(t)) \to 0$

Lancaster University

PROWLER.io
the decision company

$$\frac{\mathrm{d}}{\mathrm{d}t} w_{s,\vec{u}(t)}(x(t)) = \dot{\vec{u}} \cdot D_{\vec{u}} w_{s,\vec{u}(t)}(x(t)) + \dot{x}_s \cdot D_{x_s} w_{s,\vec{u}(t)}(x(t))$$

Closed-loop dynamics

Proposition: $w_{s,\vec{u}(t)}(x(t)) \to 0$

$$\frac{\mathrm{d}}{\mathrm{d}t} w_{s,\vec{u}(t)}(x(t)) = \dot{\vec{u}} \cdot D_{\vec{u}} w_{s,\vec{u}(t)}(x(t)) + \dot{x}_s \cdot D_{x_s} w_{s,\vec{u}(t)}(x(t))$$

- For sufficiently large t,

$$|\dot{u}_s(t)| = \left| \frac{1}{t+1} \left\{ f_{s,\vec{u}(t)}(x(t)) - x_s^i(t) \right\} \right| \leq \epsilon.$$

- A change of continuation payoffs of at most η changes $w_{s,\vec{u}(t)}(x(t))$ by at most $2\delta\eta$, so

$$D_{\vec{u}} w_{s,\vec{u}(t)}(x(t)) \leq 2\delta$$

- Therefore $\dot{\vec{u}} \cdot D_{\vec{u}} w_{s,\vec{u}(t)}(x(t)) \leq 2\delta\epsilon$

Closed-loop dynamics

Proposition: $w_{s,\vec{u}(t)}(x(t)) \to 0$

$$\frac{\mathrm{d}}{\mathrm{d}t} w_{s,\vec{u}(t)}(x(t)) = \dot{\vec{u}} \cdot D_{\vec{u}} w_{s,\vec{u}(t)}(x(t)) + \dot{x}_s \cdot D_{x_s} w_{s,\vec{u}(t)}(x(t))$$

- The second term is a BR dynamics and energy as if we had a fixed normal form game with payoffs $f_{s,\vec{u}(t)}(\cdot)$.
- So as in Harris (1998) or Hofbauer and Sorin (2006),

$$\dot{x}_s \cdot D_{x_s} w_{s,\vec{u}(t)}(x(t)) \leq -w_{s,\vec{u}(t)}(x(t))$$

Closed-loop dynamics

Proposition: $w_{s,\vec{u}(t)}(x(t)) \to 0$

Lancaster University

PROWLER.io the decision company

$$\frac{\mathrm{d}}{\mathrm{d}t} w_{s,\vec{u}(t)}(x(t)) = \dot{\vec{u}} \cdot D_{\vec{u}} w_{s,\vec{u}(t)}(x(t)) + \dot{x}_s \cdot D_{x_s} w_{s,\vec{u}(t)}(x(t))$$

$$\begin{aligned} \dot{\vec{u}} \cdot D_{\vec{u}} w_{s,\vec{u}(t)}(x(t)) &\leq 2\delta\epsilon \\ \dot{x}_s \cdot D_{x_s} w_{s,\vec{u}(t)}(x(t)) &\leq -w_{s,\vec{u}(t)}(x(t)) \end{aligned}$$

So for sufficiently large t,

$$\frac{\mathrm{d}}{\mathrm{d}t} w_{s,\vec{u}(t)}(x(t)) \leq -w_{s,\vec{u}(t)}(x(t)) + 2\delta\epsilon.$$

Since ϵ can be arbitrarily small, the result follows.

Closed-loop dynamics

Consequence of $w_{s,\vec{u}(t)}(x(t)) \to 0$

Lancaster University

PROWLER.io the decision company

Recall (again) that $w_{s,\vec{u}(t)}(x(t)) \to 0$ implies

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

Fixing $\epsilon > 0$, this allows us to define $t_1(\epsilon)$ be such that

$$\max_s |f_{s,\vec{u}(t)}(x(t)) - v(G_{s,\vec{u}(t)})| \leq (1-\delta)\epsilon/16$$

for all $t \geq t_1(\epsilon)$

Assume for the rest of the proof that $t > t_1(\epsilon)$ so that $x_s(t)$ is always close to an equilibrium of $G_{s,\vec{u}(t)}$.

Closed-loop dynamics

Continuation payoffs converge (Lemma A)

Lancaster University

PROWLER.io the decision company

This is surprisingly tricky to show. Start with a technical lemma.
Definition: $s_f(t) \in \operatorname*{argmax}_s |f_{s,\vec{u}(t)} - u_s(t)|$.

Lemma A

If $\max_s |f_{s,\vec{u}(t)} - u_s(t)| > \epsilon$ then for any s such that

$$\left| |u_s(t) - v(G_{s,\vec{u}(t)})| - |u_{s_f(t)}(t) - v(G_{s_f(t),\vec{u}(t)})| \right| \leq \frac{(1-\delta)\epsilon}{8}$$

it follows that

$$\frac{\mathrm{d}}{\mathrm{d}t} \left| u_s(t) - v(G_{s,\vec{u}(t)}) \right| \leq -\frac{3(1-\delta)\epsilon}{4(1+t)}.$$

Closed-loop dynamics

Continuation payoffs converge (Proof of Lem A)

Lancaster University

PROWLER.io the decision company

Suppose s is such that

$$\left| |u_s(t) - v(G_{s,\vec{u}(t)})| - |u_{s_f(t)}(t) - v(G_{s_f(t),\vec{u}(t)})| \right| \leq \frac{(1-\delta)\epsilon}{8}$$

Then since $\max_s |f_{s,\vec{u}(t)}(x(t)) - v(G_{s,\vec{u}(t)})| \leq (1-\delta)\epsilon/16$,

$$|u_s(t) - f_{s,\vec{u}(t)}(x(t))| - |u_{s_f(t)}(t) - f_{s_f(t),\vec{u}(t)}(x(t))| \geq -\frac{(1-\delta)\epsilon}{4}$$

Dividing by $t+1$ we see that

$$|\dot{u}_s(t)| \geq |\dot{u}_{s_f(t)}(t)| - \frac{(1-\delta)\epsilon}{4(t+1)}$$

Closed-loop dynamics

Continuation payoffs converge (Proof of Lem A)

Lancaster University

PROWLER.io the decision company

We have that $|\dot{u}_s(t)| \geq |\dot{u}_{s_f(t)}(t)| - \frac{(1-\delta)\epsilon}{4(t+1)}$.
Now

$$\max_s |f_{s,\vec{u}(t)} - u_s(t)| > \epsilon \quad \Leftrightarrow \quad |\dot{u}_{s_f(t)}| > \epsilon/(t+1)$$

Therefore

$$|\dot{u}_s(t)| \geq |\dot{u}_{s_f(t)}(t)| - \frac{1-\delta}{4}|\dot{u}_{s_f(t)}(t)| = \frac{3+\delta}{4}|\dot{u}_{s_f(t)}(t)|$$

and

$$\left|\frac{\mathrm{d}}{\mathrm{d}t} v(G_{s,\vec{u}(t)})\right| \leq \delta \max_{s'} |\dot{u}_{s'}(t)| = \delta |\dot{u}_{s_f(t)}(t)|$$

Closed-loop dynamics

Continuation payoffs converge (Proof of Lem A)

Lancaster University

PROWLER.io the decision company

If $u_s(t) > v(G_{s,\vec{u}(t)})$, our conditions give that $\dot{u}_s(t) < 0$.
Since $|\dot{u}_s(t)| > \frac{3+\delta}{4}|\dot{u}_{s_f(t)}(t)|$, it follows that

$$\dot{u}_s(t) < -\frac{3+\delta}{4}|\dot{u}_{s_f(t)}(t)|$$

Furthermore,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left|u_s(t) - v(G_{s,\vec{u}(t)})\right| &\leq \dot{u}_s(t) + \frac{\mathrm{d}}{\mathrm{d}t}|v(G_{s,\vec{u}(t)})| \\
&< -\frac{3+\delta}{4}|\dot{u}_{s_f(t)}(t)| + \delta|\dot{u}_{s_f(t)}(t)| \\
&= -\frac{3}{4}(1-\delta)|\dot{u}_{s_f(t)}(t)| \\
&< -\frac{3}{4}(1-\delta)\frac{\epsilon}{t+1}
\end{aligned}$$

Closed-loop dynamics

Continuation payoffs converge (Lemma B)

Lancaster University

PROWLER.io the decision company

Lemma B

For sufficiently large t,

- $\max_s |v(G_{s,\vec{u}(t)}) - u_s(t)| < \left(1 + \frac{3(1-\delta)}{16}\right)\epsilon$
- $\max_s |f_{s,\vec{u}(t)}(x(t)) - u_s(t)| < \left(\frac{3-\delta}{2}\right)\epsilon$

We have already seen that

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

so the two statements are equivalent

If $\max_s |f_{s,\vec{u}(t)}(x(t)) - u_s(t)| < \epsilon$ the result follows immediately.

Closed-loop dynamics

Continuation payoffs converge (Proof of Lem B)

Lancaster University

PROWLER.io the decision company

Definition: $s_v(t) \in \underset{s}{\operatorname{argmax}} |v(G_{s,\vec{u}(t)}) - u_s(t)|$. We can show

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

So the condition of Lemma A applies to $s_v(t)$ automatically,

$$\max_s |f_{s,\vec{u}(t)} - u_s(t)| > \epsilon \quad \Rightarrow \quad \frac{\mathrm{d}}{\mathrm{d}t} \left| u_{s_v(t)}(t) - v(G_{s_v(t),\vec{u}(t)}) \right| \leq -\frac{3(1-\delta)\epsilon}{4(1+t)}$$

and $\exists t_2$ s.t. $\max_s |f_{s,\vec{u}(t_2)} - u_s(t_2)| \leq \epsilon$

Closed-loop dynamics

Continuation payoffs converge (Proof of Lem B)

Lancaster University

PROWLER.io the decision company

Consider arbitrary $t > t_2$ with $\max_s |f_{s,\vec{u}(t)}(x(t)) - u_s(t)| \geq \epsilon$.

Let $t_3 = \sup\{t' \leq t : \forall \tau \in (t', t)\ \max_s |f_{s,\vec{u}(\tau)}(x(\tau)) - u_s(\tau)| \geq \epsilon\}$.

The definition gives us $\max_s |f_{s,\vec{u}(t_3^-)} - u_s(t_3^-)| \leq \epsilon$, which implies

$$\max_s |v(G_{s,\vec{u}(t_3^-)}) - u_s(t_3^-)| < \left(1 + \frac{3(1-\delta)}{16}\right)\epsilon$$

Closed-loop dynamics

Continuation payoffs converge (Proof of Lem B)

Since $\frac{\mathrm{d}}{\mathrm{d}t}\left|u_{s_v(t)}(t) - v(G_{s_v(t),\vec{u}(t)})\right| < 0$ for $t' \in (t_3, t)$,

$$\max_s |v(G_{s,\vec{u}(t')}) - u_s(t')| < \left(1 + \frac{3(1-\delta)}{16}\right)\epsilon$$

for all $t' \in [t_3, t]$.

Various simple bounds lead to

$$\max_s |f_{s,\vec{u}}(x(t)) - u_s(t)| < \left(1 + \frac{3(1-\delta)}{8}\right)\epsilon.$$

Closed-loop dynamics

Convergence

Lancaster University

PROWLER.io the decision company

We know that

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

Theorem

$u_s(t) \to \mathrm{Val}_s$, $f_{s,\vec{u}(t)}(x(t)) \to \mathrm{Val}_s$, and $x(t)$ converges to the set of Nash equilibrium strategy profiles.

Closed-loop dynamics

Rate of convergence

Lancaster University

PROWLER.io the decision company

Theorem

$\exists K(\delta)$ such that $\forall s$

$$|u_s(t) - \mathrm{Val}_s| \leq \frac{K(\delta)}{t}.$$

δ-converging dynamics

Convergence to asymptotic value

Lancaster University

PROWLER.io the decision company

Add in evolution of $\delta(t)$:

$$\dot{\delta}(t) = \frac{1 - \delta(t)}{(t+1)\log(t+1)}$$

Theorem

$u_s(t)$ and $f_{s,\vec{u}(t),\delta(t)}(x(t))$ converge to the asymptotic value of the game

Summary of results

Lancaster University

PROWLER.io
the decision company

- Convergence in open-loop dynamics proved only in δ sufficiently small
- Convergence (at rate $1/t$) of closed-loop dynamics
- Convergence to asymptotic value of δ-converging dynamics

What's still to come?

Lancaster University

PROWLER.io the decision company

In a nutshell, learning not computation

- Learning while playing; beliefs about x_s only update when actually at s
- Learning while playing; stochastic approximation
- Initial estimates u_s not the same for each player
- Different update rates at different states

Suggestions (and questions) welcome

Lancaster University

PROWLER.io
the decision company

March 13, 2018

David S. Leslie, with Steve Perkins and Zibo Xu