LEARNING IN GAMES WITH NOISY PAYOFF OBSERVATIONS

Mario Bravo[1] Panayotis Mertikopoulos[2]

[1]Universidad de Santiago de Chile

[2]CNRS – Laboratoire d'Informatique de Grenoble

ADGO 2016 – Santiago, January 28, 2016

Outline

Learning in Games

The basic context:

- Decision-making: *agents choose actions, each seeking to optimize some objective.*
- Payoffs: *rewards are determined by the decisions of all interacting agents.*
- Learning: the agents adjust their decisions and the process continues.

Learning in Games

The basic context:

- *Decision-making: agents choose actions, each seeking to optimize some objective.*

  Example: *a trader chooses asset proportions in an investment portfolio.*

- *Payoffs*: *rewards are determined by the decisions of all interacting agents.*

  Example: asset placements determine returns.

- *Learning:* the agents adjust their decisions and the process continues.

  Example: change asset proportions based on performance.

Learning in Games

The basic context:

- *Decision-making: agents choose actions, each seeking to optimize some objective.*
  Example: *a trader chooses asset proportions in an investment portfolio.*
- *Payoffs: rewards are determined by the decisions of all interacting agents.*
  Example: asset placements determine returns.
- *Learning:* the agents adjust their decisions and the process continues.
  Example: change asset proportions based on performance.

When does the agents' learning process lead to a "reasonable" outcome?

Motivation

- In many applications, decisions taken at very fast time-scales.
- Regulations/physical constraints limit changes in decisions.
- Fast time-scales have adverse effects on quality of feedback.

Motivation

- In many applications, decisions taken at very fast time-scales.

  Example: in high-frequency trading (HFT), decision times $\approx 100\ \mu s$.

- Regulations/physical constraints limit changes in decisions.

  Example: the SEC requires small differences in HFT orders to reduce volatility.

- Fast time-scales have adverse effects on quality of feedback.

  Example: volatility estimates highly inaccurate at the $100\ \mu s$ time-scale.

The Flash Crash of 2010

A trillion-dollar NYSE crash (and partial rebound) that lasted 35 minutes (14:42–15:07)

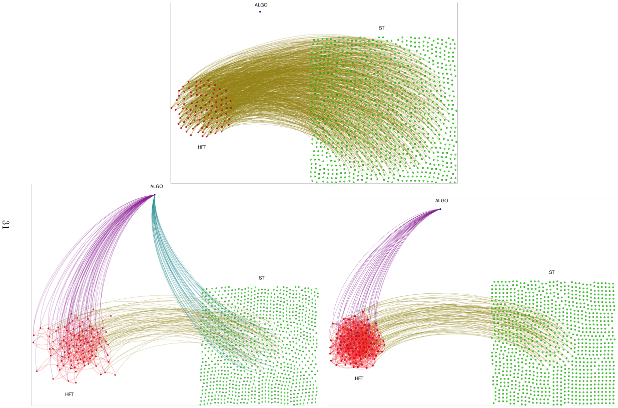


Figure 5: Network snapshots of the market behaving normally (top), when ALGO starts selling and HFTs absorb the initial sell pressure a moment before the hot-potato effect starts (bottom left), and when the price reaches its trough (bottom right).

Aggressive selling due to imperfect volatility estimates induced a huge drop in liquidity and precipitated the crash (Vuorenmaa and Wang, 2014)

What this talk is about:

Examine the robustness of a class of continuous-time learning schemes with noisy feedback.

Outline

Game setup

Throughout this talk, we focus on *finite games:*

- Finite set of *players*: $\mathcal{N} = \{1, \dots, N\}$
- Finite set of *actions* per player: $\mathcal{A}_k = \{\alpha_{k,1}, \alpha_{k,2}, \dots\}$
- Reward of player k determined by corresponding *payoff function* $u_k \colon \prod_k \mathcal{A}_k \to \mathbb{R}$:

$$(\alpha_1, \dots, \alpha_n) \mapsto u_k(\alpha_1, \dots, \alpha_N)$$

Game setup

Throughout this talk, we focus on *finite games:*

- Finite set of *players*: $\mathcal{N} = \{1, \dots, N\}$
- Finite set of *actions* per player: $\mathcal{A}_k = \{\alpha_{k,1}, \alpha_{k,2}, \dots\}$
- Reward of player k determined by corresponding *payoff function* $u_k \colon \prod_k \mathcal{A}_k \to \mathbb{R}$:

$$(\alpha_1, \dots, \alpha_n) \mapsto u_k(\alpha_1, \dots, \alpha_N)$$

- *Mixed strategies* $x_k \in \mathcal{X}_k \equiv \Delta(\mathcal{A}_k)$ yield *expected payoffs*

$$u_k(x_1, \dots, x_N) = \sum_{\alpha_1} \dots \sum_{\alpha_N} x_{1,\alpha_1} \cdots x_{N,\alpha_N} \, u_k(\alpha_1, \dots, \alpha_N)$$

- *Strategy profiles:* $x = (x_1, \dots, x_N) \in \mathcal{X} \equiv \prod_k \mathcal{X}_k$

Game setup

Throughout this talk, we focus on *finite games:*

- Finite set of *players*: $\mathcal{N} = \{1, \dots, N\}$
- Finite set of *actions* per player: $\mathcal{A}_k = \{\alpha_{k,1}, \alpha_{k,2}, \dots\}$
- Reward of player k determined by corresponding *payoff function* $u_k \colon \prod_k \mathcal{A}_k \to \mathbb{R}$:

$$(\alpha_1, \dots, \alpha_n) \mapsto u_k(\alpha_1, \dots, \alpha_N)$$

- *Mixed strategies* $x_k \in \mathcal{X}_k \equiv \Delta(\mathcal{A}_k)$ yield *expected payoffs*

$$u_k(x_1, \dots, x_N) = \sum\nolimits_{\alpha_1} \dots \sum\nolimits_{\alpha_N} x_{1,\alpha_1} \cdots x_{N,\alpha_N} \, u_k(\alpha_1, \dots, \alpha_N)$$

- *Strategy profiles:* $x = (x_1, \dots, x_N) \in \mathcal{X} \equiv \prod_k \mathcal{X}_k$
- *Payoff vector* of player k: $v_k(x) = (v_{k\alpha}(x))_{\alpha \in \mathcal{A}_k}$ where

$$v_{k\alpha}(x) = v_k(\alpha; x_{-k})$$

is the payoff to the α-th action of player k in the mixed strategy profile $x \in \mathcal{X}$.

Regret

Suppose players follow a *trajectory of play* $x(t)$ (based on some learning/adjustment rule, to be discussed later).

How does $x_k(t)$ compare on average to the "best possible" action $\alpha_k \in \mathcal{A}_k$?

$$u_k(\alpha; x_{-k}(s)) - u_k(x(s))$$

Regret

Suppose players follow a *trajectory of play* $x(t)$ (based on some learning/adjustment rule, to be discussed later).

How does $x_k(t)$ compare on average to the "best possible" action $\alpha_k \in \mathcal{A}_k$?

$$\int_0^t u_k(\alpha; x_{-k}(s)) - u_k(x(s))\, ds$$

Regret

Suppose players follow a *trajectory of play* $x(t)$ (based on some learning/adjustment rule, to be discussed later).

How does $x_k(t)$ compare on average to the "best possible" action $\alpha_k \in \mathcal{A}_k$?

$$\max_{\alpha \in \mathcal{A}_k} \int_0^t u_k(\alpha; x_{-k}(s)) - u_k(x(s))\, ds$$

Regret

Suppose players follow a *trajectory of play* $x(t)$ (based on some learning/adjustment rule, to be discussed later).

How does $x_k(t)$ compare on average to the "best possible" action $\alpha_k \in \mathcal{A}_k$?

$$\mathrm{Reg}_k(t) = \max_{\alpha \in \mathcal{A}_k} \int_0^t u_k(\alpha; x_{-k}(s)) - u_k(x(s))\, ds$$

Regret

Suppose players follow a *trajectory of play* $x(t)$ (based on some learning/adjustment rule, to be discussed later).

How does $x_k(t)$ compare on average to the "best possible" action $\alpha_k \in \mathcal{A}_k$?

$$\mathrm{Reg}_k(t) = \max_{\alpha \in \mathcal{A}_k} \int_0^t u_k(\alpha; x_{-k}(s)) - u_k(x(s))\, ds$$

Definition

$x(t)$ leads to *no regret* if $\mathrm{Reg}_k(t) = o(t)$ for all $k \in \mathcal{N}$, i.e. if every player's average regret is non-positive in the long run.

NB: unilateral definition, no need for a game

Dominated strategies

Definition

A (pure) strategy $\alpha \in \mathcal{A}_k$ is *dominated by* $\beta \in \mathcal{A}_k$ if

$$v_{k\alpha}(x) < v_{k\beta}(x) \quad \text{for all } x \in \mathcal{X}.$$

More generally, a *mixed* strategy $p \in \mathcal{X}_k$ is *dominated by* $q \in \mathcal{X}_k$ if

$$\langle v_k(x) | p - q \rangle < 0 \quad \text{for all } x \in \mathcal{X}.$$

Variants: weakly/iteratively dominated defined analogously.

Nash equilibrium

Definition

A strategy profile $x^* \in \mathcal{X}$ is a *Nash equilibrium* if

$$u_k(x_k^*; x_{-k}^*) \geq u_k(x_k; x_{-k}^*) \quad \text{for all } x_k \in \mathcal{X}_k,\ k \in \mathcal{N}, \tag{NE}$$

i.e. when no player has an incentive to deviate from x^*.

Variants:

- Pure: x^* is a corner of $\mathcal{X}$ (the support of x^* is a singleton)
- Strict: (NE) holds as an equality iff $x_k = x_k^*$ for all $k \in \mathcal{N}$; equivalently, x^* is strict iff x^* is pure and

  $$u_k(\alpha; x_{-k}^*) < u_k(x^*) \quad \text{for all } \alpha \notin \operatorname{supp}(x_k^*)$$
- Restricted: (NE) holds for all x_k whose support is contained in that of x_k^* (like Nash equilibrium but players not allowed to deviate to actions not present in x^*)

strict ⊆ pure ⊆ Nash ⊆ restricted

Some basic questions

- Does $x(t)$ lead to no regret?
- Are dominated strategies eliminated along $x(t)$?
- What are the possible limit points of $x(t)$?
- Does $x(t)$ converge to Nash equilibrium?
- If not, do time averages converge?
- …

Outline

Exponential reinforcement learning

A well-known strategy adjustment process is *exponential learning*:

$$
\begin{aligned}
\dot{y}_{k\alpha} &= v_{k\alpha}(x) \\
x_{k\alpha}(t) &= \frac{\exp(y_{k\alpha}(t))}{\sum_\beta \exp(y_{k\beta}(t))}
\end{aligned}
\tag{XL}
$$

Exponential reinforcement learning

A well-known strategy adjustment process is *exponential learning*:

$$
\begin{aligned}
y_{k\alpha}(t) &= \int_0^t v_{k\alpha}(x(s))\,ds \\
x_{k\alpha}(t) &= \frac{\exp(y_{k\alpha}(t))}{\sum_\beta \exp(y_{k\beta}(t))}
\end{aligned}
\tag{XL}
$$

In words:

- Score actions based on their cumulative payoffs.
- Assign probability weights exponentially proportionally to these scores.

(Exponential reinforcement of highest scoring strategies).

Exponential reinforcement learning

A well-known strategy adjustment process is *exponential learning*:

$$
\begin{aligned}
y_{k\alpha}(t) &= \int_0^t v_{k\alpha}(x(s))\, ds \\
x_{k\alpha}(t) &= \frac{\exp(y_{k\alpha}(t))}{\sum_\beta \exp(y_{k\beta}(t))}
\end{aligned}
\tag{XL}
$$

In words:

- Score actions based on their cumulative payoffs.
- Assign probability weights exponentially proportionally to these scores.

(Exponential reinforcement of highest scoring strategies).

Continuous-time analogue of EXP3/EWA class of online learning algorithms (Vovk, 1990; Littlestone and Warmuth, 1994; Sorin, 2009;...)

Links with evolutionary game theory

Trajectories of play under (XL) follow the replicator dynamics (Taylor & Jonker, 1978):

$$\dot{x}_{k\alpha} = x_{k\alpha}\left[v_{k\alpha}(x) - \sum\nolimits_{\beta} x_{k\beta} v_{k\beta}(x)\right] \tag{RD}$$

Most widely studied dynamics in evolutionary game theory; known properties include:

- Dominated strategies become extinct under interior solutions of (RD)
- Nash equilibria are stationary under (RD); stationary points of (RD) are restricted equilibria
- Limit points of interior solutions are Nash equilibria
- Strict Nash equilibria are locally stable and attracting
- Convergence to restricted equilibria in potential games.
- …

An alternative characterization of exponential learning

The logit map $y_\alpha \mapsto e^{y_\alpha} / \sum_\beta e^{y_\beta}$ can be equivalently characterized as

$$y \mapsto \arg\max_{x \in \Delta} \{ \langle y | x \rangle - h(x) \}$$

where $h(x) = -\sum_\beta x_\beta \log x_\beta$ is the (negative) Gibbs entropy.

In words:

Agents play mixed strategies that maximize their expected cumulative payoff minus a penalty.

Interpretation:

The entropic penalty promotes exploration (contrast to greedily playing $\arg\max \langle y | x \rangle$)

Reinforcement learning via regularization

A general reinforcement principle:

- Score actions by keeping track of their cumulative payoffs over time.
- Play an "approximate" best response to the resulting score vector

Reinforcement learning via regularization

A general reinforcement principle:

- Score actions by keeping track of their cumulative payoffs over time.
- Play an "approximate" best response to the resulting score vector

Formally:

$$\begin{aligned} \dot{y}_k &= v_k(x) \\ x_k(t) &= Q_k(y_k(t)) \end{aligned} \tag{RL}$$

where the *approximate best response* (or *choice map*) Q_k is defined as

$$Q_k(y_k) = \operatorname*{arg\,max}_{x_k \in \mathcal{X}_k} \{\langle y_k | x_k \rangle - h_k(x_k)\}$$

for some *penalty function* $h_k \colon \mathcal{X}_k \to \mathbb{R}$

Assumptions for h:
Continuous on $\mathcal{X}$; smooth on interiors of faces; strongly convex:

$$h(tx + (1-t)x) \leq th(x) + (1-t)h(x) - \tfrac{1}{2}Kt(1-t)\|x - x'\|^2 \quad \text{for all } t \in [0,1]$$

Examples

Ex. 1. Entropic penalty:

$$h(x) = \sum_\beta x_\beta \log x_\beta$$

Induces the logit map

$$G_\alpha(v) = \frac{\exp(v_\alpha)}{\sum_\beta \exp(v_\beta)}$$

Examples

Ex. 1. Entropic penalty:

$$h(x) = \sum_\beta x_\beta \log x_\beta$$

Induces the logit map

$$G_\alpha(v) = \frac{\exp(v_\alpha)}{\sum_\beta \exp(v_\beta)}$$

Ex. 2. Quadratic penalty:

$$h(x) = \frac{1}{2} \sum_\beta x_\beta^2$$

Induces the closest point projection map

$$\Pi(v) = \arg\min_{x \in \Delta} \|v - x\| = \operatorname{proj}_\Delta v$$

Examples

Ex. 1. Entropic penalty:

$$h(x) = \sum_\beta x_\beta \log x_\beta$$

Induces the logit map

$$G_\alpha(v) = \frac{\exp(v_\alpha)}{\sum_\beta \exp(v_\beta)}$$

Ex. 2. Quadratic penalty:

$$h(x) = \frac{1}{2} \sum_\beta x_\beta^2$$

Induces the closest point projection map

$$\Pi(v) = \operatorname*{arg\,min}_{x \in \Delta} \|v - x\| = \operatorname{proj}_\Delta v$$

Important dichotomy: h *is steep* $\rightsquigarrow \operatorname{im} Q = \Delta^\circ$; h *is non-steep* $\rightsquigarrow \operatorname{im} Q = \Delta$

Examples of dynamics

Ex. 1 The entropic penalty leads to exponential reinforcement learning:

$$\begin{aligned} \dot{y}_{k\alpha} &= v_{k\alpha}(x) \\ x_{k\alpha} &= \frac{\exp(y_{k\alpha})}{\sum_\beta \exp(y_{k\beta})} \end{aligned} \tag{XL}$$

Trajectories of (XL) satisfy the replicator dynamics

Examples of dynamics

Ex. 1 The entropic penalty leads to exponential reinforcement learning:

$$\begin{aligned} \dot{y}_{k\alpha} &= v_{k\alpha}(x) \\ x_{k\alpha} &= \frac{\exp(y_{k\alpha})}{\sum_\beta \exp(y_{k\beta})} \end{aligned} \tag{XL}$$

Trajectories of (XL) satisfy the replicator dynamics

Ex. 2 The quadratic penalty $h(x) = \frac{1}{2}\sum_\beta x_\beta^2$ leads to *projected reinforcement learning*:

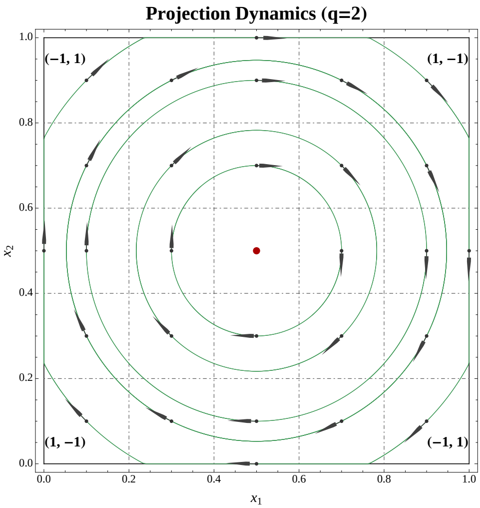
$$\begin{aligned} \dot{y}_k &= v_k(x) \\ x &= \operatorname{proj}_{\mathcal{X}} y \end{aligned} \tag{PL}$$

Closely related to the *projection dynamics* of Friedman (1991):

$$\dot{x}_{k\alpha} = \begin{cases} v_{k\alpha}(x) - |\operatorname{supp}(x_k)|^{-1} \sum_{\beta\in\operatorname{supp}(x_k)} v_{k\beta}(x) & \text{if } \alpha \in \operatorname{supp}(x_k) \\ 0 & \text{otherwise} \end{cases} \tag{PD}$$

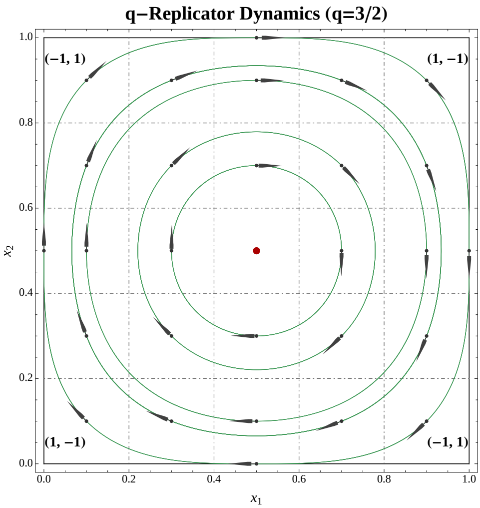
The x-orbits of (PL) satisfy (PD) on an open dense set of times (M & Sandholm, 2015).

Example portraits



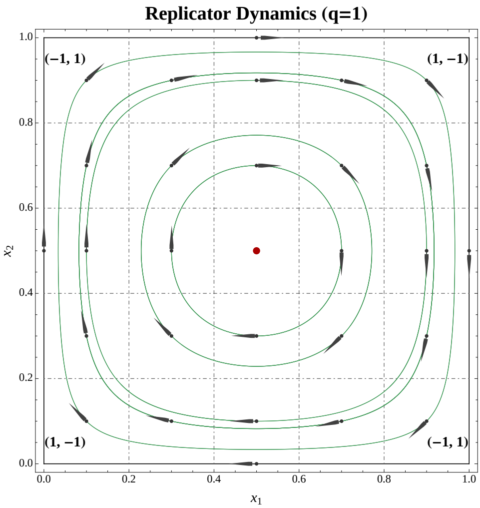
$$h(x) = \frac{1}{2} \sum_\beta x_\beta^2$$

Example portraits



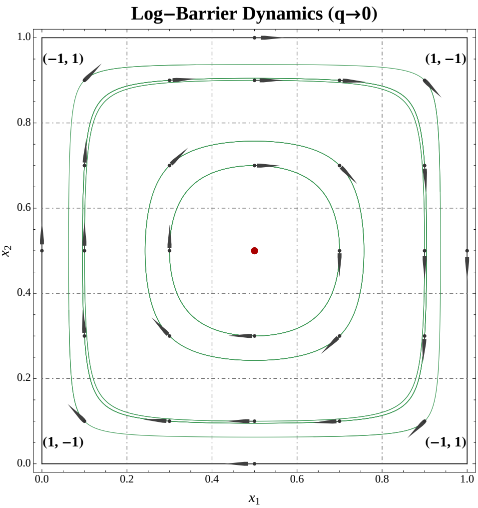
$$h(x) = \frac{4}{3} \sum_\beta x_\beta^{3/2}$$

Example portraits



$$h(x) = \sum_\beta x_\beta \log x_\beta$$

Example portraits



$$h(x) = -\sum_\beta \log x_\beta$$

Extinction of Dominated Strategies

Recall:

- p_k is *dominated* by p'_k if $\langle v_k(x) | p_k - p'_k \rangle < 0$ for all $x \in \mathcal{X}$.
- A strategy $p_k \in \mathcal{X}_k$ *becomes extinct* along $x(t)$ if

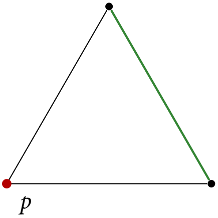
$$\min\{x_{k\alpha}(t) : \alpha \in \operatorname{supp}(p_k)\} \to 0 \quad \text{as } t \to \infty$$

Extinction of Dominated Strategies

Recall:

- p_k is *dominated* by p'_k if $\langle v_k(x) | p_k - p'_k \rangle < 0$ for all $x \in \mathcal{X}$.
- A strategy $p_k \in \mathcal{X}_k$ *becomes extinct* along $x(t)$ if

$$\min\{x_{k\alpha}(t) : \alpha \in \operatorname{supp}(p_k)\} \to 0 \quad \text{as } t \to \infty$$

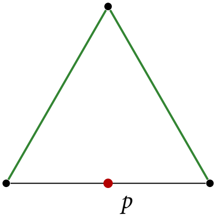


Extinction of Dominated Strategies

Recall:

- p_k is *dominated* by p'_k if $\langle v_k(x) | p_k - p'_k \rangle < 0$ for all $x \in \mathcal{X}$.
- A strategy $p_k \in \mathcal{X}_k$ *becomes extinct* along $x(t)$ if

$$\min\{x_{k\alpha}(t) : \alpha \in \mathrm{supp}(p_k)\} \to 0 \quad \text{as } t \to \infty$$

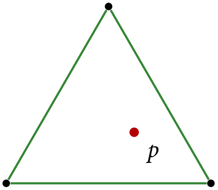


Extinction of Dominated Strategies

Recall:

- p_k is *dominated* by p'_k if $\langle v_k(x) | p_k - p'_k \rangle < 0$ for all $x \in \mathcal{X}$.
- A strategy $p_k \in \mathcal{X}_k$ *becomes extinct* along $x(t)$ if

$$\min\{x_{k\alpha}(t) : \alpha \in \operatorname{supp}(p_k)\} \to 0 \quad \text{as } t \to \infty$$



Theorem (M & Sandholm, 2015)

Dominated strategies become extinct under the reinforcement learning dynamics (RL).

Stability and convergence analysis

Recall:

- x^* is a *Nash equilibrium* iff $u_k(x^*) \geq u_k(x_k; x^*_{-k})$ for all $x_k \in \mathcal{X}_k$, $k \in \mathcal{N}$.
- A Nash equilibrium is *strict* if the above inequality is strict for all $x_k \neq x^*_k$.

Stability and convergence analysis

Recall:

- x^* is a *Nash equilibrium* iff $u_k(x^*) \geq u_k(x_k; x^*_{-k})$ for all $x_k \in \mathcal{X}_k$, $k \in \mathcal{N}$.
- A Nash equilibrium is *strict* if the above inequality is strict for all $x_k \neq x^*_k$.

Theorem (M & Sandholm '15)

Let $x(t) = Q(y(t))$ *be an orbit of* (RL).

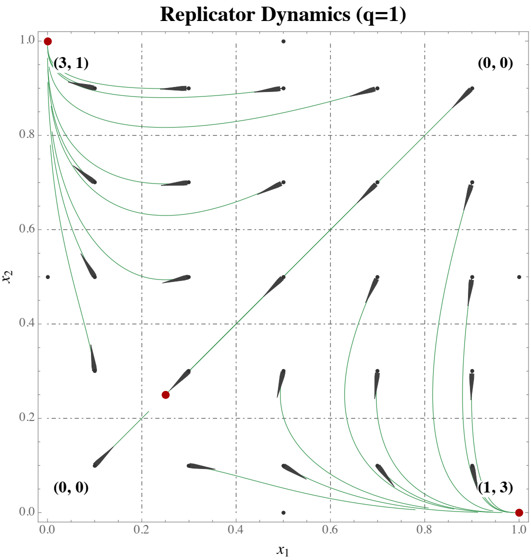
I. *If* $x(t) \to x^*$, *then* x^* *is a Nash equilibrium.*

II. x^* *is stable and attracting iff it is a strict Nash equilibrium.*

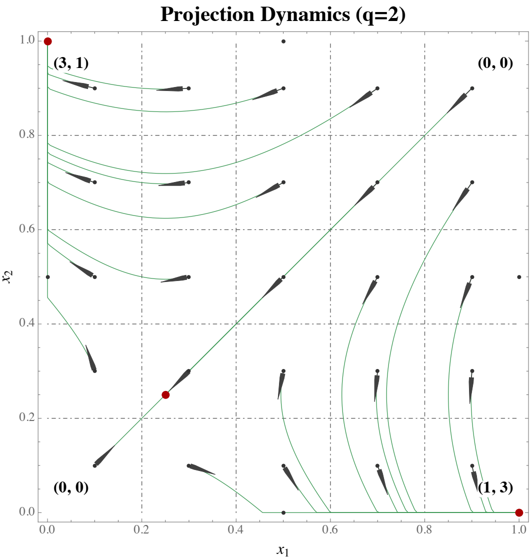
III. $x(t)$ *converges to Nash equilibrium in potential games.*

Special case: EGT "folk theorem" for the replicator dynamics

Convergence to Equilibrium



Convergence to Equilibrium



Outline

The model

Noisy payoff observations lead to the stochastically perturbed learning model

$$\begin{aligned} dY_k &= v_k(X)\,dt + dZ_k \\ X_k &= Q_k(\eta_k Y_k) \end{aligned} \tag{SRL}$$

where:

- the *noise process* Z_k is an Itô martingale (think Brownian motion) with covariance

  $$dZ_{k\alpha} \cdot dZ_{\ell\beta} = \Sigma_{\alpha\beta}\,dt$$

  (noise possibly *state-dependent* and/or *correlated* across players and strategies)
- $\eta_k \equiv \eta_k(t)$ is a (possibly variable) *learning parameter*, introduced for flexibility
- the rest, as before

Assumptions for the noise (Z) and the learning parameter (η)

- $\sup_t \|\Sigma(t)\| < \infty$
- $\eta(t)$ smooth, nonincreasing, and $\eta(t) = \omega(t)$ (i.e. $\lim_{t\to\infty} t\eta(t) = \infty$)

Evolution of mixed strategies

How do mixed strategies evolve under (SRL)?

Proposition

Suppose that the penalty function of player k is of the form $h_k(x_k) = \sum_\alpha \theta_k(x_{k\alpha})$ and Z_k is a Wiener process. Then, $X(t)$ locally follows the stochastic differential equation

$$
\begin{aligned}
dX_{k\alpha} = {} & \frac{\eta_k}{\theta''_{k\alpha}} \left[v_{k\alpha} - \Theta''_k \textstyle\sum_\beta v_{k\beta}/\theta''_{k\beta} \right] dt \\
& + \frac{\eta_k}{\theta''_{k\alpha}} \left[\sigma_{k\alpha}\, dW_{k\alpha} - \Theta''_k \textstyle\sum_\beta \sigma_{k\beta}/\theta''_{k\beta}\, dW_{k\beta} \right] \\
& + \frac{\dot\eta_k}{\eta_k} \frac{1}{\theta''_{k\alpha}} \left[\theta'_{k\alpha} - \Theta''_k \textstyle\sum_\beta \theta'_{k\beta}/\theta''_{k\beta} \right] dt \\
& - \frac{1}{2} \frac{1}{\theta''_{k\alpha}} \left[\theta'''_{k\alpha} U^2_{k\alpha} - \Theta''_k \textstyle\sum_\beta \theta'''_{k\beta}/\theta''_{k\beta}\, U^2_{k\beta} \right] dt,
\end{aligned}
$$

where:

a) $\Theta''_k = \left(\sum_\beta 1/\theta''_{k\beta} \right)^{-1}$,

b) $U^2_{k\alpha} = \left(\dfrac{\eta_k}{\theta''_{k\alpha}} \right)^2 \left[\sigma^2_{k\alpha} \left(1 - \Theta''_k/\theta''_{k\alpha}\right)^2 + \sum_{\beta\neq\alpha} \left(\Theta''_k/\theta''_{k\beta}\right)^2 \sigma^2_{k\beta} \right]$.

Evolution of mixed strategies

How do mixed strategies evolve under (SRL)?

Proposition

Suppose that the penalty function of player k is of the form $h_k(x_k) = \sum_\alpha \theta_k(x_{k\alpha})$ and Z_k is a Wiener process. Then, $X(t)$ locally follows the stochastic differential equation

$$\begin{aligned} dX_{k\alpha} &= \frac{\eta_k}{\theta''_{k\alpha}} \left[v_{k\alpha} - \Theta''_k \textstyle\sum_\beta v_{k\beta}/\theta''_{k\beta} \right] dt \\ &+ \frac{\eta_k}{\theta''_{k\alpha}} \left[\sigma_{k\alpha}\, dW_{k\alpha} - \Theta''_k \textstyle\sum_\beta \sigma_{k\beta}/\theta''_{k\beta}\, dW_{k\beta} \right] \\ &+ \frac{\dot\eta_k}{\eta_k} \frac{1}{\theta''_{k\alpha}} \left[\theta'_{k\alpha} - \Theta''_k \textstyle\sum_\beta \theta'_{k\beta}/\theta''_{k\beta} \right] dt \\ &- \frac{1}{2} \frac{1}{\theta''_{k\alpha}} \left[\theta'''_{k\alpha} U^2_{k\alpha} - \Theta''_k \textstyle\sum_\beta \theta'''_{k\beta}/\theta''_{k\beta}\, U^2_{k\beta} \right] dt, \end{aligned}$$

AWFUL!

where:

a) $\Theta''_k = \left(\sum_\beta 1/\theta''_{k\beta} \right)^{-1}$,

b) $U^2_{k\alpha} = \left(\frac{\eta_k}{\theta''_{k\alpha}} \right)^2 \left[\sigma^2_{k\alpha} \left(1 - \Theta''_k/\theta''_{k\alpha}\right)^2 + \sum_{\beta\neq\alpha} \left(\Theta''_k/\theta''_{k\beta}\right)^2 \sigma^2_{k\beta} \right]$.

Examples

The entropic penalty $h(x) = \sum_\alpha x_\alpha \log x_\alpha$ yields the *stochastic replicator dynamics*

$$\begin{aligned}
dX_{k\alpha} &= \eta_k X_{k\alpha}\left[v_{k\alpha} - \textstyle\sum_\beta^k X_{k\beta}\, v_{k\beta}\right] dt && \text{(drift)}\\
&+ \eta_k X_{k\alpha}\left[\sigma_{k\alpha}\, dW_{k\alpha} - \textstyle\sum_\beta^k \sigma_{k\beta} X_{k\beta}\, dW_{k\beta}\right] && \text{(noise)}\\
&+ \frac{\dot\eta_k}{\eta_k} X_{k\alpha}\left[\log X_{k\alpha} - \textstyle\sum_\beta^k X_{k\beta} \log X_{k\beta}\right] dt && \text{(due to } \dot\eta\text{)}\\
&+ \frac{1}{2} X_{k\alpha}\left[\sigma_{k\alpha}^2 (1 - 2X_{k\alpha}) - \textstyle\sum_\beta^k \sigma_{k\beta}^2 X_{k\beta}\, (1 - 2X_{k\beta})\right] dt. && \text{(Itô)}
\end{aligned}$$

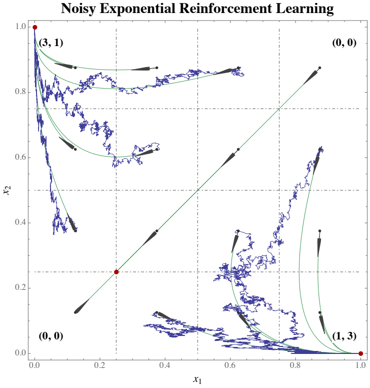
Examples

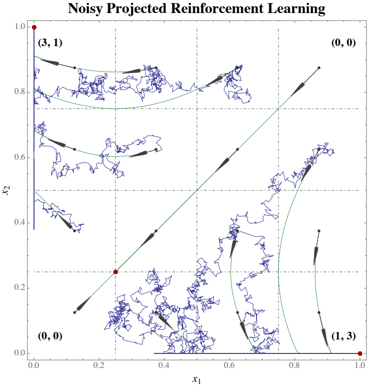
The quadratic penalty $h(x) = \frac{1}{2} \sum_\alpha x_\alpha^2$ yields the *stochastic projection dynamics*

$$\begin{aligned} dX_{k\alpha} &= \Big[v_{k\alpha} - |\mathrm{supp}(X_k)|^{-1} \textstyle\sum_{\beta\in\mathrm{supp}(X_k)} v_{k\beta}\Big]\,dt && \text{(drift)} \\ &+ \Big[\sigma_{k\alpha}\,dW_{k\alpha} - |\mathrm{supp}(X_k)|^{-1} \textstyle\sum_{\beta\in\mathrm{supp}(X_k)} \sigma_{k\beta}\,dW_{k\beta}\Big] && \text{(noise)} \\ &+ \frac{\dot\eta_k}{\eta_k}\Big[X_{k\alpha} - |\mathrm{supp}(X_k)|^{-1}\Big]\,dt. && \text{(due to } \dot\eta\text{)} \end{aligned}$$

NB: There is no Itô correction, but $X(t)$ follows this SDE only locally

Examples





Evolution of play under (SRL) with logit and projection choice maps ($\sigma = 1$)

Consistency and regret

(XL) leads to no regret (Sorin, 2009); in fact, so does (RL) (Kwon & M, 2014). *Is this still true in the presence of noise?*

Consistency and regret

(XL) leads to no regret (Sorin, 2009); in fact, so does (RL) (Kwon & M, 2014). *Is this still true in the presence of noise?*

Yes, provided that the learning parameter $\eta(t)$ tends to zero.

Theorem (Bravo & M, 2015)

If a player runs (SRL) *with* $\eta(t)$ *such that* $\lim_{t\to\infty} \eta(t) = 0$*, then*

$$\operatorname{Reg}(t) \leq \frac{\Omega}{\eta(t)} + \sigma_{\max}^2 \frac{|\mathcal{A}|}{2K} \int_0^t \eta(s)\, ds + \mathcal{O}\left(\sigma_{\max}\sqrt{t \log\log t}\right) \quad \textit{(a.s.)},$$

where Ω and K are constants related to the player's penalty function.

Corollary

If $\eta(t) \sim t^{-\gamma}$*, optimal regret bound obtained for* $\gamma = 1/2$ *and is of order* $\mathcal{O}(\sqrt{t \log\log t})$*; subleading term is* $2\sigma_{\max}\sqrt{\frac{\Omega|\mathcal{A}|}{2K} t}$.

Proof

Sketch of proof.

- Introduce the (primal-dual) *Fenchel coupling*

$$F(x,y) = h(x) + h^*(y) - \langle y|x\rangle$$

- Fix some test strategy $p \in \mathcal{X}$ and consider the rate-adjusted coupling

$$H(t) = \frac{1}{\eta(t)} F(p, \eta(t)Y(t))$$

- Use Itô's lemma to calculate $dH(t)$
- Bound each of the resulting terms (iterated logarithm for the noise, strong convexity for the Itô correction, etc.)
- Maximize over all $p \in \mathcal{X}$ to obtain bound on the regret. □

Extinction of dominated strategies

Are dominated strategies eliminated under (SRL)?

Extinction of dominated strategies

Are dominated strategies eliminated under (SRL)?

Yes, with no vanishing parameter assumptions on $\eta(t)$

Theorem (Bravo & M, 2015)

If $p_k \in \mathcal{X}_k$ is dominated (even iteratively), then it becomes extinct along $X(t)$ almost surely.

Extinction of dominated strategies

Are dominated strategies eliminated under (SRL)?

Yes, with no vanishing parameter assumptions on $\eta(t)$

Theorem (Bravo & M, 2015)

If $p_k \in \mathcal{X}_k$ is dominated (even iteratively), then it becomes extinct along $X(t)$ almost surely.

Extinction rate of a pure dominated strategy $\alpha \in \mathcal{A}_k$:

- If η_k is constant, $h_k(x_k) = \sum_\beta \theta(x_{k\beta})$ and $\tau_\delta = \inf\{t > 0 : X_{k\alpha}(t) < \delta\}$, then

$$\mathbb{E}[\tau_\delta] \leq \frac{C_k - \theta_k'(\delta)}{\eta_k m_k} \quad \text{for some } C_k > 0,\ m_k > 0$$

- If θ_k is non-steep, *dominated strategies become extinct in finite time* (a.s.)

Stability and convergence properties

What is the dynamics' long-term behavior in regards to Nash equilibria?

Stability and convergence properties

What is the dynamics' long-term behavior in regards to Nash equilibria?

Theorem

Let $x^* \in \mathcal{X}$. Then:

- *If a trajectory $X(t)$ converges to x^* with positive probability, x^* is a Nash equilibrium.*
- *If x^* is a strict Nash equilibrium, it is stochastically stable and attracting: for all $\varepsilon > 0$ and for every neighborhood U_0 of x^*, there exists a neighborhood $U \subseteq U_0$ of x^* such that*

$$\mathbb{P}\Big(X(t) \in U_0 \text{ for all } t \geq 0 \text{ and } \lim_{t\to\infty} X(t) = x^*\Big) \geq 1 - \varepsilon.$$

NB: no vanishing parameter assumptions on $\eta(t)$

Long-term time averages

In zero-sum games, the dynamics do not converge to a Nash equilibrium, but their time-averages do (Hofbauer et al., 2009; M & Sandholm, 2015). Is this still true for (SRL)?

Long-term time averages

In zero-sum games, the dynamics do not converge to a Nash equilibrium, but their time-averages do (Hofbauer et al., 2009; M & Sandholm, 2015). Is this still true for (SRL)?

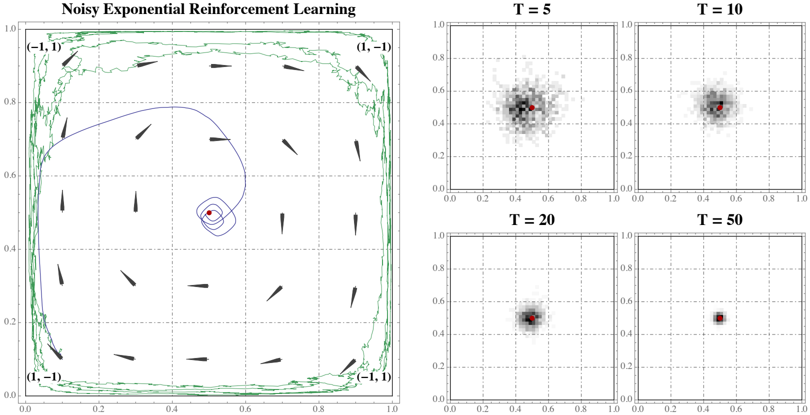
Yes, provided that the learning parameter $\eta(t)$ tends to zero.

Theorem (Bravo & M, 2015)

Let $\mathcal{G}$ be a zero-sum 2-player game with an interior equilibrium. If both players run (SRL) *with vanishing learning parameters* ($\eta_k(t) \to 0$)*, the time averages $\bar{X}(t) = t^{-1} \int_0^t X(s)\, ds$ converge to the Nash set of $\mathcal{G}$.*

(Corollary of more general result linking time averages of (SRL) to the best-response dynamics)

Time averages



(a) A sample trajectory and its time average.

(b) Distribution of time averages at time T.

Concluding remarks

- Dichotomy between "converging to a face" (undom. strategies, strict equilibria) and "average" results (regret, time-averages, ...): constant η better for the former, vanishing η better for the latter
- Itô's formula introduces second-order terms: same control trade-offs as in discrete time
- Some results extend to more general games (e.g. continuous action sets); others trickier
- Possible to handle more intense noise processes (semimartingale noise, fractional Brownian motion), but results different
- Other directions???