ALGORITHMIC GAME THEORY

FROM MULTI-AGENT OPTIMIZATION TO ONLINE LEARNING

Roberto Cominetti[1] Panayotis Mertikopoulos[2 3]

[1]Universidad Adolfo Ibáñez

[2]French National Center for Scientific Research (CNRS)

[3]Criteo AI Lab

Journées SMAI-MODE – September 10–11, 2020

ALGORITHMIC GAME THEORY

LEC. 5: LEARNING IN FINITE GAMES AND MULTI-ARMED BANDITS

Roberto Cominetti[1] Panayotis Mertikopoulos[2 3]

[1]Universidad Adolfo Ibáñez

[2]French National Center for Scientific Research (CNRS)

[3]Criteo AI Lab

Journées SMAI-MODE – September 10–11, 2020

Outline

Overview

Learning in finite games

- **Frequencies** (population shares) $\rightsquigarrow$ **Choice probabilities** (mixed strategies)
- **Dynamics** (continuous time) $\rightsquigarrow$ **Algorithms** (discrete time)
- Information available to the players:
  - **Perfect payoff vector**
  - **Noisy payoff vector**
  - **Bandit (only rewards)**

- **Big picture:** Focus on concepts + selected deep dives
- **Multi-agent** (game-theoretic) v. **online** ("playing against anything")
- **Notation:** losses ("ℓ") $\leftrightsquigarrow$ utilities ("u"); actions $\leftrightsquigarrow$ pure strategies; etc.

Learning with a finite number of actions

Online decision-making with **mixed strategies**

---

repeat

At each epoch $t \geq 0$

Choose **mixed strategy** $x_t \in \mathcal{X} := \Delta(\mathcal{A})$

Encounter **payoff vector** $v_t \in \mathbb{R}^{\mathcal{A}}$ [depends on context]

Get **mean payoff** $u_t(x_t) = \langle v_t, x_t \rangle$

Receive **feedback** [depends on context]

until end

Learning with a finite number of actions

Online decision-making with **mixed strategies**

---

repeat

At each epoch $t \geq 0$

Choose **mixed strategy** $x_t \in \mathcal{X} := \Delta(\mathcal{A})$

Encounter **payoff vector** $v_t \in \mathbb{R}^{\mathcal{A}}$ [depends on context]

Get **mean payoff** $u_t(x_t) = \langle v_t, x_t \rangle$

Receive **feedback** [depends on context]

until end

Key considerations

- **Time:** continuous or discrete?
- **Players:** ~~continuous~~ discrete
- **Actions:** ~~continuous~~ discrete
- **Payoffs:** determined by other players or "Nature"?
- **Feedback:** full info? payoff-based?

Online v. multi-agent learning

How are payoffs generated?

Online v. multi-agent learning

How are payoffs generated?

- Online viewpoint
  - Single, focal agent
  - Different payoff function encountered at each stage
  - Agnostic: no assumptions on mechanism generating u_t (dispassionate Nature)

Online v. multi-agent learning

How are payoffs generated?

- **Online viewpoint**
  - Single, focal agent
  - Different payoff function encountered at each stage
  - **Agnostic:** no assumptions on mechanism generating u_t (dispassionate Nature)
- **Multi-agent viewpoint**
  - Several agents
  - Individual payoff functions depend on actions of other agents
  - **Game-theoretic:** underlying mechanism is a (finite) game

Online v. multi-agent learning

How are payoffs generated?

- **Online viewpoint**
  - Single, focal agent
  - Different payoff function encountered at each stage
  - **Agnostic:** no assumptions on mechanism generating u_t (dispassionate Nature)
- **Multi-agent viewpoint**
  - Several agents
  - Individual payoff functions depend on actions of other agents
  - **Game-theoretic:** underlying mechanism is a (finite) game

What is the interplay between online and multi-agent learning?

Outline

Online viewpoint: regret minimization

The most widely used online performance measure is the agent's regret

$$u_t(x) - u_t(x_t)$$

Online viewpoint: regret minimization

The most widely used online performance measure is the agent's regret

$$\int_0^T [u_t(x) - u_t(x_t)]\, dt$$

Online viewpoint: regret minimization

The most widely used online performance measure is the agent's regret

$$\max_{x \in \mathcal{X}} \int_0^T [u_t(x) - u_t(x_t)]\, dt$$

Online viewpoint: regret minimization

The most widely used online performance measure is the agent's regret

$$\mathrm{Reg}(T) = \max_{x \in \mathcal{X}} \int_0^T \left[u_t(x) - u_t(x_t) \right] dt = \max_{x \in \mathcal{X}} \int_0^T \langle v_t, x - x_t \rangle \, dt$$

Online viewpoint: regret minimization

The most widely used online performance measure is the agent's **regret**

$$\mathrm{Reg}(T) = \max_{x \in \mathcal{X}} \int_0^T \left[u_t(x) - u_t(x_t) \right] dt = \max_{x \in \mathcal{X}} \int_0^T \langle v_t, x - x_t \rangle \, dt$$

No regret: $\mathrm{Reg}(T) = o(T)$ [the smaller the better]

"The chosen policy is as good as the best fixed strategy in hindsight."

Online viewpoint: regret minimization

The most widely used online performance measure is the agent's **regret**

$$\mathrm{Reg}(T) = \max_{x \in \mathcal{X}} \int_0^T [u_t(x) - u_t(x_t)]\, dt = \max_{x \in \mathcal{X}} \int_0^T \langle v_t, x - x_t \rangle\, dt$$

No regret: $\mathrm{Reg}(T) = o(T)$ [the smaller the better]

"The chosen policy is as good as the best fixed strategy in hindsight."

Prolific literature:

- Economics [Hannan; Fudenberg & Levine; Hart & Mas-Colell…]
- Mathematics [Robinson; Blackwell; Hofbauer; Sorin…]
- Computer science [Littlestone & Warmuth; Vovk; Cesa-Bianchi & Lugosi …]

Learning with exponential weights

The "exponential weights" dynamics

$$\dot{y}_t = v_t \qquad x_t = \Lambda(y_t) \tag{EWD}$$

where Λ is the logit map

$$\Lambda(y) = \frac{(\exp(y_a))_{a\in\mathcal{A}}}{\sum_{a\in\mathcal{A}} \exp(y_a)} \quad \text{for all } y \in \mathbb{R}^{\mathcal{A}}$$

Learning with exponential weights

The "exponential weights" dynamics

$$\dot{y}_t = v_t \qquad x_t = \Lambda(y_t) \tag{EWD}$$

where Λ is the logit map

$$\Lambda(y) = \frac{(\exp(y_a))_{a\in\mathcal{A}}}{\sum_{a\in\mathcal{A}} \exp(y_a)} \quad \text{for all } y \in \mathbb{R}^{\mathcal{A}}$$

- KL divergence relative to a target strategy $x \in \mathcal{X}$

$$D_t := D_{\mathrm{KL}}(x, x_t) = \sum_{a\in\mathcal{A}} x_a \log \frac{x_a}{x_{a,t}}$$

Learning with exponential weights

The "exponential weights" dynamics

$$\dot{y}_t = v_t \qquad x_t = \Lambda(y_t) \tag{EWD}$$

where Λ is the logit map

$$\Lambda(y) = \frac{(\exp(y_a))_{a\in\mathcal{A}}}{\sum_{a\in\mathcal{A}} \exp(y_a)} \quad \text{for all } y \in \mathbb{R}^{\mathcal{A}}$$

- KL divergence relative to a target strategy $x \in \mathcal{X}$

$$D_t := D_{\mathrm{KL}}(x, x_t) = \sum_{a\in\mathcal{A}} x_a \log \frac{x_a}{x_{a,t}}$$

- Evolution over time

$$\dot{D}_t = \cdots = \langle v_t, x_t - x \rangle = u_t(x_t) - u_t(x)$$

Learning with exponential weights

The "exponential weights" dynamics

$$\dot{y}_t = v_t \qquad x_t = \Lambda(y_t) \tag{EWD}$$

where Λ is the logit map

$$\Lambda(y) = \frac{(\exp(y_a))_{a\in\mathcal{A}}}{\sum_{a\in\mathcal{A}} \exp(y_a)} \quad \text{for all } y \in \mathbb{R}^{\mathcal{A}}$$

- KL divergence relative to a target strategy $x \in \mathcal{X}$

$$D_t := D_{\mathrm{KL}}(x, x_t) = \sum_{a\in\mathcal{A}} x_a \log \frac{x_a}{x_{a,t}}$$

- Evolution over time

$$\dot{D}_t = \cdots = \langle v_t, x_t - x \rangle = u_t(x_t) - u_t(x)$$

- Integrate:

$$\mathrm{Reg}(T) \le \max_{x\in\mathcal{X}} D_{\mathrm{KL}}(x, x_0) = \mathcal{O}(1)$$

Follow the regularized leader

Are the no-regret propeties of (EWD) a “fluke”?

Follow the regularized leader

Are the no-regret propeties of (EWD) a "fluke"?

- $\Lambda(y)$ approximates the best response correspondence (the "*leader*")

$$y \mapsto \arg\max_{x \in \mathcal{X}} \langle y, x \rangle$$

Follow the regularized leader

Are the no-regret propeties of (EWD) a "fluke"?

- $\Lambda(y)$ approximates the best response correspondence (the "*leader*")

$$y \mapsto \arg\max_{x \in \mathcal{X}} \{\langle y, x \rangle - h(x)\}$$

where $h(x) = \sum_{a \in \mathcal{A}} x_a \log x_a$ is the (negative) entropy of $x \in \mathcal{X}$

Follow the regularized leader

Are the no-regret propeties of (EWD) a "fluke"?

- $\Lambda(y)$ approximates the best response correspondence (the "*leader*")

$$y \mapsto \arg\max_{x \in \mathcal{X}} \{\langle y, x \rangle - h(x)\}$$

where $h(x) = \sum_{a \in \mathcal{A}} x_a \log x_a$ is the (negative) entropy of $x \in \mathcal{X}$

- Regularized best responses

$$Q(y) = \arg\max_{x \in \mathcal{X}} \{\langle y, x \rangle - h(x)\}$$

where $h: \mathcal{X} \to \mathbb{R}$ is a (strictly) convex regularizer function

Follow the regularized leader

Are the no-regret propeties of (EWD) a "fluke"?

- $\Lambda(y)$ approximates the best response correspondence (the "*leader*")

$$y \mapsto \arg\max_{x \in \mathcal{X}} \{\langle y, x \rangle - h(x)\}$$

  where $h(x) = \sum_{a \in \mathcal{A}} x_a \log x_a$ is the (negative) entropy of $x \in \mathcal{X}$

- Regularized best responses

$$Q(y) = \arg\max_{x \in \mathcal{X}} \{\langle y, x \rangle - h(x)\}$$

  where $h: \mathcal{X} \to \mathbb{R}$ is a (strictly) convex regularizer function

- Follow the regularized leader

$$\begin{aligned} \dot{y}_t &= v_t \\ x_t &= Q(y_t) \end{aligned} \qquad \text{(FTRL)}$$

The projection dynamics

Example: Quadratic (Euclidean) regularization

$$h(x) = \frac{1}{2}\sum_a x_a^2$$

The projection dynamics

Example: Quadratic (Euclidean) regularization

$$h(x) = \frac{1}{2}\sum_a x_a^2$$

Choice map $\rightsquigarrow$ closest point projection:

$$\Pi(y) = \arg\max_{x\in\mathcal{X}}\{\langle y, x\rangle - (1/2)\|x\|_2^2\} = \arg\min_{x\in\mathcal{X}}\|y - x\|$$

The projection dynamics

Example: Quadratic (Euclidean) regularization

$$h(x) = \frac{1}{2}\sum_a x_a^2$$

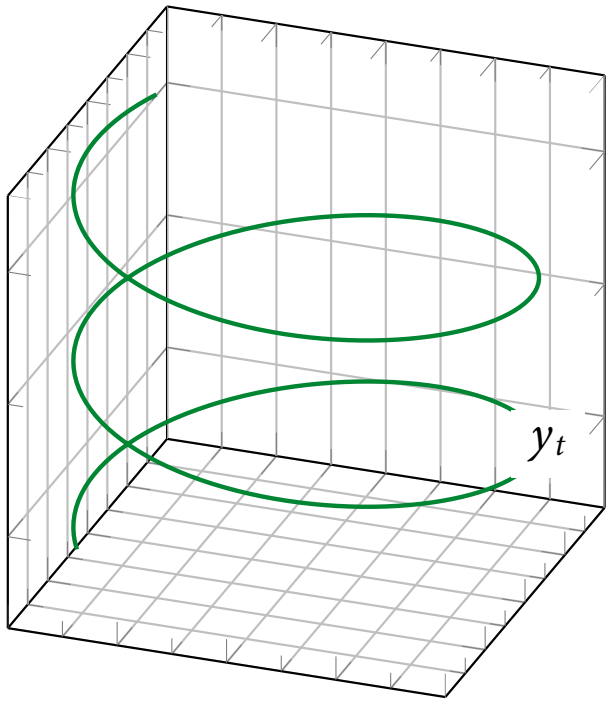
Choice map $\rightsquigarrow$ closest point projection:

$$\Pi(y) = \arg\max_{x\in\mathcal{X}}\{\langle y, x\rangle - (1/2)\|x\|_2^2\} = \arg\min_{x\in\mathcal{X}}\|y - x\|$$

Projection dynamics [M & Sandholm, 2016]

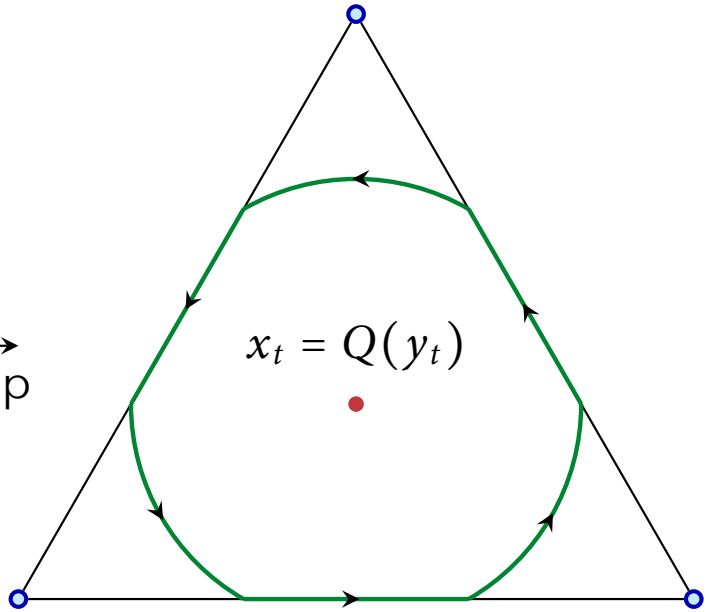
$$\begin{aligned}\dot{y}_t &= v_t\\ x_t &= \Pi(y_t)\end{aligned} \qquad \text{(PL)}$$

In and out of the boundary



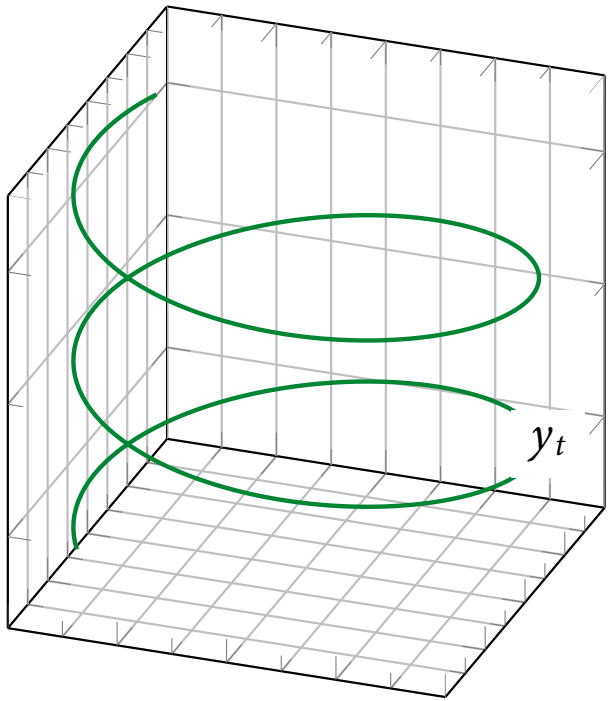
Payoff space

$$\xrightarrow[\text{choice map}]{Q}$$



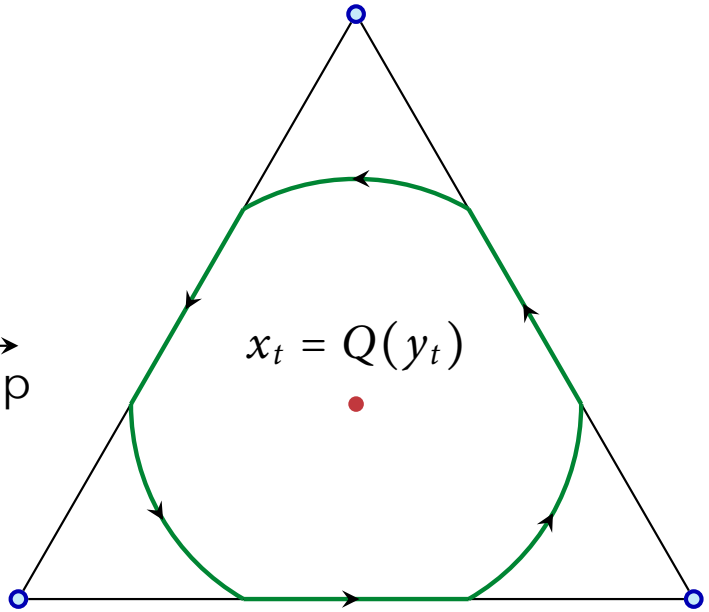
Strategy space

In and out of the boundary



Payoff space

Q
choice map



Strategy space

Key difference with replicator: faces no longer forward invariant

Portraits and examples

The Tsallis-Havrda -Charvát kernel: $h(x) = [q(1-q)]^{-1} \sum_a (x_a - x_a^q)$

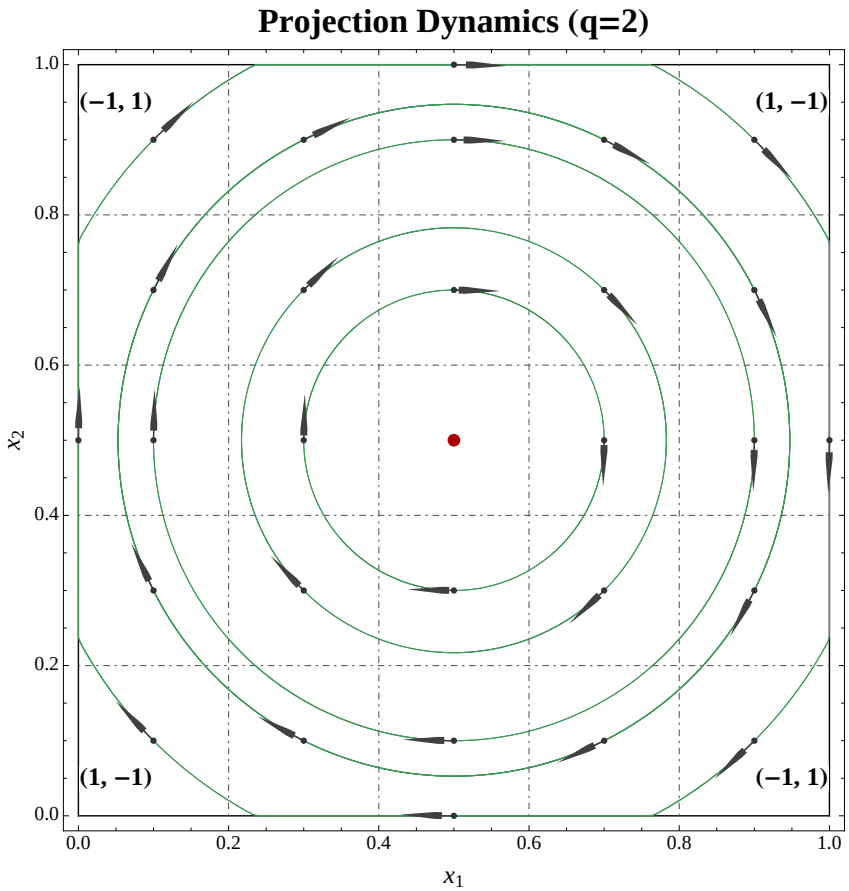


Figure: Phase portraits of (FTRL) in Matching Pennies for diferrent values of $q > 0$

Portraits and examples

The Tsallis-Havrda -Charvát kernel: $h(x) = [q(1-q)]^{-1} \sum_a (x_a - x_a^q)$

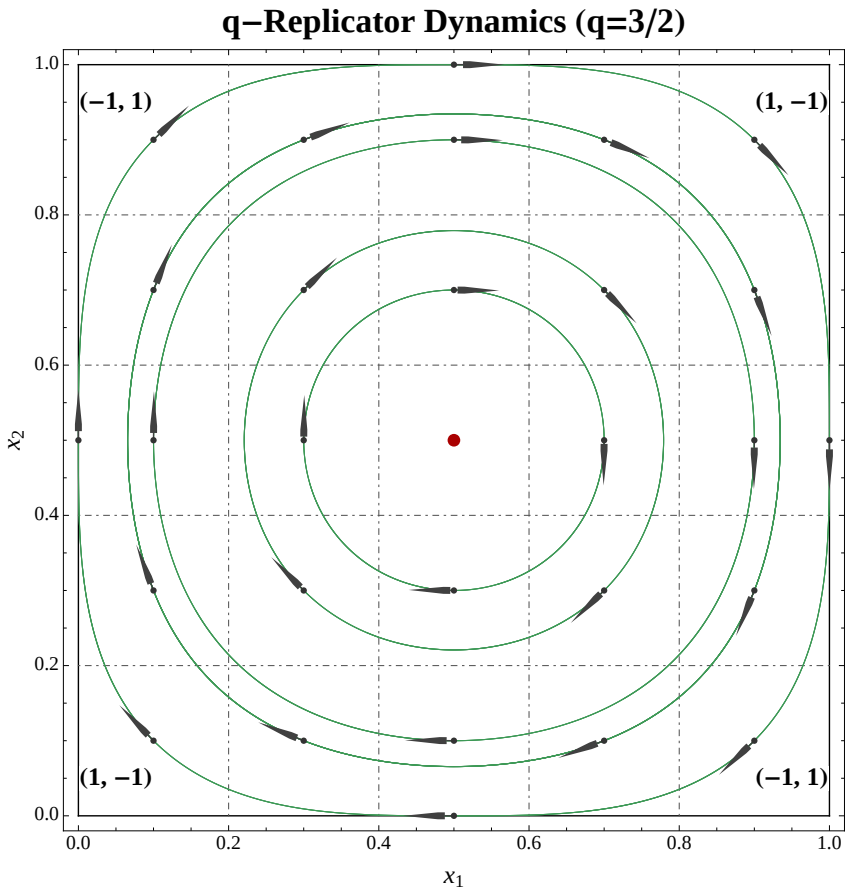


Figure: Phase portraits of (FTRL) in Matching Pennies for diferrent values of $q > 0$

Portraits and examples

The Tsallis-Havrda -Charvát kernel: $h(x) = [q(1-q)]^{-1} \sum_a (x_a - x_a^q)$

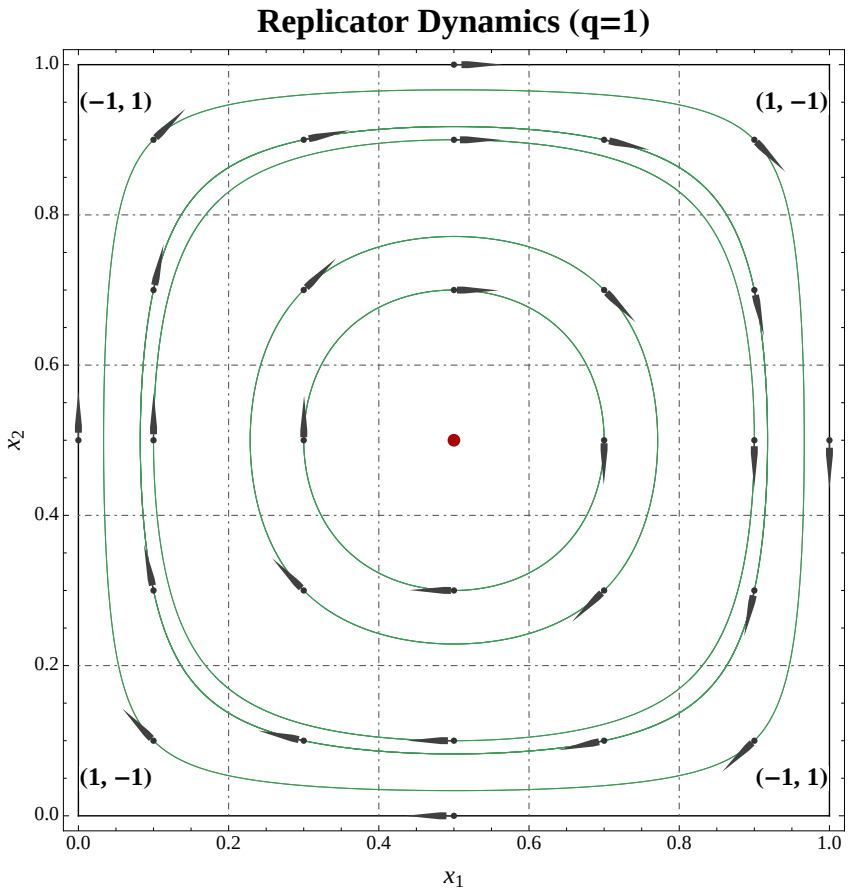


Figure: Phase portraits of (FTRL) in Matching Pennies for diferrent values of $q > 0$

Portraits and examples

The Tsallis-Havrda -Charvát kernel: $h(x) = [q(1-q)]^{-1} \sum_a (x_a - x_a^q)$

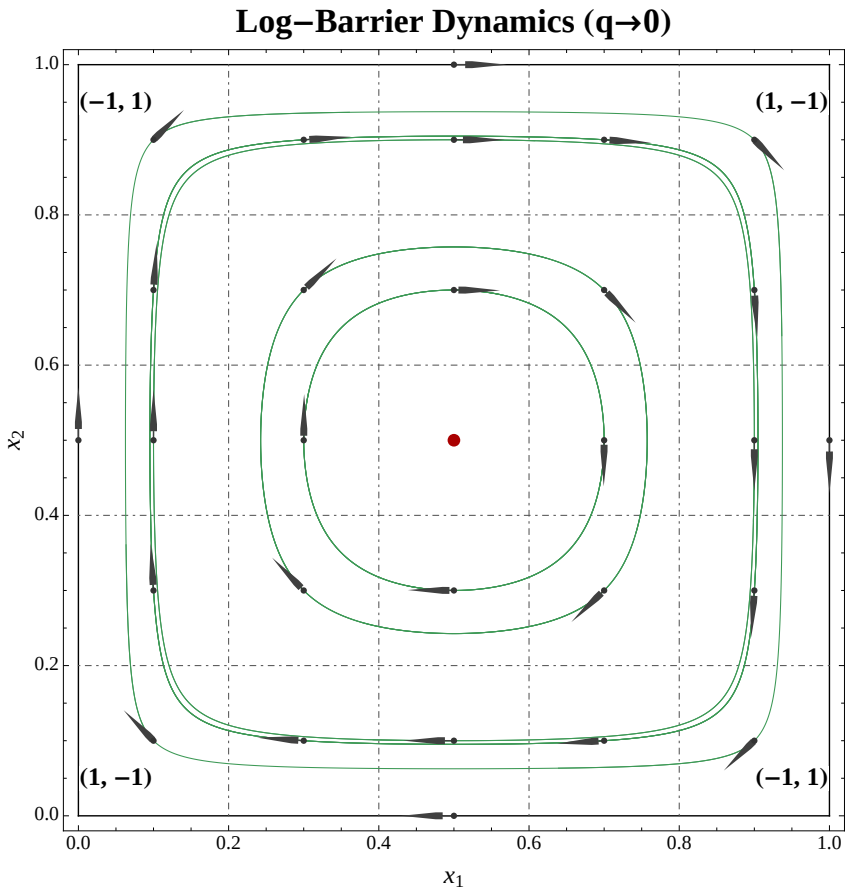


Figure: Phase portraits of (FTRL) in Matching Pennies for diferrent values of $q > 0$

No regret under FTRL

Do the no-regret properties of (EWD) extend to (FTRL)?

No regret under FTRL

Do the no-regret properties of (EWD) extend to (FTRL)?

- Require primal-dual analogue of KL divergence

No regret under FTRL

Do the no-regret properties of (EWD) extend to (FTRL)?

- Require primal-dual analogue of KL divergence
- **Fenchel coupling** [M & Sandholm, 2016; M & Zhou, 2019]

$$F_t = h(x) + h^*(y_t) - \langle y_t, x \rangle$$

where $h^*(y) = \max_{x \in \mathcal{X}} \{\langle y, x \rangle - h(x)\}$ is the convex conjugate of h

No regret under FTRL

Do the no-regret properties of (EWD) extend to (FTRL)?

- Require primal-dual analogue of KL divergence
- **Fenchel coupling** [M & Sandholm, 2016; M & Zhou, 2019]

$$F_t = h(x) + h^*(y_t) - \langle y_t, x \rangle$$

where $h^*(y) = \max_{x \in \mathcal{X}} \{\langle y, x \rangle - h(x)\}$ is the convex conjugate of h

- By Danskin's theorem: $[\nabla h^*(y) = Q(y)]$

$$\dot{F}_t = \langle \dot{y}_t, Q(y_t) \rangle - \langle \dot{y}_t, x \rangle = \langle v_t, x_t - x \rangle$$

No regret under FTRL

Do the no-regret properties of (EWD) extend to (FTRL)?

- Require primal-dual analogue of KL divergence
- **Fenchel coupling** [M & Sandholm, 2016; M & Zhou, 2019]

$$F_t = h(x) + h^*(y_t) - \langle y_t, x \rangle$$

where $h^*(y) = \max_{x \in \mathcal{X}} \{\langle y, x \rangle - h(x)\}$ is the convex conjugate of h

- By Danskin's theorem: $[\nabla h^*(y) = Q(y)]$

$$\dot{F}_t = \langle \dot{y}_t, Q(y_t) \rangle - \langle \dot{y}_t, x \rangle = \langle v_t, x_t - x \rangle$$

Theorem (Kwon & M, 2017)

Under (FTRL)*, the optimizer enjoys the regret bound*

$$\mathrm{Reg}(T) \leq \max_{x \in \mathcal{X}} F(x, y_0) = \mathcal{O}(1)$$

Outline

Multi-agent learning

- **Multiple** agents, individual objectives
- Payoffs determined by actions of **all** agents
- Agents receive payoffs, **adjust actions**, and the process repeats

Multi-agent learning

- **Multiple** agents, individual objectives

  Example: *select a route from home to work*

- Payoffs determined by actions of **all** agents

  Example: *outcome of auction revealed*

- Agents receive payoffs, **adjust actions**, and the process repeats

  Example: *change bid next time*

Multi-agent learning

- **Multiple** agents, individual objectives

  Example: *select a route from home to work*

- Payoffs determined by actions of **all** agents

  Example: *outcome of auction revealed*

- Agents receive payoffs, **adjust actions**, and the process repeats

  Example: *change bid next time*

Does no-regret learning lead to equilibrium?

Finite games

- **Players:** $\mathcal{N} = \{1, \dots, N\}$ [atomic player roles]
- **Actions:** finite action sets $\mathcal{A}_i = \{a_{i,1}, a_{i,2}, \dots\}$ [routes, bids, products,…]
- **Payoffs:** depend on all players' strategies
  - *Action profiles* $(a_i; a_{-i}) := (a_1, \dots, a_i, \dots, a_N) \in \mathcal{A} = \prod_i \mathcal{A}_i$
  - *Mixed strategies*

$$x_{ia_i} = \text{probability that player } i \text{ chooses } a_i \in \mathcal{A}_i$$
$$x_i = (x_{ia_i})_{a_i \in \mathcal{A}_i} \in \mathcal{X}_i := \Delta(\mathcal{A}_i)$$
$$x = (x_1, \dots, x_N) \in \mathcal{X} := \prod_i \mathcal{X}_i$$

  - *Payoff functions*

$$u_i(a_i; a_{-i}) = \text{payoff to player } i \text{ when playing } a_i \text{ against } a_{-i}$$

  - *Mean payoff per strategy*

$$u_{ia_i}(x) := u_i(a_i; x_{-i}) = \sum_{a_{-i}} x_{-i,a_{-i}} u_i(a_i; a_{-i})$$

  - *Payoff vector*

$$v_i(x) = (u_{ia_i}(x))_{a_i \in \mathcal{A}_i}$$

Correlated strategies

Instead of mixing, *correlated strategies* respond to the "state of the world"

$$\chi_a = \chi_{a_1,\ldots,a_N} \in \Delta(\mathcal{A})$$

[NB: $\prod_i \Delta(\mathcal{A}_i) \ll \Delta(\prod_i \mathcal{A}_i)$]

Correlated strategies

Instead of mixing, *correlated strategies* respond to the "state of the world"

$$\chi_a = \chi_{a_1,\ldots,a_N} \in \Delta(\mathcal{A})$$

[NB: $\prod_i \Delta(\mathcal{A}_i) \ll \Delta(\prod_i \mathcal{A}_i)$]

Marginals of χ:

$$x_{ia_i} = \sum_{a_{-i} \in \mathcal{A}_{-i}} \chi_{a_i;a_{-i}}$$

[NB: χ mixed $\iff$ $\chi_a = \prod_i x_{ia_i}$]

Correlated strategies

Instead of mixing, *correlated strategies* respond to the "state of the world"

$$\chi_a = \chi_{a_1,\ldots,a_N} \in \Delta(\mathcal{A})$$

[NB: $\prod_i \Delta(\mathcal{A}_i) \ll \Delta(\prod_i \mathcal{A}_i)$]

Marginals of χ:

$$x_{ia_i} = \sum_{a_{-i} \in \mathcal{A}_{-i}} \chi_{a_i;a_{-i}}$$

[NB: χ mixed $\iff$ $\chi_a = \prod_i x_{ia_i}$]

Correlated equilibrium: [Aumann, 1974, 1987]

$$\sum_{a_{-i} \in \mathcal{A}_{-i}} \chi^*_{a_i;a_{-i}} u_i(a_i; a_{-i}) \geq \sum_{a_{-i} \in \mathcal{A}_{-i}} \chi^*_{a_i;a_{-i}} u_i(a'_i; a_{-i}) \quad \text{for all } a_i, a'_i$$

Correlated strategies

Instead of mixing, *correlated strategies* respond to the "state of the world"

$$\chi_a = \chi_{a_1,\ldots,a_N} \in \Delta(\mathcal{A})$$

[NB: $\prod_i \Delta(\mathcal{A}_i) \ll \Delta(\prod_i \mathcal{A}_i)$]

Marginals of χ:

$$x_{ia_i} = \sum_{a_{-i} \in \mathcal{A}_{-i}} \chi_{a_i;a_{-i}}$$

[NB: χ mixed $\iff$ $\chi_a = \prod_i x_{ia_i}$]

Correlated equilibrium: [Aumann, 1974, 1987]

$$\sum_{a_{-i} \in \mathcal{A}_{-i}} \chi^*_{a_i;a_{-i}} u_i(a_i; a_{-i}) \geq \sum_{a_{-i} \in \mathcal{A}_{-i}} \chi^*_{a_i;a_{-i}} u_i(a'_i; a_{-i}) \quad \text{for all } a_i, a'_i$$

Coarse correlated equilibrium: [Hannan, 1957]

$$\sum_{a_i \in \mathcal{A}_i} \sum_{a_{-i} \in \mathcal{A}_{-i}} \chi^*_{a_i;a_{-i}} u_i(a_i; a_{-i}) \geq \sum_{a_i \in \mathcal{A}_i} \sum_{a_{-i} \in \mathcal{A}_{-i}} \chi^*_{a_i;a_{-i}} u_i(a'_i; a_{-i})$$

No regret and equilibrium

No-regret learning converges to equilibrium!

No regret and equilibrium

Under no-regret learning, empirical frequencies converge to equilibrium …

No regret and equilibrium

Under no-regret learning, empirical frequencies of play converge to coarse correlated equilibrium

No regret and equilibrium

Under no-regret learning, empirical frequencies of play converge to coarse correlated equilibrium ¯_(ツ)_/¯

No regret and equilibrium

Under no-regret learning, empirical frequencies of play converge to coarse correlated equilibrium ¯_(ツ)_/¯

✗ Very weak notion of "convergence"

$\rightsquigarrow$ stray arbitrarily far from equilibrium infinitely often

[Hart and Mas-Colell, 2000, 2003]

No regret and equilibrium

Under no-regret learning, empirical frequencies of play converge to coarse correlated equilibrium ¯_(ツ)_/¯

✗ Very weak notion of "convergence"

⇝ stray arbitrarily far from equilibrium infinitely often

[Hart and Mas-Colell, 2000, 2003]

✗ Very weak notion of "equilibrium"

⇝ assign positive weight only to strictly dominated strategies

[Viossat and Zapechelnyuk, 2013]

No-regret learning and rationality

What is the interplay between online and multi-agent learning?

No-regret learning and rationality

What is the interplay between online and multi-agent learning?

- Do dominated strategies die out under no-regret learning?
- Are Nash equilibria stationary?
- Are they stable? Are they attracting?
- What other behaviors can occur?

Dominated strategies

Suppose $a \in \mathcal{A}$ is *dominated* by $a' \in \mathcal{A}$

- Consistent difference in payoffs/scores:

$$u_a(x) \leq u_{a'}(x) - \varepsilon \quad \text{for some } \varepsilon > 0$$

$$y_{a,t} = \int_0^t u_a(x_\tau)\, d\tau \leq \int_0^t \left[u_{a'}(x_\tau) - \varepsilon\right] d\tau = y_{a',t} - \varepsilon t$$

Dominated strategies

Suppose $a \in \mathcal{A}$ is *dominated* by $a' \in \mathcal{A}$

- Consistent difference in payoffs/scores:

$$u_a(x) \leq u_{a'}(x) - \varepsilon \quad \text{for some } \varepsilon > 0$$

$$y_{a,t} = \int_0^t u_a(x_\tau)\, d\tau \leq \int_0^t \left[u_{a'}(x_\tau) - \varepsilon\right] d\tau = y_{a',t} - \varepsilon t$$

- Translation to choice probabilities not clear

Want: large score difference $y_{a',t} - y_{a,t} \implies x_{a,t} \to 0$ (???)

Dominated strategies

Suppose $a \in \mathcal{A}$ is *dominated* by $a' \in \mathcal{A}$

- Consistent difference in payoffs/scores:

$$u_a(x) \leq u_{a'}(x) - \varepsilon \quad \text{for some } \varepsilon > 0$$

$$y_{a,t} = \int_0^t u_a(x_\tau)\, d\tau \leq \int_0^t \left[u_{a'}(x_\tau) - \varepsilon\right] d\tau = y_{a',t} - \varepsilon t$$

- Translation to choice probabilities not clear

Want: large score difference $y_{a',t} - y_{a,t} \implies x_{a,t} \to 0$ (???)

Theorem (M & Sandholm, 2016)

Under (FTRL)*:*

- $\lim_{t\to\infty} x_{ia_i,t} = 0$ *whenever* a_i *is dominated*
- *If* h *is (sub)differentiable on* $\mathcal{X}$*,* **elimination occurs in finite time**

Stability and convergence

Primal-dual nature of dynamics requires redefinition:

Definition

1. x^* is **stable** if $Q(y_t)$ stays close to x^* when $Q(y_0)$ starts close enough to x^*
2. x^* is **attracting** if $Q(y_t) \to x^*$ whenever $Q(y_0)$ starts close enough to x^*
3. x^* is **asymptotically stable** if it is stable and attracting

Stability and convergence

Primal-dual nature of dynamics requires redefinition:

Definition

1. x^* is **stable** if $Q(y_t)$ stays close to x^* when $Q(y_0)$ starts close enough to x^*
2. x^* is **attracting** if $Q(y_t) \to x^*$ whenever $Q(y_0)$ starts close enough to x^*
3. x^* is **asymptotically stable** if it is stable and attracting

Theorem (M & Sandholm, 2016; Flokas et al., 2020)

I. *If $x_t \to x^*$, then x^* is a Nash equilibrium.*

II. *If $x^* \in \mathcal{X}$ is stable, then x^* is Nash.*

III. *x^* is asymptotically stable if and only if it is a strict Nash equilibrium.*

[Special case: "folk theorem" of EGT]

Non-convergence in zero-sum games

In bilinear zero-sum games:

x^* is full-support equilibrium $\implies$ (FTRL) admits constant of motion

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

Non-convergence in zero-sum games

In bilinear zero-sum games:

$$x^* \text{ is full-support equilibrium} \implies \text{(FTRL) admits constant of motion}$$

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

Theorem (M & Sandholm, 2016; M, Piliouras & Papadimitriou, 2018)

Assume (FTRL) *is run in a bilinear zero-sum game with an interior equilibrium. Then:*

- *The dynamics are* Poincaré recurrent
- *Time-averages* $\bar{x}_t = t^{-1} \int_0^t x_\tau \, d\tau$ converge to Nash equilibrium

Outline

Learning with a finite number of actions

Online decision-making with **mixed strategies**

---

repeat

At each epoch $t = 1, 2, \ldots$

Choose **mixed strategy** $X_t \in \mathcal{X} := \Delta(\mathcal{A})$

Choose **action** $a_t \sim X_t$

Encounter **payoff vector** $v_t \in \mathbb{R}^{\mathcal{A}}$ [depends on context]

Get **payoff** $u_t(a_t) = v_{a_t,t}$

Receive **feedback** [maybe]

until end

Learning with a finite number of actions

Online decision-making with **mixed strategies**

---

repeat

At each epoch $t = 1, 2, \ldots$

Choose **mixed strategy** $X_t \in \mathcal{X} := \Delta(\mathcal{A})$

Choose **action** $a_t \sim X_t$

Encounter **payoff vector** $v_t \in \mathbb{R}^{\mathcal{A}}$ [depends on context]

Get **payoff** $u_t(a_t) = v_{a_t,t}$

Receive **feedback** [maybe]

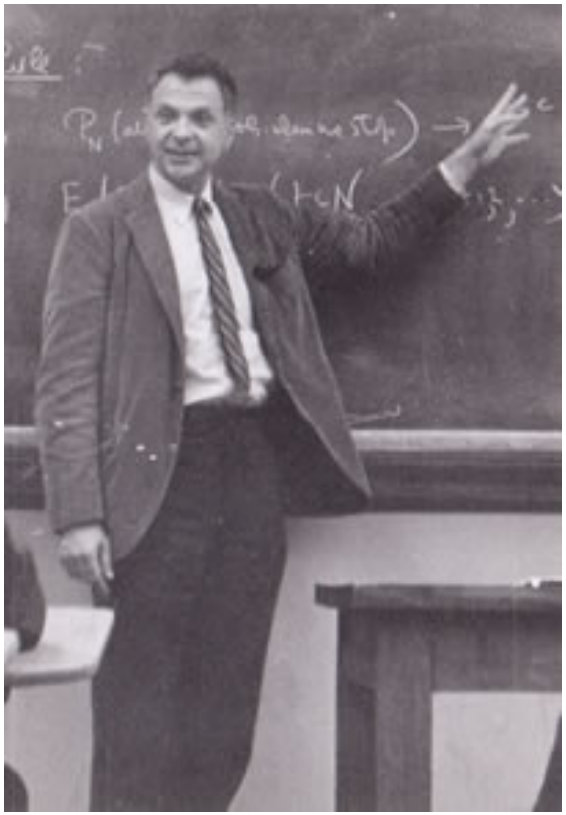
until end

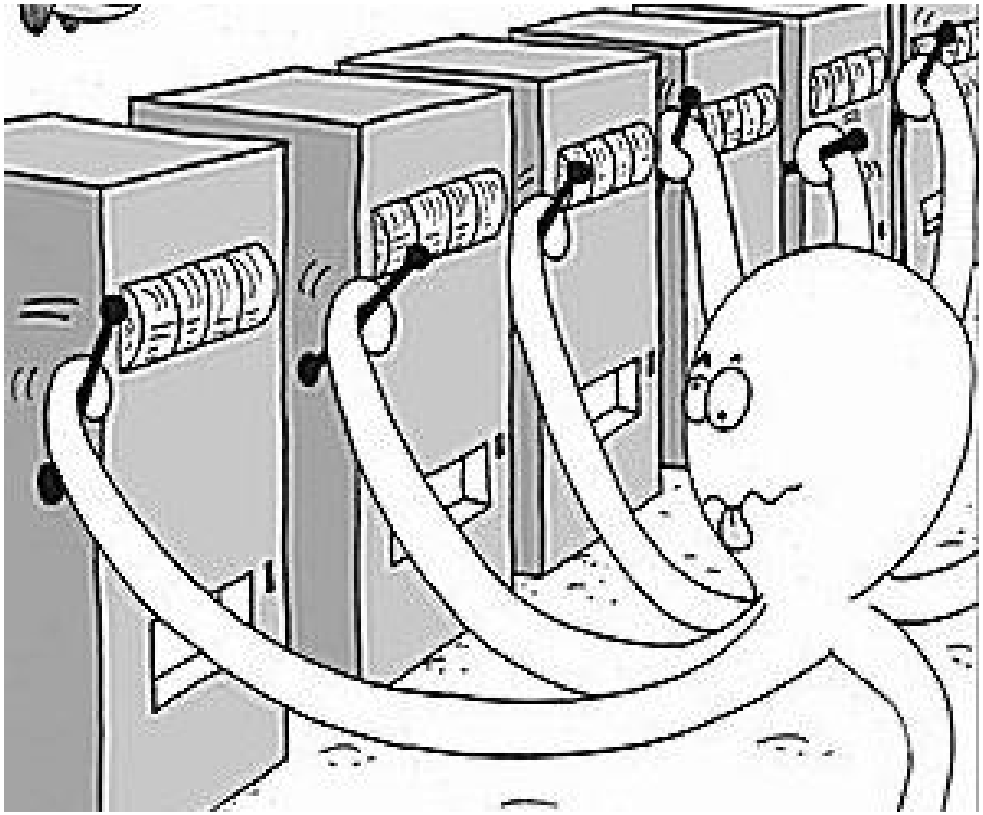
Key considerations

- Time: ~~continuous~~ or **discrete**
- Players: ~~continuous~~ or **discrete**
- Actions: ~~continuous~~ or **discrete**
- **Losses:** determined by other players or "Nature"?
- **Feedback:** full info? payoff-based?

Multi-armed bandits

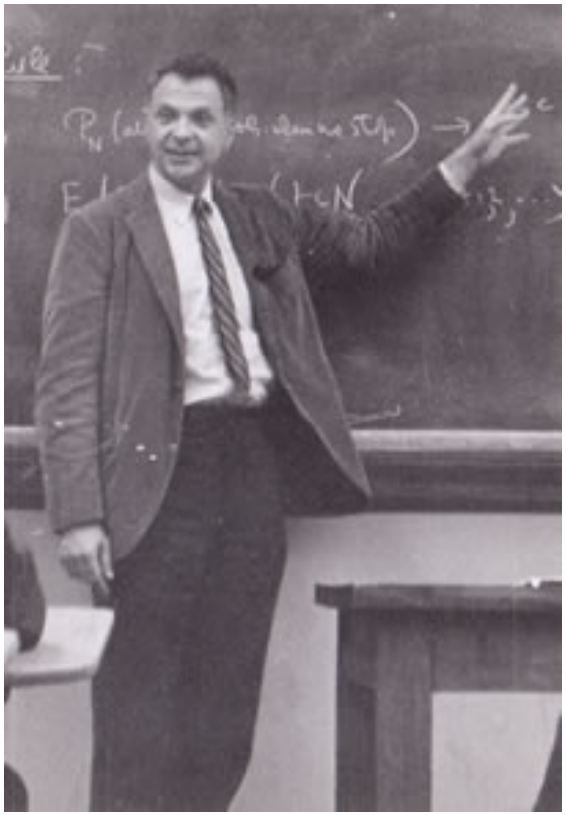
Robbins' multi-armed bandit problem: how to play in a (rigged) casino?

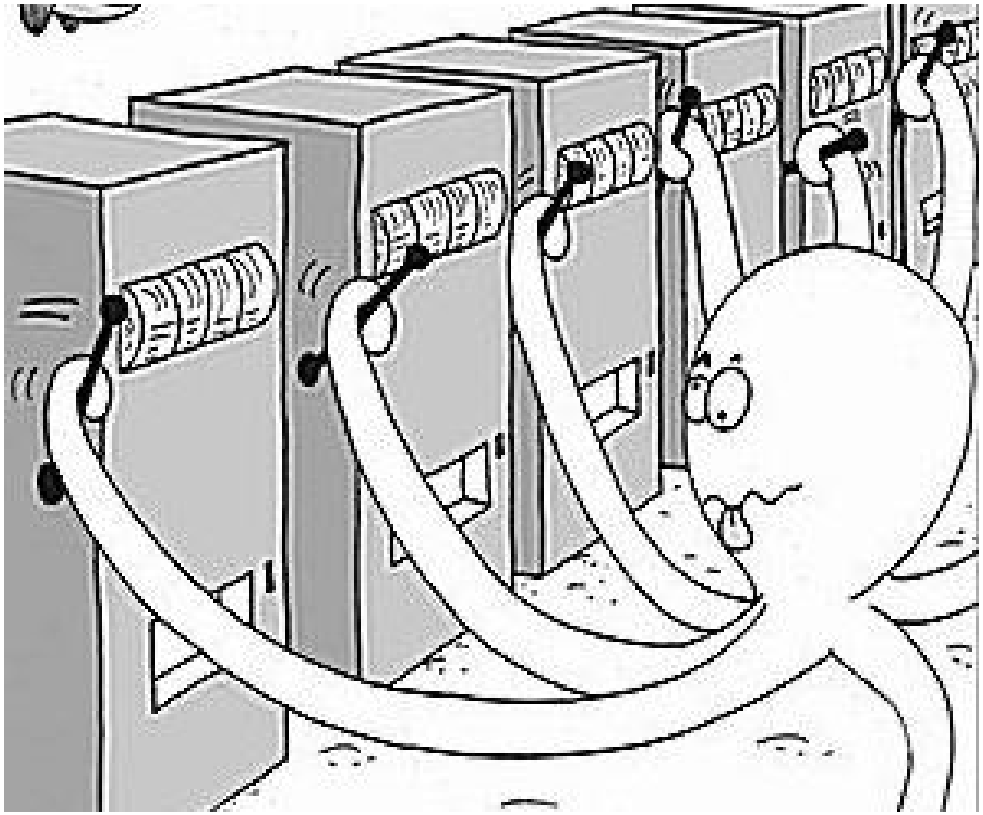




Multi-armed bandits

Robbins' multi-armed bandit problem: **how to play in a (rigged) casino?**





[Lec. 6: What if the arms are players themselves?]

Online viewpoint: regret minimization

The agent's **regret** in discrete time

Realized regret: $$\mathrm{Reg}(T) = \max_{a \in \mathcal{A}} \sum_{t=1}^{T} [u_t(a) - u_t(a_t)]$$

Mean regret: $$\overline{\mathrm{Reg}}(T) = \max_{x \in \mathcal{X}} \sum_{t=1}^{T} [u_t(x) - u_t(X_t)] = \max_{x \in \mathcal{X}} \sum_{t=1}^{T} \langle v_t, x - X_t \rangle$$

Online viewpoint: regret minimization

The agent's **regret** in discrete time

Realized regret: $$\mathrm{Reg}(T) = \max_{a \in \mathcal{A}} \sum_{t=1}^{T} [u_t(a) - u_t(a_t)]$$

Mean regret: $$\overline{\mathrm{Reg}}(T) = \max_{x \in \mathcal{X}} \sum_{t=1}^{T} [u_t(x) - u_t(X_t)] = \max_{x \in \mathcal{X}} \sum_{t=1}^{T} \langle v_t, x - X_t \rangle$$

- **Adversarial framework:** regret guarantees against *any* given sequence v_t
- No distinction between *mean* regret and *pseudo*-regret

[Bubeck and Cesa-Bianchi, 2012]

- **Not here:** stochastic, Markovian, oblivious/non-oblivious,…

[Cesa-Bianchi and Lugosi, 2006]

Feedback

Three types of feedback (from best to worst):

- **Full, exact information**: observe entire payoff vector v_t
- **Full, inexact information**: observe estimate V_t of v_t
- **Partial information / Bandit:** only chosen component $u_t(a_t) = v_{a_t,t}$

Feedback

Three types of feedback (from best to worst):

- **Full, exact information**: observe entire payoff vector v_t
- **Full, inexact information**: observe estimate V_t of v_t
- **Partial information / Bandit:** only chosen component $u_t(a_t) = v_{a_t,t}$

Typically V_t

$$V_t = v_t + U_t + b_t$$

where U_t is **zero-mean** and b_t is the **bias** of V_t

Feedback

Three types of feedback (from best to worst):

- **Full, exact information**: observe entire payoff vector v_t
- **Full, inexact information**: observe estimate V_t of v_t
- **Partial information / Bandit:** only chosen component $u_t(a_t) = v_{a_t,t}$

Typically V_t

$$V_t = v_t + U_t + b_t$$

where U_t is **zero-mean** and b_t is the **bias** of V_t

Assumptions

- **Bias:** $\|b_t\| \leq B_t$ (a.s.)
- **Variance:** $\mathbb{E}[\|U_t\|^2 \mid \mathcal{F}_t] \leq \sigma_t^2$ (a.s.)
- **Second moment:** $\mathbb{E}[\|V_t\|^2 \mid \mathcal{F}_t] \leq M_t^2$ (a.s.)

Follow the regularized leader

Implementing FTRL with full information (exact or inexact):

$$\begin{aligned} Y_{t+1} &= Y_t + \gamma_t V_t \\ X_{t+1} &= Q(Y_{t+1}) \end{aligned} \qquad \text{(FTRL)}$$

where γ_t is a variable step-size parameter

Follow the regularized leader

Implementing FTRL with full information (exact or inexact):

$$\begin{aligned} Y_{t+1} &= Y_t + V_t \\ X_{t+1} &= Q(\eta_{t+1} Y_{t+1}) \end{aligned} \qquad \text{(FTRL)}$$

where η_t is a variable learning rate parameter

Follow the regularized leader

Implementing FTRL with full information (exact or inexact):

$$\begin{aligned} Y_{t+1} &= Y_t + \gamma_t V_t \\ X_{t+1} &= Q(Y_{t+1}) \end{aligned} \qquad \text{(FTRL)}$$

where γ_t is a variable step-size parameter

Technical: Will need Q Lipschitz continuous $\iff$ h is strongly convex

$$h(x') \geq h(x) + \langle \nabla h(x), x' - x \rangle + \frac{K}{2} \|x' - x\|^2$$

Follow the regularized leader

Implementing FTRL with full information (exact or inexact):

$$\begin{aligned} Y_{t+1} &= Y_t + \gamma_t V_t \\ X_{t+1} &= Q(Y_{t+1}) \end{aligned} \qquad \text{(FTRL)}$$

where γ_t is a variable step-size parameter

Technical: Will need Q Lipschitz continuous $\iff$ h is strongly convex

$$h(x') \geq h(x) + \langle \nabla h(x), x' - x \rangle + \frac{K}{2} \|x' - x\|^2$$

Example: Multiplicative / Exponential Weights algorithm

$$\begin{aligned} Y_{t+1} &= Y_t + \gamma_t V_t \\ X_{t+1} &= \frac{(\exp(Y_{a,t+1}))_{a \in \mathcal{A}}}{\sum_{a \in \mathcal{A}} \exp(Y_{a,t+1})} \end{aligned} \qquad \text{(EW)}$$

[Vovk, 1990; Littlestone and Warmuth, 1994; Auer et al., 1995; Freund and Schapire, 1999; Sorin, 2009; Arora et al., 2012]

Regret guarantees of FTRL

Work as in continuous-time case

- Fenchel coupling

$$F_t = h(x) + h^*(Y_t) - \langle Y_t, x \rangle$$

Regret guarantees of FTRL

Work as in continuous-time case

- Fenchel coupling

$$F_t = h(x) + h^*(Y_t) - \langle Y_t, x \rangle$$

- Discrete-time evolution

$$F_{t+1} \leq F_t - \gamma \langle V_t, X_t - x \rangle + \frac{\gamma^2}{2K} \| V_t \|_*^2$$

Regret guarantees of FTRL

Work as in continuous-time case

- Fenchel coupling

$$F_t = h(x) + h^*(Y_t) - \langle Y_t, x \rangle$$

- Discrete-time evolution

$$F_{t+1} \leq F_t - \gamma \langle V_t, X_t - x \rangle + \frac{\gamma^2}{2K} \| V_t \|_*^2$$

- Aggregate/Telescope:

$$\overline{\mathrm{Reg}}(T) = \mathcal{O}\left(\frac{\max h - \min h}{\gamma} + \sum_{t=1}^{T} B_t + \gamma \sum_{t=1}^{T} M_t^2 \right)$$

Regret guarantees of FTRL

Work as in continuous-time case

- Fenchel coupling

$$F_t = h(x) + h^*(Y_t) - \langle Y_t, x \rangle$$

- Discrete-time evolution

$$F_{t+1} \leq F_t - \gamma \langle V_t, X_t - x \rangle + \frac{\gamma^2}{2K} \| V_t \|_*^2$$

- Aggregate/Telescope:

$$\overline{\mathrm{Reg}}(T) = \mathcal{O}\left(\frac{\max h - \min h}{\gamma} + \sum_{t=1}^{T} B_t + \gamma \sum_{t=1}^{T} M_t^2 \right)$$

- Take $\gamma \propto 1/\sqrt{T}$: [Why?]

$$\overline{\mathrm{Reg}}(T) = \mathcal{O}\left(\sqrt{T} + \sum_{t=1}^{T} B_t + \frac{\sum_{t=1}^{T} M_t^2}{\sqrt{T}} \right)$$

Regret guarantees of FTRL

Theorem (Shalev-Shwartz and Singer, 2007; Shalev-Shwartz, 2011)

- Assume:
  - *feedback unbiased and bounded in mean square* ($B_t = 0$, $\sup_t M_t < M$)
  - $\gamma = (2/M)\sqrt{KH/T}$ *with* $H = \max h - \min h$
- Then: *FTRL enjoys the bound*

$$\overline{\mathrm{Reg}}(T) \leq 2M\sqrt{(H/K)T} = \mathcal{O}(\sqrt{T})$$

Regret guarantees of FTRL

Theorem (Shalev-Shwartz and Singer, 2007; Shalev-Shwartz, 2011)

- Assume:
  - *feedback unbiased and bounded in mean square* ($B_t = 0$, $\sup_t M_t < M$)
  - $\gamma = (2/M)\sqrt{KH/T}$ *with* $H = \max h - \min h$
- Then: *FTRL enjoys the bound*

$$\overline{\mathrm{Reg}}(T) \leq 2M\sqrt{(H/K)T} = \mathcal{O}(\sqrt{T})$$

Observe:

- This bound is tight [Nesterov, 2004; Abernethy et al., 2008; Bubeck, 2015]
- Cannot achieve $\mathcal{O}(1)$ regret as in continuous time [Why?]
- How to do if T is unknown? [Exercise]

Which regularizer to pick?

- Assume perfect info, $v_{a,t} \in [0,1]$ [for simplicity]

Which regularizer to pick?

- Assume perfect info, $v_{a,t} \in [0,1]$ [for simplicity]
- Euclidean regularization
  - L^2-norm bound $M = |\mathcal{A}|^{1/2}$
  - Strong convexity modulus $K = 1$; $H \leq 1/2$
  - Optimal tuning gives

$$\overline{\mathrm{Reg}}(T) \leq 2\sqrt{|\mathcal{A}| \cdot T}$$

Which regularizer to pick?

- Assume perfect info, $v_{a,t} \in [0,1]$ [for simplicity]
- Euclidean regularization
  - L^2-norm bound $M = |\mathcal{A}|^{1/2}$
  - Strong convexity modulus $K = 1$; $H \leq 1/2$
  - Optimal tuning gives

$$\overline{\mathrm{Reg}}(T) \leq 2\sqrt{|\mathcal{A}| \cdot T}$$

- Entropic regularization / Exponential weights
  - L^∞-norm bound $M = 1$
  - Strong convexity modulus $K = 1$; $H = \log|\mathcal{A}|$
  - Optimal tuning gives

$$\overline{\mathrm{Reg}}(T) \leq 2\sqrt{\log|\mathcal{A}| \cdot T}$$

Which regularizer to pick?

- Assume perfect info, $v_{a,t} \in [0,1]$ [for simplicity]
- Euclidean regularization
  - L^2-norm bound $M = |\mathcal{A}|^{1/2}$
  - Strong convexity modulus $K = 1$; $H \leq 1/2$
  - Optimal tuning gives

$$\overline{\mathrm{Reg}}(T) \leq 2\sqrt{|\mathcal{A}| \cdot T}$$

- Entropic regularization / Exponential weights
  - L^∞-norm bound $M = 1$
  - Strong convexity modulus $K = 1$; $H = \log|\mathcal{A}|$
  - Optimal tuning gives

$$\overline{\mathrm{Reg}}(T) \leq 2\sqrt{\log|\mathcal{A}| \cdot T}$$

- Huge reduction in dimensionality!

Learning with bandit feedback

The bandit / partial info case:

- Play action $a_t \in \mathcal{A}$ according to mixed strategy $X_t \in \mathcal{X}$
- Receive payoff $u_t(a_t) = v_{a_t,t} \in [0,1]$

Learning with bandit feedback

The bandit / partial info case:

- Play action $a_t \in \mathcal{A}$ according to mixed strategy $X_t \in \mathcal{X}$
- Receive payoff $u_t(a_t) = v_{a_t,t} \in [0,1]$

- Importance weighted estimator:

$$V_{a,t} = \frac{\mathbb{1}(a_t = a)}{\mathbb{P}(a_t = a)} u_t(a_t) = \begin{cases} 0 & \text{if } a \neq a_t \\ \dfrac{u_t(a_t)}{X_{a,t}} & \text{if } a = a_t \end{cases}$$

Learning with bandit feedback

The bandit / partial info case:

- Play action $a_t \in \mathcal{A}$ according to mixed strategy $X_t \in \mathcal{X}$
- Receive payoff $u_t(a_t) = v_{a_t,t} \in [0,1]$

- Importance weighted estimator:

$$V_{a,t} = \frac{\mathbb{1}(a_t = a)}{\mathbb{P}(a_t = a)} u_t(a_t) = \begin{cases} 0 & \text{if } a \neq a_t \\ \dfrac{u_t(a_t)}{X_{a,t}} & \text{if } a = a_t \end{cases}$$

✓ Unbiased estimator [Verify this]

Learning with bandit feedback

The bandit / partial info case:

- Play action $a_t \in \mathcal{A}$ according to mixed strategy $X_t \in \mathcal{X}$
- Receive payoff $u_t(a_t) = v_{a_t,t} \in [0,1]$

- Importance weighted estimator:

$$V_{a,t} = \frac{\mathbb{1}(a_t = a)}{\mathbb{P}(a_t = a)} u_t(a_t) = \begin{cases} 0 & \text{if } a \neq a_t \\ \dfrac{u_t(a_t)}{X_{a,t}} & \text{if } a = a_t \end{cases}$$

✓ Unbiased estimator [Verify this]

✗ Not bounded in mean square! [$X_{a,t}$ can become arbitrarily small]

Possible outlets

Two approaches:

1. **Adjust the algorithm:** [valid for all regularizers]
   - Reduce variance by increasing exploration

$$X_t \leftarrow (1-\varepsilon_t)Q(Y_t) + \varepsilon_t \mathrm{unif}$$

   - Still unbiased; variance = $\mathcal{O}(1/\varepsilon_t)$
   - Adaptation of FTRL bounds yields

$$\overline{\mathrm{Reg}}(T) = \mathcal{O}(T^{2/3})$$

Possible outlets

Two approaches:

1. **Adjust the algorithm:** [valid for all regularizers]
   - Reduce variance by increasing exploration
   $$X_t \leftarrow (1-\varepsilon_t)Q(Y_t) + \varepsilon_t \text{unif}$$
   - Still unbiased; variance = $\mathcal{O}(1/\varepsilon_t)$
   - Adaptation of FTRL bounds yields
   $$\overline{\text{Reg}}(T) = \mathcal{O}(T^{2/3})$$

2. **Adjust the analysis:** [only valid for (EW)]
   - Derive refined bound on KL divergence [not clear for other FTRL]
   - Refined bound suitable for variance growth up to $\mathcal{O}(1/\min|X_{a,t}|)$
   - Almost tight bound:
   $$\overline{\text{Reg}}(T) = \mathcal{O}(\sqrt{|\mathcal{A}|\log|\mathcal{A}| \cdot T})$$

[Log factor can be shaved off; Audibert and Bubeck, 2010]

Recap

Quick recap:

- In general, no-regret learning does not converge to equilibrium ✗
- Multi-agent FTRL echoes replicator properties
- Discrete-time analysis $\rightsquigarrow$ next lecture
- **Regret guarantees:** $\mathcal{O}(1)$ in continuous time, $\mathcal{O}(\sqrt{T})$ in discrete [both tight]
- Non-Euclidean regularization can be **very** beneficial [EW algo]
- Bandit framework much harder, but **still possible to achieve** $\mathcal{O}(\sqrt{T})$ ✓

References I

Abernethy, Jacob, Peter L. Bartlett, Alexander Rakhlin, Ambuj Tewari. 2008. Optimal strategies and minimax lower bounds for online convex games. *COLT '08: Proceedings of the 21st Annual Conference on Learning Theory*.

Arora, Sanjeev, Elad Hazan, Satyen Kale. 2012. The multiplicative weights update method: A meta-algorithm and applications. *Theory of Computing* **8**(1) 121–164.

Audibert, Jean-Yves, Sébastien Bubeck. 2010. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research* **11** 2635–2686.

Auer, Peter, Nicolò Cesa-Bianchi, Yoav Freund, Robert E. Schapire. 1995. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*.

Aumann, Robert J. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* **1**(1) 67–96.

Aumann, Robert J. 1987. Correlated equilibrium as an expression of Baeysian rationality. *Econometrica* **55**(1) 1–18.

Bubeck, Sébastien. 2015. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning* **8**(3-4) 231–358.

Bubeck, Sébastien, Nicolò Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* **5**(1) 1–122.

References II

Cesa-Bianchi, Nicolò, Gábor Lugosi. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

Coucheney, Pierre, Bruno Gaujal, Panayotis Mertikopoulos. 2015. Penalty-regulated dynamics and robust learning procedures in games. *Mathematics of Operations Research* **40**(3) 611–633.

Freund, Yoav, Robert E. Schapire. 1999. Adaptive game playing using multiplicative weights. *Games and Economic Behavior* **29** 79–103.

Hannan, James. 1957. Approximation to Bayes risk in repeated play. Melvin Dresher, Albert William Tucker, P. Wolfe, eds., *Contributions to the Theory of Games, Volume III*, *Annals of Mathematics Studies*, vol. 39. Princeton University Press, Princeton, NJ, 97–139.

Hart, Sergiu, Andreu Mas-Colell. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* **68**(5) 1127–1150.

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

Littlestone, Nick, Manfred K. Warmuth. 1994. The weighted majority algorithm. *Information and Computation* **108**(2) 212–261.

Mertikopoulos, Panayotis, William H. Sandholm. 2016. Learning in games via reinforcement and regularization. *Mathematics of Operations Research* **41**(4) 1297–1324.

Nesterov, Yurii. 2004. *Introductory Lectures on Convex Optimization: A Basic Course*. No. 87 in Applied Optimization, Kluwer Academic Publishers.

References III

Shalev-Shwartz, Shai. 2011. Online learning and online convex optimization. *Foundations and Trends in Machine Learning* **4**(2) 107–194.

Shalev-Shwartz, Shai, Yoram Singer. 2007. Convex repeated games and Fenchel duality. *Advances in Neural Information Processing Systems 19*. MIT Press, 1265–1272.

Sorin, Sylvain. 2009. Exponential weight algorithm in continuous time. *Mathematical Programming* **116**(1) 513–528.

Viossat, Yannick, Andriy Zapechelnyuk. 2013. No-regret dynamics and fictitious play. *Journal of Economic Theory* **148**(2) 825–842.

Vovk, Vladimir G. 1990. Aggregating strategies. *COLT '90: Proceedings of the 3rd Workshop on Computational Learning Theory*. 371–383.