GAMES, DYNAMICS & LEARNING

Panayotis Mertikopoulos[1]

joint with

A. Giannou[2] T. Lianeas[2] E. V. Vlatakis-Gkaragkounis[3]

[1]French National Center for Scientific Research (CNRS) & Criteo AI Lab

[2]NTUA

[3]Columbia University

ECE–NTUA – May 28, 2021

GAMES, DYNAMICS & LEARNING

3. LEARNING IN FINITE GAMES AND BANDITS

Panayotis Mertikopoulos[1]

joint with

A. Giannou[2] T. Lianeas[2] E. V. Vlatakis-Gkaragkounis[3]

[1]French National Center for Scientific Research (CNRS) & Criteo AI Lab

[2]NTUA

[3]Columbia University

ECE-NTUA – May 28, 2021

Outline

Overview

Online learning – cont. time

Multi-agent learning – cont. time

Learning in discrete time

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** $x_t \in \mathcal{X} \coloneqq \Delta(\mathcal{A})$

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 or~~ discrete
- **Actions:** ~~continuous or~~ 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

Overview

Online learning – cont. time

Multi-agent learning – cont. time

Learning in discrete time

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**

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

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."

Online viewpoint: regret minimization

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

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

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

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

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

Possible approach: Look at distance between x_t and benchmark x

$$D_t = \frac{1}{2}\|x_t - x\|^2$$

$$\dot{D}_t = \langle x_t - x, \dot{x}_t \rangle = \text{Ugly!}$$

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 $x_a = \frac{\exp(y_a)}{Z} \Rightarrow \log x_a = y_a - \log Z$

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

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

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

$$\dot{D}_t = -\sum_a x_a \frac{d}{dt} \log x_a$$

$$\frac{d}{dt} \log x_a = \frac{d}{dt}\left[y_a - \log \sum_\beta \exp(y_\beta)\right] = \dot{y}_a - \frac{\sum_\beta \boxed{\dot{y}_\beta} \exp(y_\beta)}{\sum_\beta \exp(y_\beta)}$$

$\dot{y}_a \to V_a$, $\dot{y}_\beta \to V_\beta$, x_β

$$= V_a - \sum_\beta x_\beta V_\beta = V_a - \langle V_t, X_t \rangle$$

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

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

$$D_t = D_0 + \int_0^t [u_t(x_s) - u_t(x)]\, ds$$

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 $x \in \mathcal{X}$

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

- Evolution over time

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

- Integrate:

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

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*")

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

Observe $v = (v_1, \dots, v_n)$

$$\langle v, x \rangle = \sum_a v_a x_a$$

$$\max_a v_a = \max_{x \in \mathcal{X}} \langle v, x \rangle = \max_{x \in \mathcal{X}} \sum_a v_a x_a$$

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 $x \in \mathcal{X}$

Exercise: Show that $\Lambda(y)$ maximizes $\langle y, x\rangle - \sum_a x_a \log x_a$
s.t. $\sum_a x_a = 1$, $x_a \geq 0$

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 $h\colon \mathcal{X} \to \mathbb{R}$ 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 $x \in \mathcal{X}$ Arkadi Nemirovski 1983

- Regularized best responses

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

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

Nesterov 2004

- Follow the regularized leader

Shalev Shwartz & Singer 2006

$$\begin{aligned} \dot{y}_t &= V_t \\ x_t &= Q(y_t) \end{aligned} \tag{FTRL}$$

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

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

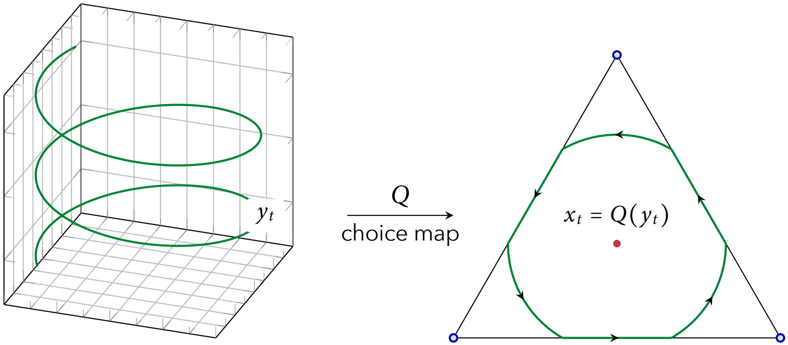
Choice map $\rightsquigarrow$ closest point projection:

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

Projection dynamics [M & Sandholm, 2016]

$$\begin{aligned} \dot{y}_t &= V_t \\ x_t &= \Pi(y_t) \end{aligned} \tag{PL}$$

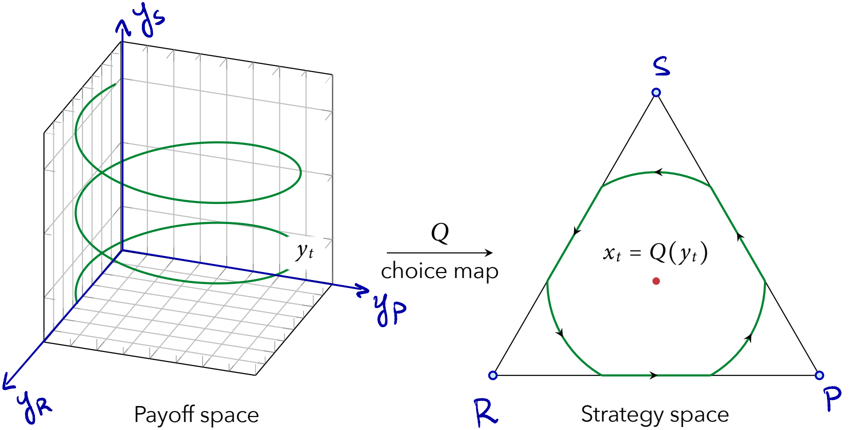
In and out of the boundary



Payoff space

Strategy space

In and out of the boundary



Key difference with replicator: faces no longer forward invariant

Portraits and examples

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

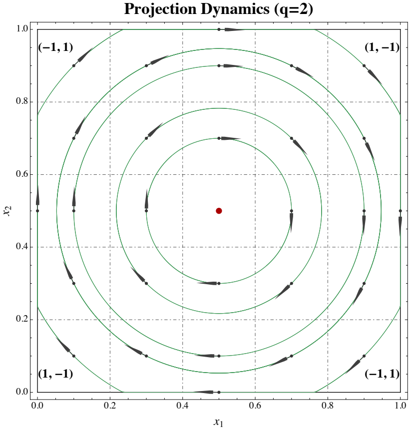


Figure: Phase portraits of (FTRL) in Matching Pennies for diferrent values of $q > 0$

Portraits and examples

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

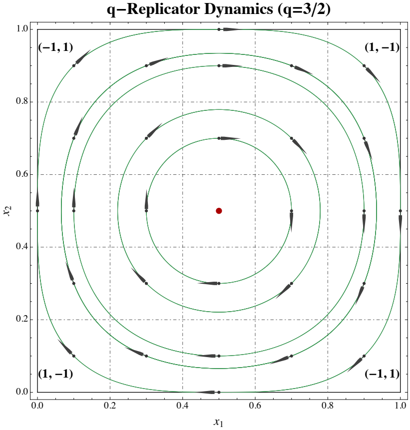


Figure: Phase portraits of (FTRL) in Matching Pennies for diferrent values of $q > 0$

Portraits and examples

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

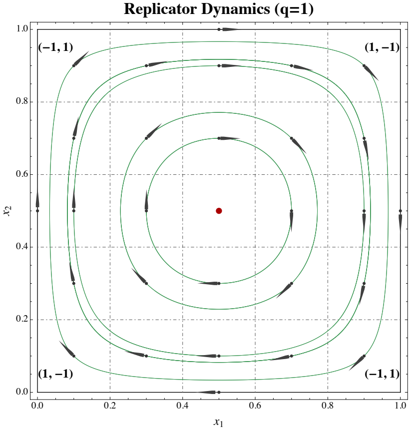


Figure: Phase portraits of (FTRL) in Matching Pennies for diferrent values of $q > 0$

Portraits and examples

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

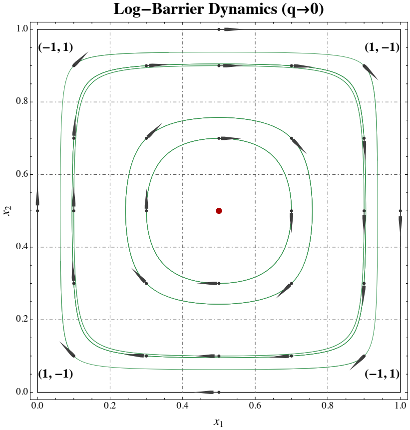


Figure: Phase portraits of (FTRL) in Matching Pennies for diferrent values of $q > 0$

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)?

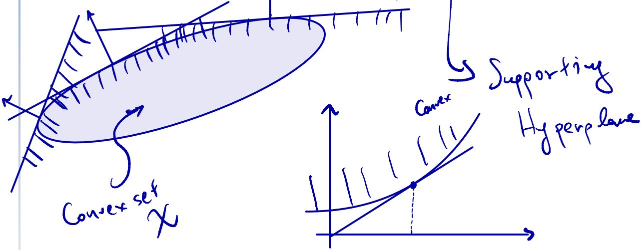
- Require primal-dual analogue of KL divergence
- Fenchel coupling [M & Sandholm, 2016; M & Zhou, 2019]

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

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*

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

Outline

Overview

Online learning – cont. time

Multi-agent learning – cont. time

Learning in discrete time

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* $(a_i; a_{-i}) \coloneqq (a_1, \dots, a_i, \dots, a_N) \in \mathcal{A} = \prod_i \mathcal{A}_i$
  - *Mixed strategies*

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

  - *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*

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

  - *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"

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

[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"

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

[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"

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

[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"

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

[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"

$\leadsto$ stray arbitrarily far from equilibrium infinitely often

[Hart and Mas-Colell, 2000, 2003]

✗ Very weak notion of "equilibrium"

$\leadsto$ 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:

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

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

Dominated strategies

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

- Consistent difference in payoffs/scores:

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

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

- Translation to choice probabilities not clear

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

Dominated strategies

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

- Consistent difference in payoffs/scores:

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

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

- Translation to choice probabilities not clear

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

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

REPLICATOR DYNAMICS

- Lyap. stable $\Longrightarrow$ x^* is a N. Eq.
- x^* is strict NE $\Rightarrow$ asym. stable under RD

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.*

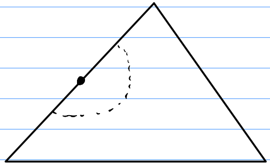
II. *If $x^* \in \mathcal{X}$ is stable, then x^* is Nash.*

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

[Special case: "folk theorem" of EGT]

REPLICATOR DYNAMICS

① Lyap. Stable ⇒ Nash



1. Argued by contra
2. Consider small nhd where non-sup strat > supp strat
3. Consider trajectory remaining in the nhd
4. Showed that 2 ⇒ scores would differ linearly in t
5. Leveraged structure of logit choice to conclude that supported strategy dies out

② Use same approach to show that Strict Nash are AS

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^* is full-support equilibrium $\implies$ (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

Overview

Online learning - cont. time

Multi-agent learning - cont. time

Learning in discrete time

Learning with a finite number of actions

Online decision-making with mixed strategies

repeat

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

Choose mixed strategy $X_t \in \mathcal{X} \coloneqq \Delta(\mathcal{A})$

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

STOCHASTIC PROCESS

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]

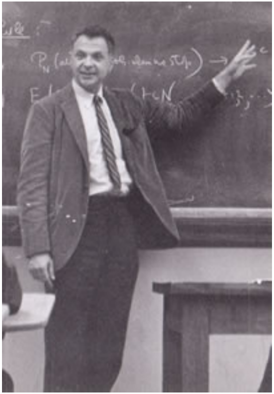
until end

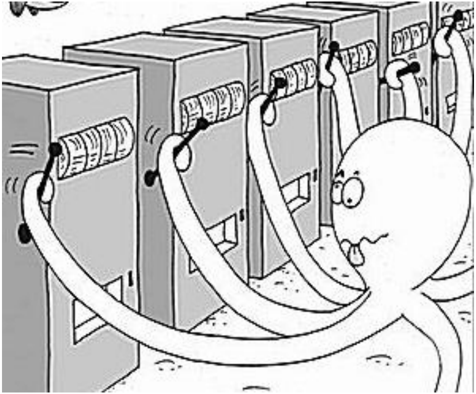
Key considerations

- **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

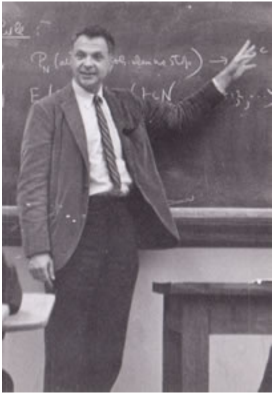
Robbins' multi-armed bandit problem: how to play in a (rigged) casino?

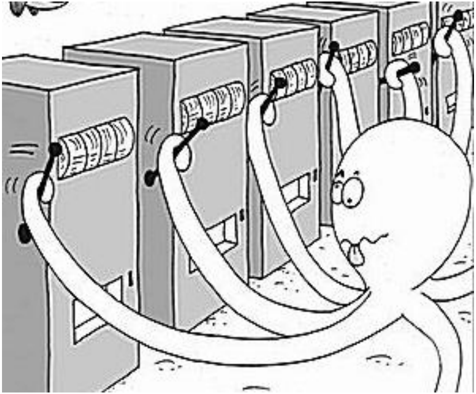




Multi-armed bandits

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

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

Online viewpoint: regret minimization

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

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

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

- **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]

MDPs, Gittins Index (1950's)
Katehakis Burnetas

Feedback

Three types of feedback (from best to worst):

- **Full, exact information**: observe entire payoff vector V_t
- **Full, inexact information**: observe estimate $\hat{V}_t$ 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 $\hat{V}_t$

$$\hat{V}_t = V_t + Z_t + b_t$$

where Z_t is zero-mean and b_t is the bias of $\hat{V}_t$

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

$$V_t = V_t + Z_t + b_t$$

where Z_t is zero-mean and b_t is the bias of V_t

$\mathbb{E}[Z_t \mid \mathcal{F}_t] = 0$

Assumptions

$\mathcal{F}_t = \sigma(X_1, \dots, X_t)$

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

ONLINE MIRROR DESCENT / LAZY (FTRL)

NEMIROVSKI YUDIN 1983

$$Y_{t+1} = Y_t + \gamma_t V_t$$
$$X_{t+1} = Q(Y_{t+1})$$

where γ_t is a variable step-size parameter

Follow the regularized leader

Implementing FTRL with full information (exact or inexact):

Nesterov 2004, 2009

DUAL AVERAGING

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

where η_t is a variable learning rate parameter

Follow the regularized leader

Implementing FTRL with full information (exact or inexact):

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

where γ_t is a variable step-size parameter

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

$\in$ Convex Analysis

h is str.cvx

Hess(h) $\succcurlyeq$ m I

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

Follow the regularized leader

Implementing FTRL with full information (exact or inexact):

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

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

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

$\cancel{\quad} \dfrac{Y_{a,t+1}}{\sum_{b \in \mathcal{A}} Y_{a,t+1}}$

[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

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

Cont. time

$$\dot{F}_t = -\langle V_t, X_t - x \rangle$$

- Discrete-time evolution

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

PRF.

$$F_{t+1} = h(x) + h^*(Y_{t+1}) - \langle Y_{t+1}, x \rangle$$

$$= h(x) + \underbrace{h^*(Y_t + \gamma_t V_t)}_{h^*(Y_t + \gamma_t V_t) - h^*(Y_t)} - \langle Y_t, x \rangle - \gamma_t \langle V_t, x \rangle + h^*(Y_t)$$

F_t

$$F_{t+1} = F_t - \gamma_t \langle V_t, x \rangle + \underbrace{h^*(Y_t + \gamma_t V_t) - h^*(Y_t)}$$

$\nabla h^* = Q$

By LC of Q : $h^*(Y_t + \gamma_t V_t) - h^*(V_t) \leq \gamma_t \langle Q(Y_t), V_t \rangle + \frac{\gamma_t^2}{2K} \| V_t \|^2$

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

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

Reg(T)

Did not exist in the cont. time analysis

- Aggregate/Telescope:

Telescoping

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

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, 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, 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*

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

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]

References I

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

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

J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2635–2686, 2010.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, 1995.

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

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

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

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

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

References II

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

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

D. Fudenberg and D. K. Levine. *The Theory of Learning in Games*, volume 2 of *Economic learning and social evolution*. MIT Press, Cambridge, MA, 1998.

D. Fudenberg and J. Tirole. *Game Theory*. The MIT Press, 1991.

J. Hannan. Approximation to Bayes risk in repeated play. In M. Dresher, A. W. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games, Volume III*, volume 39 of *Annals of Mathematics Studies*, pages 97–139. Princeton University Press, Princeton, NJ, 1957.

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

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

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

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020.

References III

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

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

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Number 87 in Applied Optimization. Kluwer Academic Publishers, 2004.

N. Nisan, T. Roughgarden, É. Tardos, and V. V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, 2007.

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

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

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

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

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

J. W. Weibull. *Evolutionary Game Theory*. MIT Press, Cambridge, MA, 1995.