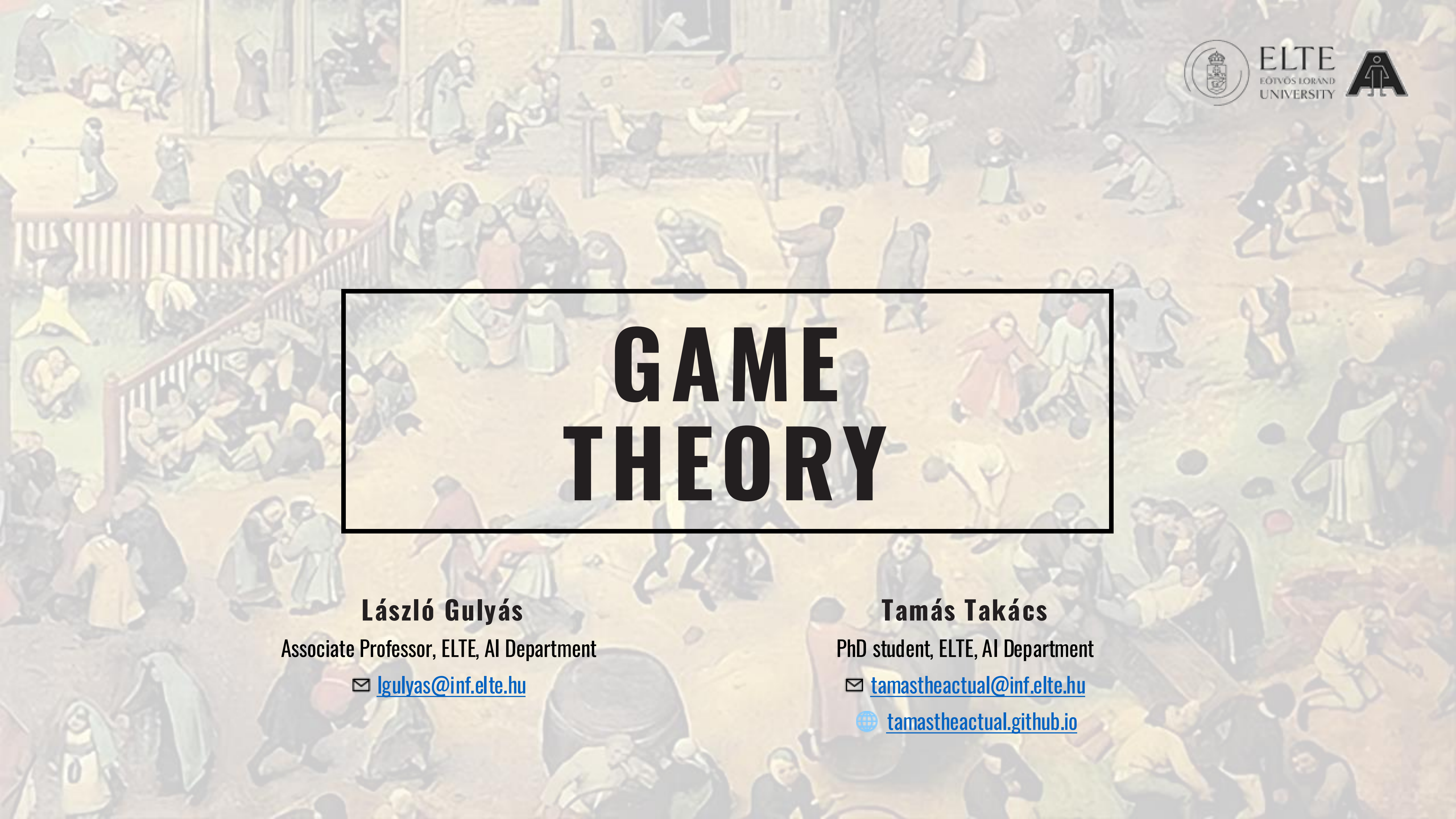
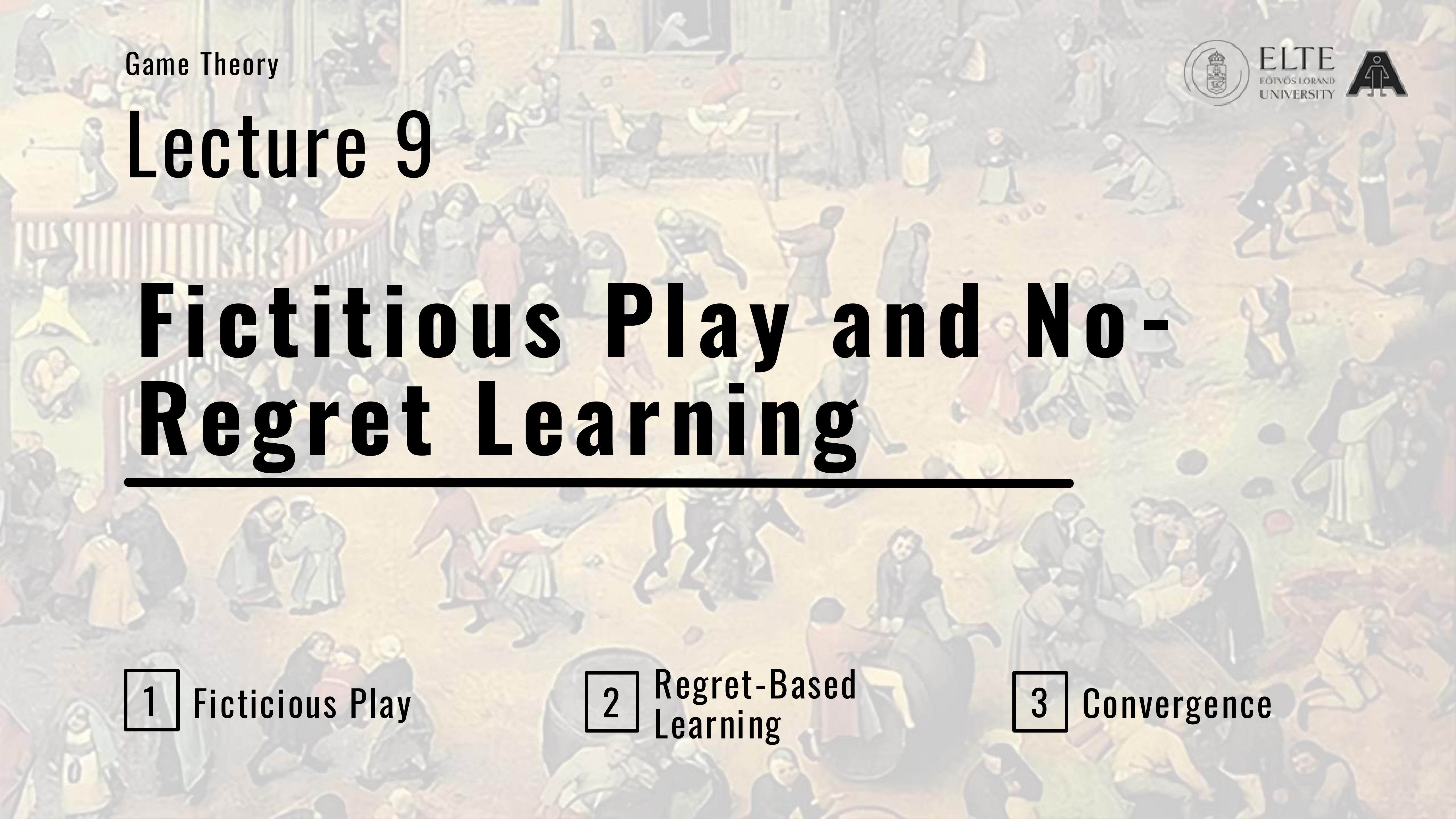


GAME THEORY

László Gulyás
Associate Professor, ELTE, AI Department
✉ lgulyas@inf.elte.hu

Tamás Takács
PhD student, ELTE, AI Department
✉ tamastheactual@inf.elte.hu
🌐 tamastheactual.github.io



Game Theory

Lecture 9

Fictitious Play and No-Regret Learning

1 Ficticious Play

2 Regret-Based Learning

3 Convergence

Correlated Equilibrium

A **correlated equilibrium (CE)** is a distribution $\mu \in \Delta(A)$ such that for every player i and every pair of actions $a_i, a_i' \in A_i$:

$$\sum_{a_{-i}} \mu(a_i, a_{-i})[u_i(a_i, a_{-i}) - u_i(a_i', a_{-i})] \geq 0$$

Notation:

- $\mu(a_i, a_{-i}) =$ probability mediator recommends profile (a_i, a_{-i})
- Sum over a_{-i} gives expectation conditional on i receiving a_i

CE: Mediator Interpretation

Interpretation (mediator form):

1. Mediator samples $a = (a_1, \ldots, a_n)$ from μ
2. Privately recommends a_i to player i
3. Each player finds it optimal to **obey** recommendation

Obedience constraint:

- Conditional on receiving recommendation a_i, player i's expected utility from playing a_i is at least as good as deviating to any other action a_i'
- Mediator must be **trusted** and recommendations **private**

Example: Prisoner's Dilemma (Dominance)

	C	D
C	$(3, 3)$	$(0, 4)$
D	$(4, 0)$	$(1, 1)$

Unique Nash equilibrium: (D, D) → payoff $(1,1)$
Correlated equilibrium:

- Can \mu put weight on (C, C) to achieve $(3,3)$?
- **No!** D strictly dominates C for both players
- CE constraints require:

$$\sum_{a_{-i}} \mu(C, a_{-i}), [u_i(C, a_{-i}) - u_i(D, a_{-i})] \geq 0$$

- But $u_i(C, a_{-i}) < u_i(D, a_{-i})$ for all a_{-i}(dominance!)
- So $\mu(C, a_{-i}) = 0$ for all a_{-i}

CE cannot overcome **dominance** (credibility issue)

Logit Response: Foundation

Random-utility model (Gumbel noise):

- Player i perceives utility: $u_i(a_i; \sigma_{-i}) + \varepsilon_{a_i}$
- Noise $\varepsilon_{a_i} \sim$ type-I extreme value (i.i.d.), scaled by $1/\lambda$
- Player maximizes perceived utility

Choice probabilities are logit:

$$\sigma_i(a_i) = \frac{\exp\{\lambda \cdot u_i(a_i; \sigma_{-i})\}}{\sum_{a_i' \in A_i} \exp\{\lambda \cdot u_i(a_i'; \sigma_{-i})\}}$$

Intuition: Higher-payoff actions get higher probability, not certainty

Classic Learning Dynamics

Three classic dynamics:

1. **Fictitious Play (FP):** Best-respond to empirical frequency
2. **Replicator Dynamics (RD):** Evolutionary selection (above-average payoffs grow)
3. **No-regret learning:** Minimize regret over time

Connection to equilibria:

- FP converges in some games (zero-sum, potential)
- RD stationary points include NE
- No-regret → CCE; no swap-regret → CE

Blackwell Approachability

Blackwell's theorem (1956):
A convex set $S \subseteq R^d$ is **approachable** if player can ensure average vector payoff converges to S regardless of opponent's play

Connection to CE:

- Define **regret vector**: components $=$ regrets for each possible swap
- **Target set**: Non-positive orthant (all regrets ≤ 0)
- Approachability $\Leftrightarrow$ achieving no swap regret

Blackwell's algorithm:

1. Compute current average payoff vector $\overline{v}_t$
2. Project $\overline{v}_t$ to nearest point in $S \rightarrow getv^*$
3. Play action that moves expected vector toward v^*
4. Repeat $\rightarrow$ converge to S

PoA: Interpretation

Interpretation:

- $\text{PoA} = 1$ → equilibrium is socially optimal
- $\text{PoA} = 2$ → equilibrium achieves half of optimal welfare
- Measures **worst-case inefficiency** of decentralized play

Note: Basic PoA definition introduced in previous lecture. Here we extend to correlated equilibria and develop the smoothness framework.

Robust definition:

- Can define PoA for **NE, CE, or CCE**
- Smoothness framework gives **unified bounds**

Game Theory

Lecture 9

Fictitious Play and No-Regret Learning

1 Ficticious Play

2 Regret-Based Learning

3 Convergence

Learning Dynamics vs Static Equilibria

Consider these scenarios:

- **Repeated game**: Players observe history and adjust strategies over time
- **Traffic routing**: Commuters learn congestion patterns through experience
- **Online advertising**: Bidders adjust bids based on observed outcomes

How do strategic agents learn to play equilibria without centralized coordination?

Why Learning Matters

Static equilibrium has limitations:

- Assumes common knowledge of game structure
- Requires infinite computational resources
- Doesn't explain how players reach equilibrium

Learning dynamics provide:

- Plausible behavioral foundations
- Algorithms for finding equilibria
- Robustness to model misspecification
- Connection between process and outcome

Question: Which learning rules converge? To which equilibria? At what rate?

Fictitious Play: The Classic Approach

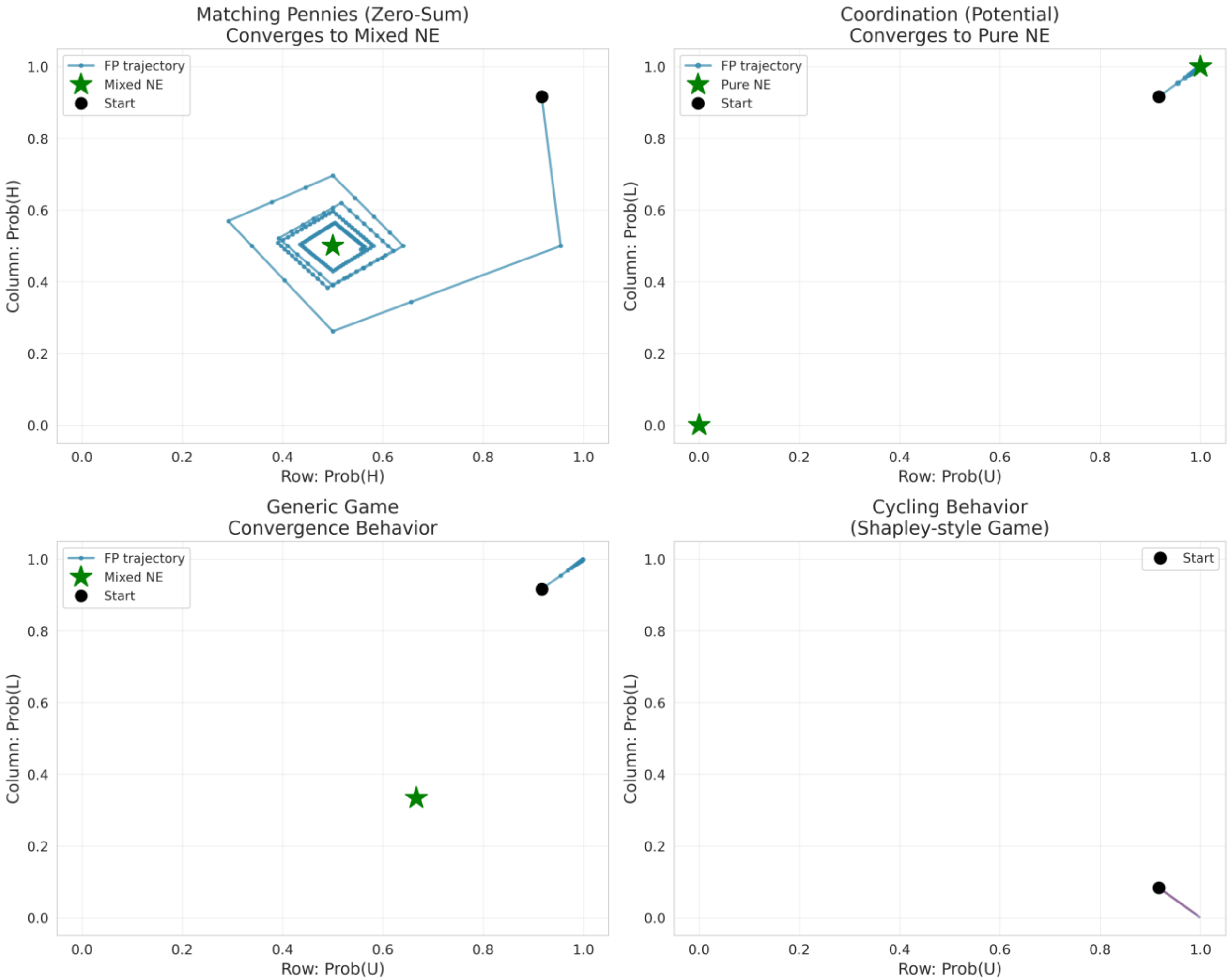
Brown 1951; Robinson 1951:

- Players maintain beliefs = empirical frequencies of past play
- Each player best-responds to current beliefs
- Simple, intuitive, behaviorally plausible

Key questions:

- Does FP converge to Nash equilibrium?
- How fast?
- In which classes of games?

FP Trajectories Across Game Types



Fictitious Play

Finite normal form game:

$$\mathcal{G} = (N, \{A_i\}_{i=1}^{n}, \{u_i\}_{i=1}^{n})$$

- A_i finite action set for player
- joint action space
- $u_i : A \to R$ payoff function
- Mixed strategies $\Delta(A_i)$, product distributions $\Delta(A) = \prod_i \Delta(A_i)$

Empirical frequencies:

$$\widehat{\pi_{i,t}} = \frac{1}{t}\sum_{\tau=1}^{t} \delta_{a_i^\tau} \in \Delta(A_i)$$

Joint empirical distribution: $\widehat{\mu_t} = \frac{1}{t}\sum_{\tau=1}^{t} \delta_{a^\tau}$

Fictitious Play Algorithm

Definition (Fictitious Play):

In round $t \geq 1$, each player i:

1. **Form belief:** Opponents play independently according to empirical frequencies $\widehat{\pi_{-i,t-1}}$
2. **Best respond:** Choose

$$a_i^t \in \arg\max_{a_i \in A_i} u_i\left(a_i, \widehat{\pi_{-i,t-1}}\right)$$

Initialization: Choose any $a^1 \in A$, set $\widehat{\pi_{i,1}} = \delta_{a_i^1}$

Tie-breaking: Arbitrary or via fixed selection rule

Best Response Dynamics

Continuous-time formulation:

Let $x_i(t) \in \Delta(A_i)$ be continuous strategy. Best response dynamics:

$$\dot{x}_i(t) \in \mathrm{BR}_i\big(x_{-i}(t)\big) - x_i(t)$$

where $\mathrm{BR}_i(\cdot)$ is set of mixed best responses.

Connection: Discrete FP is Euler discretization of BRD with step size $1/t$ on beliefs

Smooth Best Response

Regularized variant:

Replace BR with smooth response, e.g., **Gibbs/logit:**

$$\mathrm{SBR}_i(x_{-i};\eta) \propto \exp\big(\eta \cdot u_i(\cdot, x_{-i})\big)$$

where $\eta > 0$ is inverse temperature

Logit dynamics:

$$x_i^{t+1} = (1-\alpha_t)x_i^t + \alpha_t \cdot \mathrm{SBR}_i\big(x_{-i}^t;\eta\big)$$

Benefits: Unique response, connection to mirror descent and FTRL

Convergence: Positive Results

Theorem (Robinson 1951): In two-player zero-sum games, FP empirical frequencies converge to minimax equilibrium set.

Theorem (Monderer & Shapley 1996): In potential games, FP converges to pure Nash equilibrium.

Theorem: In dominance-solvable games, FP reaches unique NE in finite time.

Theorem: In 2-player $2 \times N$ games with diagonal strict concavity, FP converges.

Proof Sketch: Zero-Sum Convergence

Setup: Row payoff matrix A, column payoff $-A$

Duality gap at $(\boldsymbol{x}, \boldsymbol{y})$:

$$\mathrm{Gap}(x, y) = \max_{x'} (x')^\top Ay - \min_{y'} x^\top A\, y'$$

Key observations:

- Row best-responds to $\hat{y}_t$ → maximizes against belief
- Column best-responds to $\hat{x}_t$ → minimizes against belief
- Duality gap $\mathrm{Gap}(\hat{x}_t, \hat{y}_t)$ is non-increasing

Conclusion: $\mathrm{Gap}(\hat{x}_t, \hat{y}_t) \to 0 \Rightarrow$ empirical frequencies approach minimax set

Potential Games and FP

Potential function: $\Phi: A \to R$ such that for all i, a_i, a_i', a_{-i}:

$$u_i(a_i', a_{-i}) - u_i(a_i, a_{-i}) = \Phi(a_i', a_{-i}) - \Phi(a_i, a_{-i})$$

Theorem: In potential games:

- Best response increases Φ weakly
- FP generates sequence with non-decreasing Φ values
- Φ bounded above → sequence converges to pure NE (local max of Φ)

Examples: Congestion games, coordination games

Shapley's Cycling Example

Shapley (1964): Constructed 3×3 game where:

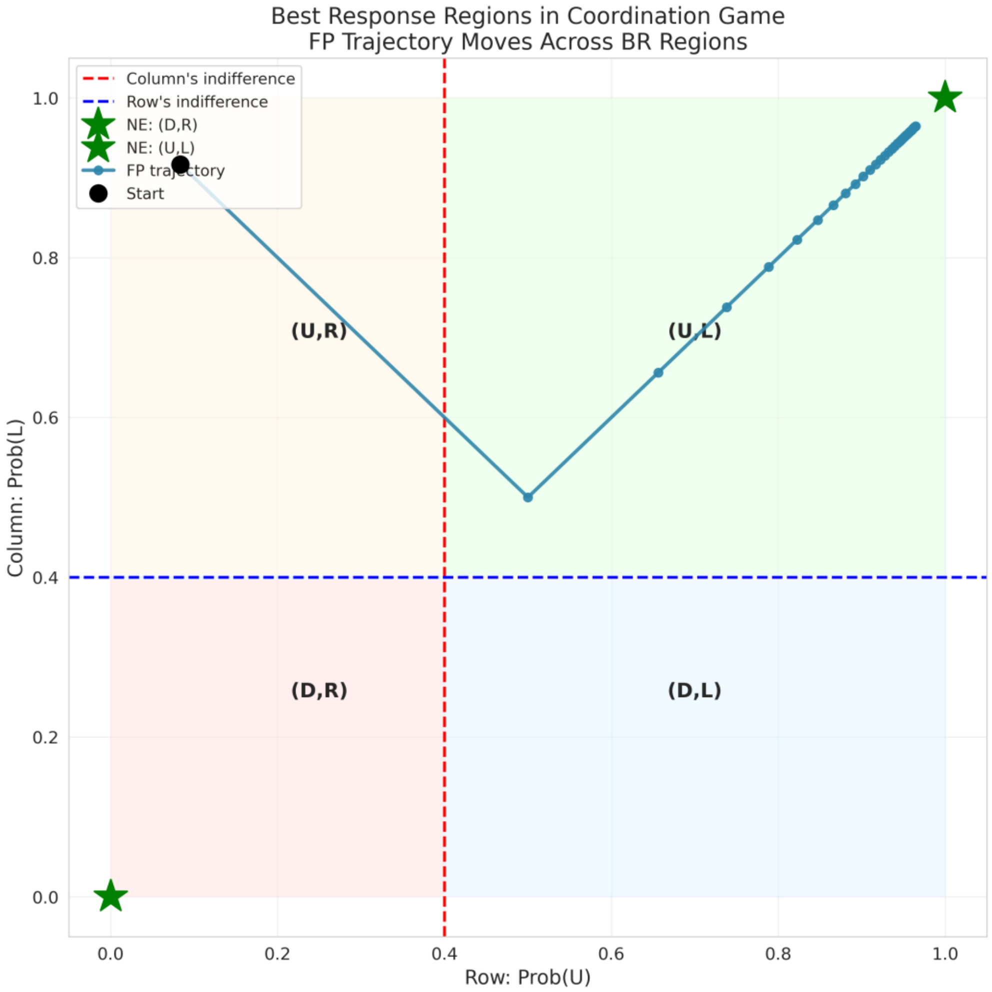
- FP does **not** converge pointwise
- Beliefs cycle around unique mixed NE
- Time-average may converge to NE

Implication: FP convergence is not universal

Example payoffs (simplified):

	L	**C**	**R**
U	$(0,0)$	$(1,0)$	$(0,1)$
M	$(0,1)$	$(0,0)$	$(1,0)$
D	$(1,0)$	$(0,1)$	$(0,0)$

Best Response Regions



Stochastic Fictitious Play

Perturbation: Add noise or use SBR with temperature $1/\eta$

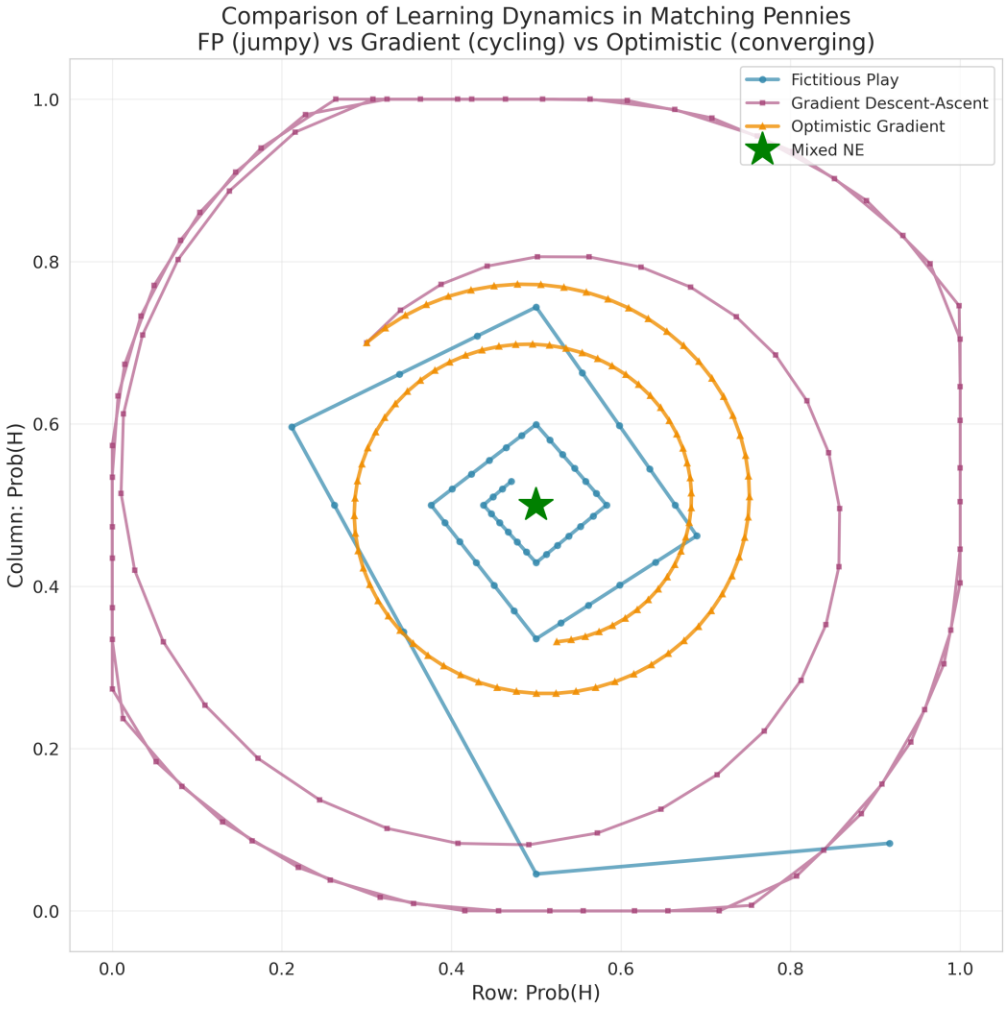
$$x_i^{t+1} = (1 - \alpha_t)x_i^t + \alpha_t \cdot \mathrm{SBR}_i\left(x_{-i}^t; \eta\right)$$

with decreasing step size $\alpha_t \to 0$

Convergence: Under suitable noise and structure:

- Converges to **quantal response equilibria** (fixed η)
- Converges to NE as $\eta \to \infty$ (under additional assumptions)
- Connection to stochastic approximation theory

FP vs Other Dynamics



Game Theory

Lecture 9

Fictitious Play and No-Regret Learning

1 Ficticious Play

2 Regret-Based Learning

3 Convergence

Motivation for Regret

FP limitations:

- Requires computing best responses (NP-hard in general)
- Can cycle or diverge
- No finite-time performance guarantees

Regret-based approach:

- Compare performance to counterfactuals
- Vanishing regret → empirical distribution approaches equilibrium
- Efficient algorithms with provable bounds

Regret Definitions

For sequence of play $a^1, \dots, a^T$, player i:

External regret relative to fixed action a_i':

$$R_T^{\text{ext}}(a_i') = \sum_{t=1}^{T} \left(u_i(a_i', a_{-i}^t) - u_i(a^t) \right)$$

Internal regret for conditional swap $a \to a'$:

$$R_T^{\text{int}}(a \to a') = \sum_{t=1}^{T} \mathbb{1}\{a_i^t = a\} \left(u_i(a', a_{-i}^t) - u_i(a, a_{-i}^t) \right)$$

Swap regret for replacement map $\phi: A_i \to A_i$:

$$R_T^{\text{swap}}(\phi) = \sum_{t=1}^{T} \left(u_i(\phi(a_i^t), a_{-i}^t) - u_i(a^t) \right)$$

Hannan Consistency

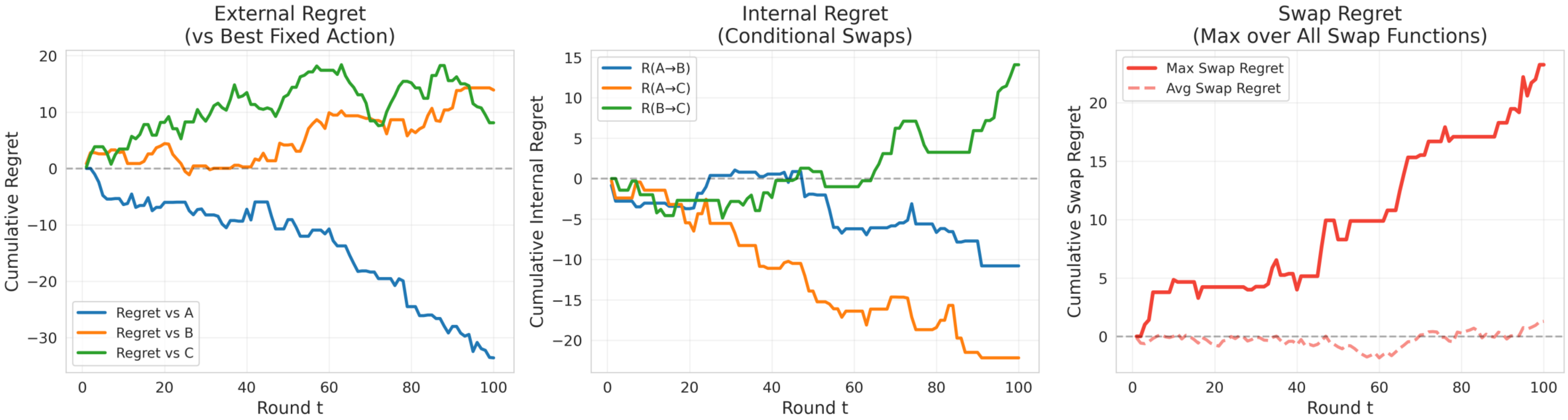
Definition: Player i is **Hannan consistent** if

$$\frac{1}{T} \max_{a_i' \in A_i} R_T^{\text{ext}}(a_i') \to 0 \quad \text{as } T \to \infty$$

Interpretation: Average external regret vanishes

Achievability: Many efficient algorithms (Hedge, FTRL, OMD, etc.)

Hannan Consistency



Regret Hierarchy

Inclusion chain:

$$\text{No swap regret} \subseteq \text{No internal regret} \subseteq \text{No external regret}$$

Reason:

- Swap regret includes all possible modifications
- Internal regret is conditional swapping
- External regret is unconditional (fixed action)

Implication: Swap regret is strongest notion

Regret and Equilibria

Theorem (Hart & Mas-Colell 2000):

If all players have vanishing external regret:

$$\widehat{\mu_T} \to \mathrm{CCE}(\mathcal{G})$$

Theorem (Foster & Vohra 1997):

If all players have vanishing internal (or swap) regret:

$$\widehat{\mu_T} \to \mathrm{CE}(\mathcal{G})$$

Special case: In two-player zero-sum, no external regret → last iterate approaches minimax value

Hedge (Multiplicative Weights)

Algorithm (Hedge/MW):

Maintain weights $w_t(a)$ for $a \in A_i$. At round t:

1. Play distribution:

$$p_t(a) = \frac{w_{t-1}(a)}{\sum_b w_{t-1}(b)}$$

1. Observe payoff vector $g_t(a) = u_i\left(a, a_{-i}^t\right)$
2. Update weights:

$$w_t(a) = w_{t-1}(a) \cdot \exp\left(\eta \cdot g_t(a)\right)$$

Initialize: $w_1(a) = 1$ for all a

Hedge: Regret Bound

Theorem: With learning rate $\eta = \sqrt{\frac{2 \log m}{T}}$ (payoffs in $[0,1]$):

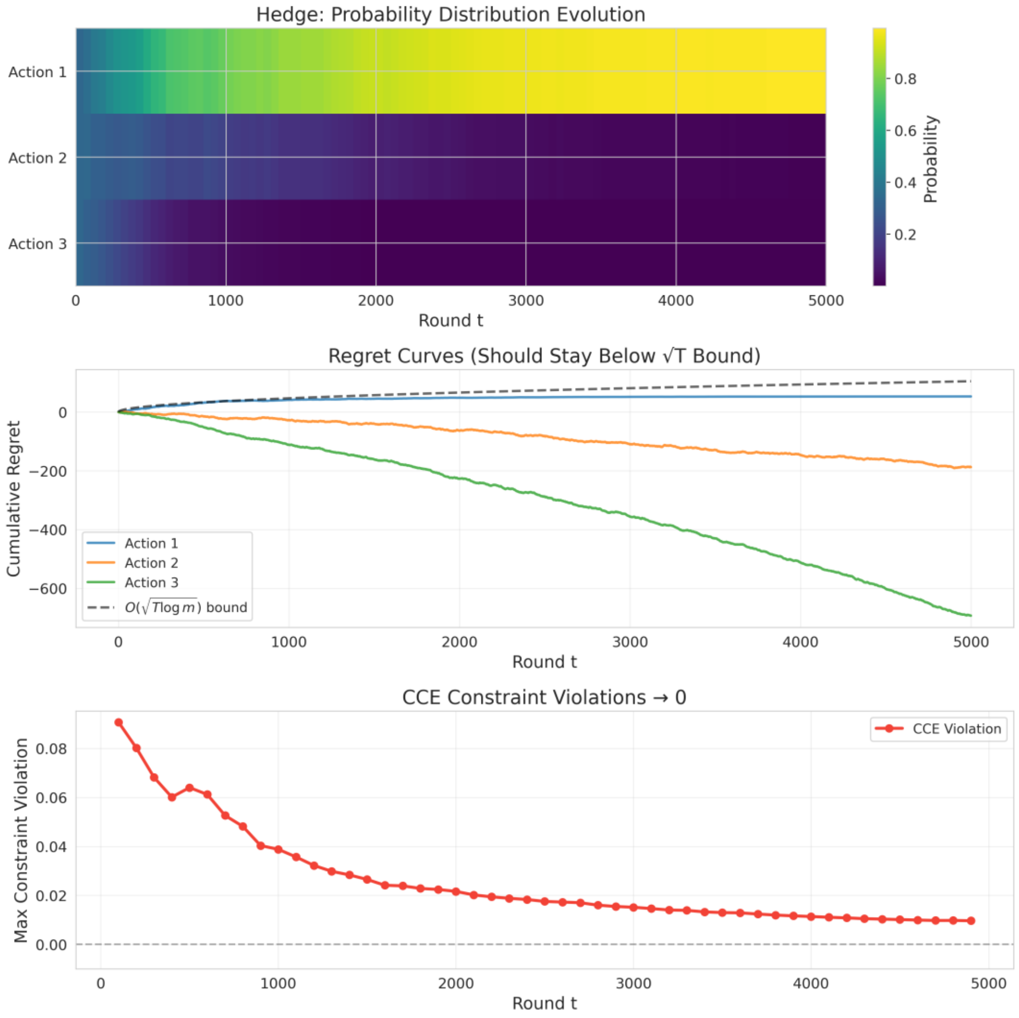
$$\max_{a' \in A_i} R_T^{\text{ext}}(a') \leq \sqrt{2T \log m}$$

Rate: $O(\sqrt{T \log m})$ regret over T rounds, m actions

Average regret: $O\left(\sqrt{\frac{\log m}{T}}\right) \to 0$

Optimality: Matches lower bound up to constants

Hedge Convergence



Hedge: Probability Distribution Evolution

Regret Curves (Should Stay Below √T Bound)

CCE Constraint Violations → 0

Follow The Regularized Leader (FTRL)

Algorithm: At round t, play

$$x_t = \arg\max_{x \in \Delta(A_i)} \left\{ \langle x, g_{1:t-1} \rangle - \frac{1}{\eta} \mathcal{R}(x) \right\}$$

where $g_{1:t-1} = \sum_{\tau=1}^{t-1} g_\tau$ and $\mathcal{R}: \Delta(A_i) \to R$ is strongly convex regularizer

Common choices:

Negative entropy: $\mathcal{R}(x) = \sum_a x(a) \log x(a) \to$ Hedge
Euclidean norm: $\mathcal{R}(x) = \frac{1}{2} |x|_2^2 \to$ gradient descent on simplex

Online Mirror Descent (OMD)

Algorithm: At round t:

1. Play $x_t = \nabla\Psi^*(\theta_{t-1})$
2. Observe payoff gradient g_t
3. Update dual parameters: $\theta_t = \theta_{t-1} + \eta g_t$

where Ψ is Legendre function (strongly convex), Ψ^* its conjugate

Connection: OMD = FTRL with specific regularizer choices

Example: Entropic regularizer → exponential weights (Hedge)

OMD: Regret Analysis

Theorem: If Ψ is λ-strongly convex w.r.t. norm $|\cdot|$ and payoffs bounded by G:

$$R_T \leq \frac{\text{diam}^2}{2\eta T} + \frac{\eta G^2 T}{2\lambda}$$

Optimal rate: Choose $\eta \sim \sqrt{\frac{\lambda \cdot \text{diam}^2}{G^2 T}}\ to\ get\ R_T = O(\sqrt{T})$

Strong convexity: Faster $O(\log T)$ regret in exp-concave settings

Optimistic OMD and Last-Iterate Convergence

Optimistic variant: Use previous gradient for prediction

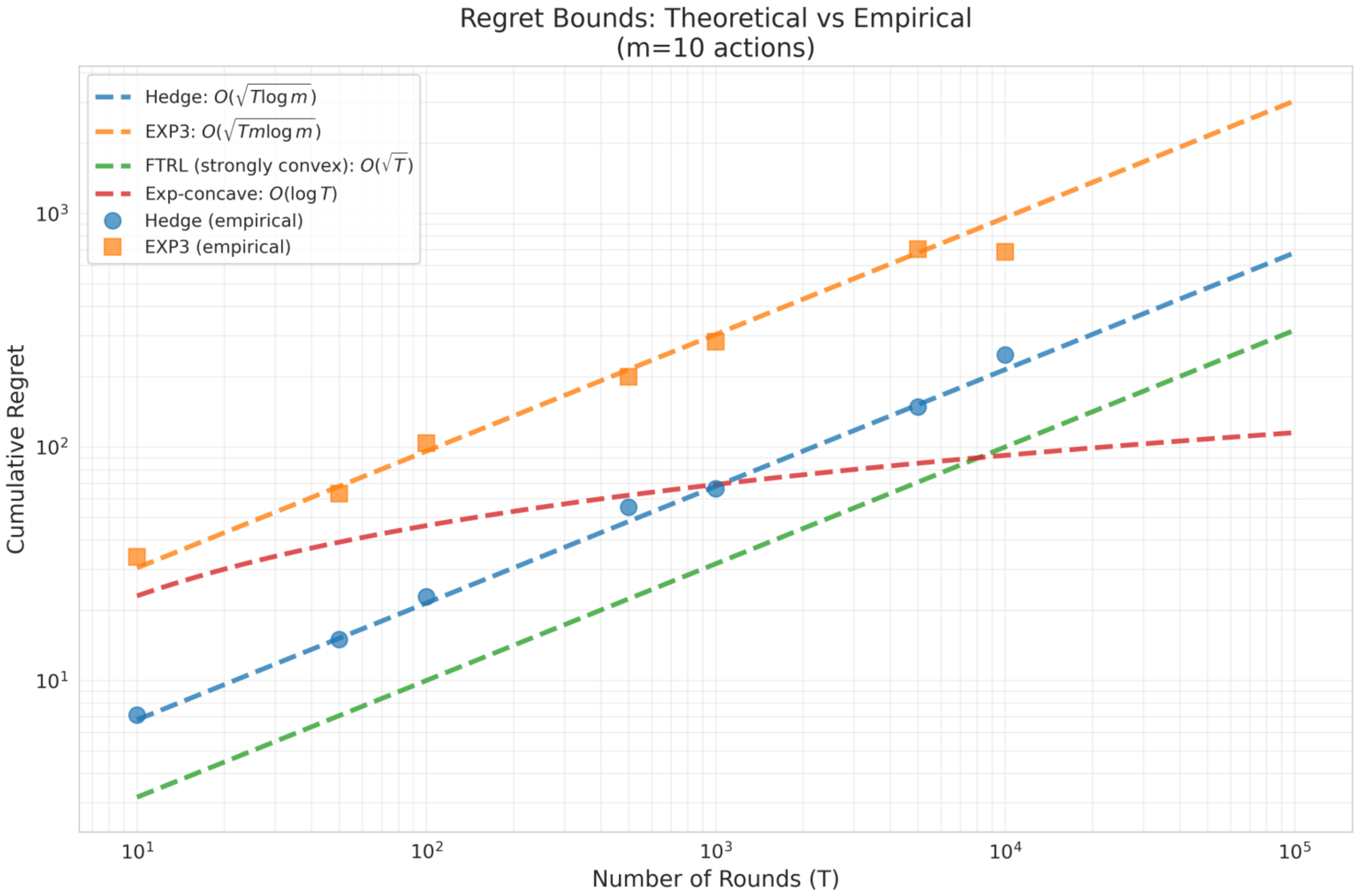
$$\theta_t = \theta_{t-1} + \eta(2g_{t-1} - g_{t-2})$$

Theorem (Daskalakis et al. 2021): In two-player zero-sum games:

- Optimistic OMD: **last iterate** converges to NE
- Standard OMD: only time-average guaranteed

Extragradient: Related method with similar last-iterate guarantees

Regret Bounds Comparison



EXP3: Bandit Feedback

Setup: Only observe payoff of chosen action (not full payoff vector)

Algorithm (EXP3): Play

$$p_t(a) = (1-\gamma)\frac{w_{t-1}(a)}{\sum_b w_{t-1}(b)} + \frac{\gamma}{m}$$

Observe reward r_t for chosen action a_t. Update:

$$w_t(a) = w_{t-1}(a) \cdot \exp\left(\eta \frac{\gamma}{m} \frac{r_t}{p_t(a)} \mathbb{1}\{a = a_t\}\right)$$

Importance weighting to estimate full payoffs

EXP3: Regret Bound

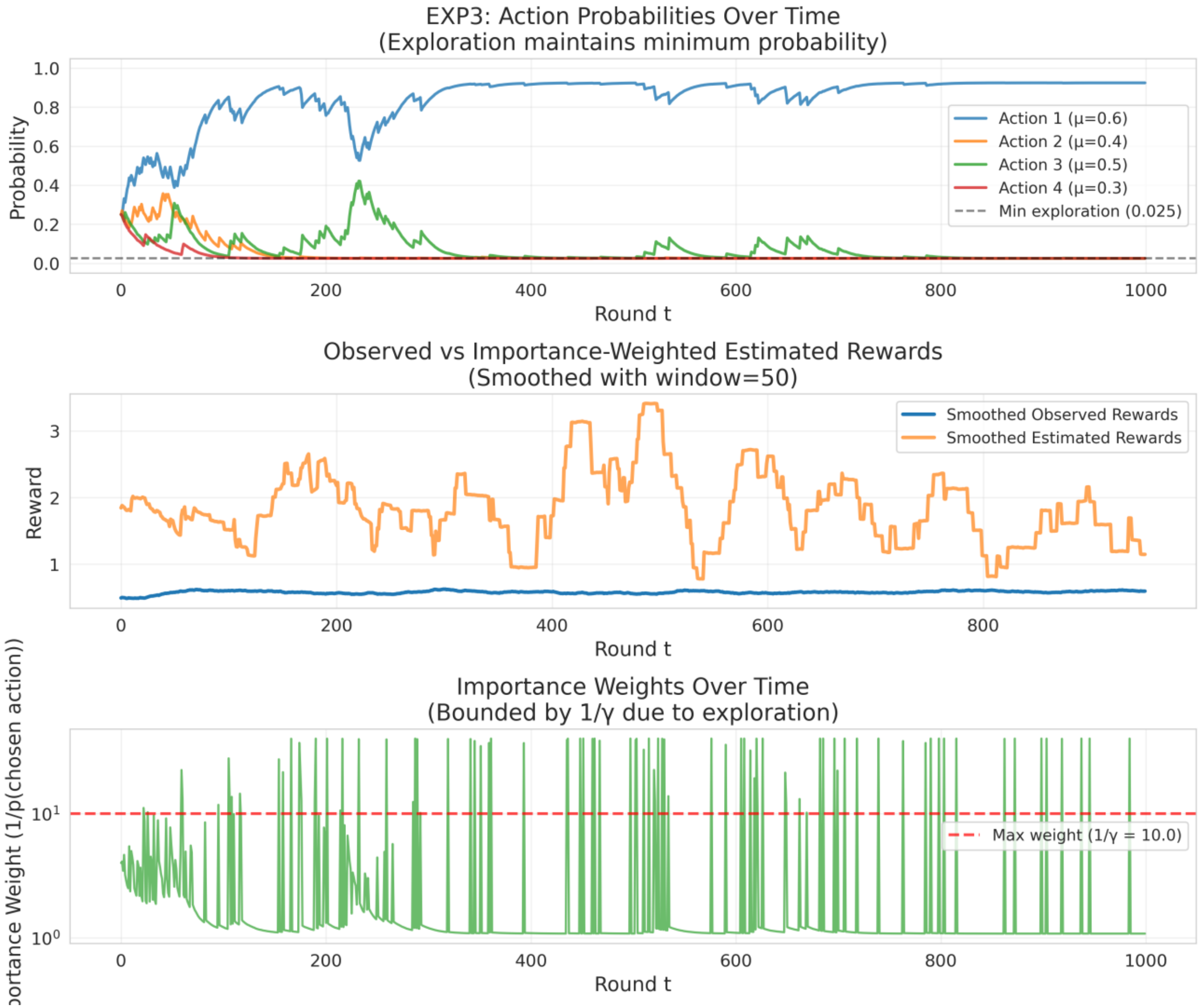
Theorem: With $\eta = \sqrt{\frac{\log m}{Tm}}$ and $\gamma = \sqrt{\frac{m \log m}{T}}$:

$$E[R_T] = O\left(\sqrt{Tm \log m}\right)$$

Cost of bandit feedback $\sqrt{m}$ factor worse than full information

Exploration-exploitation: γ controls min exploration probability

EXP3 Exploration-Exploitation



EXP3: Action Probabilities Over Time
(Exploration maintains minimum probability)

Observed vs Importance-Weighted Estimated Rewards
(Smoothed with window=50)

Importance Weights Over Time
(Bounded by 1/γ due to exploration)

Game Theory

Lecture 9

Fictitious Play and No-Regret Learning

1 Ficticious Play

2 Regret-Based Learning

3 Convergence

External Regret → CCE

Theorem: If all players use algorithms with $o(T)$external regret:

$$\widehat{\mu_T} \to T \to \infty \mathrm{CCE}(\mathcal{G})$$

Proof sketch:

- CCE requires no profitable deviation before observing recommendation
- External regret $=$ 0 means no profitable unilateral deviation
- Time-average satisfies CCE constraints asymptotically

Rate: With $O\left(\sqrt{T}\right)$ regret, $O\left(1/\sqrt{T}\right)$ constraint violation

Internal/Swap Regret → CE

Theorem: If all players use algorithms with $o(T)$ internal regret:

$$\widehat{\mu_T} \to T \to \infty \mathrm{CE}(\mathcal{G})$$

Reduction: Hart & Mas-Colell (2000) show how to convert external regret algorithm to internal regret:

- Maintain m copies of external regret minimizer
- Copy a tracks regret for swap $a \to \cdot$
- Aggregate recommendations via master algorithm

Cost: m –fold increase in computation

Zero-Sum Games: Special Case

Theorem (Freund & Schapire 1999): In two-player zero-sum:

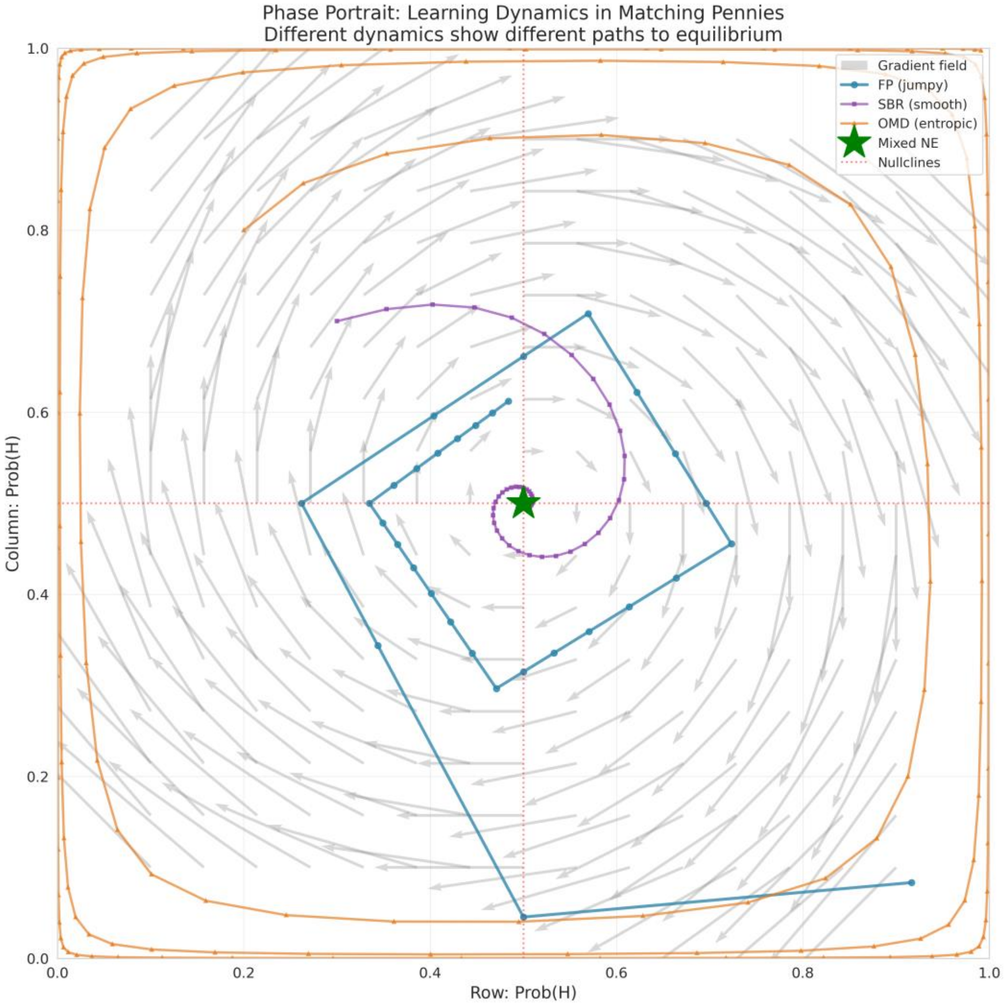
If both players use no-external-regret algorithms:

$$\frac{1}{T}\sum_{t=1}^{T} u_1(a^t) \rightarrow v^*$$

where v^* is minimax value

Stronger result (Daskalakis et al.): With optimistic methods, last **iterate** converges

Learning Dynamics Phase Portrait



Step Size Selection

Fixed step size η:

- Constant learning rate
- May not converge (cycles near equilibrium)
- Works with small enough η in strongly convex settings

Decaying step size η_t:

- Common: $\eta_t = \eta_0/\sqrt{t}$ or η_0/t
- Guarantees convergence
- Slow near convergence

Adaptive step sizes:

- AdaGrad-style: adjust per coordinate based on gradient history
- Parameter-free: no tuning needed (e.g., Coin Betting)

Regularizer Selection

Entropy regularizer:

- Natural for probability distributions
- Efficient multiplicative updates (Hedge)
- Geometry: KL divergence

Euclidean regularizer:

- Simple gradient steps
- Works in general Euclidean spaces

Time-Average vs Last-Iterate

Time-average $\widehat{\mu_T}$:

- Guaranteed to converge to CE/CCE (with no-regret)
- Robust to noise and adversarial opponents
- Standard in online learning theory

Last-iterate μ_T:

- Converges in zero-sum with optimistic methods
- Fails in general (can cycle)
- Practically relevant: current strategy, not historical average

Zero-sum → last-iterate; general sum → time-average

Exercise 1

	H	T
H	$(1,-1)$	$(-1,1)$
T	$(-1,1)$	$(1,-1)$

Tasks:

1. Run FP for 50 rounds starting from (H,H). Plot empirical frequencies.
2. Compute duality gap $\mathrm{Gap}(\hat{x}_t, \hat{y}_t)$ over time.
3. Does it converge to $(0.5,0.5)$? At what rate?

Exercise 2

Setup: 3-action game with adversarial payoffs

$$g_t = \begin{pmatrix} \sin(t/10) \\ \cos(t/10)\ 0 \end{pmatrix} + \text{noise}$$

Tasks:

1. Implement Hedge with $\eta = \sqrt{2 \log 3 / T}, T = 1000$
2. Track cumulative regret $R_T^{\text{ext}}(a)$ for each action a
3. Verify regret $\leq \sqrt{2T \log 3}$
4. Plot probability distribution evolution

Exercise 3

	Opera	Football
Opera	$(2, 1)$	$(0, 0)$
Football	$(0, 0)$	$(1, 2)$

Tasks:

1. Run Hedge for both players, 5000 rounds
2. Compute empirical distribution $\widehat{\mu_T}$
3. Check CE constraints:

$$\sum_{a_{-i}} \hat{\mu}(a_i, a_{-i})\big(u_i(a_i, a_{-i}) - u_i(a_i', a_{-i})\big) \geq 0$$

4. Estimate constraint violations

Exercise 4

Setup: Network congestion game

- 2 players, 3 routes
- Cost on route e: $c_e(x_e) = x_e$(number of users)

Tasks:

1. Formulate as potential game with $\Phi = \sum_e \sum_{k=1}^{x_e} k$
2. Run FP from random initialization
3. Verify potential Φ is non-decreasing
4. Show convergence to pure NE
5. Count rounds to convergence

Summary

- **Fictitious Play:** Simple, intuitive, converges in zero-sum and potential games, but can cycle
- **Regret notions:** External → CCE, internal/swap → CE
- **No-regret algorithms:** Hedge, FTRL, OMD achieve $O(\sqrt{T})$ regret
- **EXP3:** Handles bandit feedback with $O(\sqrt{Tm})$ regret
- **Convergence:** Time-average guaranteed, last-iterate requires optimism
- **Practical design:** Step sizes, regularizers, and feedback structure matter

Course Textbooks

- Bonanno, G. (2024). ***Game Theory (3rd ed.)***. University of California, Davis. Received from: GT Book
- Axelrod, R. (1984). ***The Evolution of Cooperation***. Basic Books. Received from: Axelrod Article
- Nisan, N., Roughgarden, T., Tardos, É., & Vazirani, V. V. (2007). ***Algorithmic Game Theory***. Cambridge University Press. Received from: AGT Book
- Myerson, R. B. (1991). ***Game Theory: Analysis of Conflict.*** Harvard University Press. Received from: GT Book 2F.
- Christianos et al., ***Multi-Agent Reinforcement Learning: Foundations and Modern Approaches***, 2023. Received from: MARL Book.pdf
- Shoham, Y., & Leyton-Brown, K. (2008). ***Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations***. Cambridge University Press. Received from: MARL Book.pdf
- `***nashpy***` documentation (readthedocs). Link: NashPy Docs

That’s All for Today!