CSCE 631: Module 9, Lecture 2

Why PFSP? The Learning Rules Behind Superhuman Self-Play:
Whom to Play, and How to Update

Fall 2026 · Alan Kuhnle · Texas A&M University

Fictitious play has no good rate. Hedge, FTRL, and mirror descent do. Which one did the superhuman systems actually run, and why?

Today: Learning Objectives and Plan

By the end you should be able to:

1. state what is and is not known about fictitious play's rate;
2. derive PFSP's $p \to \infty$ limit and say what PFSP does *not* guarantee;
3. separate *whom the learner plays* from *how the learner updates*;
4. write the NeuRD, R-NaD, and magnetic mirror descent updates, and say which iterate each makes converge.

Part	Topic (minutes)	Recall from
0	Finish L20: Pluto and the puzzle (7)	L20, Part VI
I	Fictitious play's rate story (7)	fictitious play, MWU (M3)
II	FSP, NFSP, PFSP (10)	PFSP (L20)
III	Correction: averaging neural networks (6)	CFR averages (M5); Kuhn (M4)
IV	Two axes: whom to play, how to update (6)	PSRO (M7)
V	NeuRD: Hedge in policy space (7)	Hedge, replicator (M3)
VI	R-NaD and DeepNash (9)	FTRL, cycling (M3)
VII	Magnetic mirror descent and Ataraxos (10)	mirror descent (M3)
VIII	The 2026 verdict; the answer (8)	exploitability (M5, L20)
	Buffer: questions or one discussion prompt (5)	

Part 0: Finishing L20 with Pluto

Pure self-play from current weights, and a puzzle it leaves us

Pluto (2026): A Brood War Bot From Pure Self-Play

Part 0 ▷ Competition report and the author's README; not peer-reviewed

Bot	Games	Win	Loss	Win %	Bots Beaten
Pluto	2884	2844	40	98.61	9
PurpleWave	2883	2037	846	70.66	8
Stardust	2884	2237	647	77.57	7
BananaBrain	2889	2135	754	73.9	6
Microwave	2884	1318	1566	45.7	4
McRave	2889	1313	1576	45.45	4
Steamhammer	2883	1025	1858	35.55	4
UAlbertaBot	2882	904	1978	31.37	2
Venator	2883	483	2400	16.75	1
InfestedArtosis	2889	129	2760	4.47	0
Total	**14425**	**14425**	**14425**	**N/A**	**N/A**

Figure: Churchill, CoG 2026 report (round robin, 6 of 11 columns); reused from L20.

- By Vegard Mella. Won the round robin and the final (4 to 1 over PurpleWave).
- Organizer: the first Brood War competition bot "made completely by reinforcement learning."
- One 315M-parameter network for all races. PPO-style actor-critic; both sides of every training game use the *same current weights*.
- No replays, no search, no older checkpoints, no league.

Source: Churchill (2026), competition report; Mella (2026), README, Model section. Not peer-reviewed.

Whom Do You Train Against? Three Answers

	AlphaStar (2019)	**OpenAI Five (2019)**	**Pluto (2026)**
Human data	971k replays, then RL	none	none
Opponents	league: SP, PFSP, exploiters	80% latest, 20% past	latest self only
Learner update	actor-critic (V-trace, UPGO)	PPO	PPO-style actor-critic
Diversity from	exploiters; strategy statistic	past-opponent pool	sampled openings; diversity bonus
Source	Nature	arXiv preprint	README, report

SP: self-play against the latest self. *Strategy statistic*: a summary of a build order that AlphaStar conditions on. *Diversity bonus*: Pluto's small reward for less-often-seen units and tech.

Source: Vinyals et al. (2019), MS pp. 2 to 3, 13 to 16; Berner et al. (2019) p. 5; Mella (2026), Model and Training signal sections.

L20 in One Table, and the Puzzle It Leaves

Part 0 ▷ Compressed from L20's synthesis

System	Core technique	Course concept
Libratus	CFR blueprint; subgame solving	Nash, CFR (M5, M6)
AlphaZero	MCTS plus self-play	minimax, policy iteration (M4)
AlphaStar	imitation; league with PFSP	fictitious play, PSRO (M3, M7)
OpenAI Five	PPO; 80/20 self-play	team games
Pluto	pure self-play; opening bandit	limits of self-play; bandits

The puzzle. AlphaStar reports that naive self-play "has high Elo, but is more forgetful." Pluto uses nothing else and wins. So two questions: *whom* should the learner play, and *which learning rule* should it run?

Source: L20 synthesis slide; Vinyals et al. (2019), Fig. 3 caption (MS p. 9).

Today's Question

Part 0 ▷ Asked in class after L20

"Why PFSP? FSP was one of the weakest convergence algorithms we saw in online learning, with no theoretical rate. Why not mirror descent, FTRL, MWU?"

Where we are going

1. Fictitious play's rate really is weak, and partly open (Part I).
2. PFSP chooses *opponents*; it is not the learner's update rule (Part II).
3. The guarantees we know are about *averages*; deep methods can approximate averages, at a cost (Part III).
4. Since 2020 the field has rebuilt the *update* from M3's toolkit: Hedge, regularized FTRL, mirror descent (Parts V to VII).
5. In 2026, a large study finds plain regularized policy gradient at least as good as the FP-, PSRO-, and CFR-based deep methods it tested (Part VIII).

Part I: Fictitious Play's Rate Story

Convergent since 1951; how fast is still partly open

Recall: Fictitious Play and Hedge in a Zero-Sum Matrix Game

Part I ▷ Recall Module 3

Payoff matrix $M \in \mathbb{R}^{r \times s}$ (r row, s column actions); mixed strategies lie in the simplices Δ_r, Δ_s.

- **Fictitious play (FP).** Each round, each player plays a pure best response to the opponent's *empirical average* $\bar{y}^t = \frac{1}{t}\sum_{\tau \le t} e_{j(\tau)}$ of past actions.
- **Duality gap** g_t: how much the two players could gain, in total, by best-responding to $(\bar{x}^t, \bar{y}^t)$. It is 0 exactly at a Nash equilibrium.
- **Robinson (1951):** under FP, $g_t \to 0$ in two-player zero-sum (2p0s) games. No rate is stated.
- **M3, Hedge (MWU) for both players**, payoffs scaled so losses lie in $[0, 1]$, $n = \max(r, s)$, step $\eta = \sqrt{\ln n / T}$: the average pair is an ε-Nash equilibrium with $\varepsilon \le 4\sqrt{\ln n / T}$.

Both guarantees concern *averages* $(\bar{x}, \bar{y})$. In M3's matching-pennies demo, MWU's last iterates cycled.

Source: M3 Lectures 7 to 9 (course notation); Daskalakis and Pan (2014) §I (Robinson).

How Fast Does Fictitious Play Converge?

Result	Statement	Fine print
Shapiro (1958)	value converges at $O\left(t^{-1/(r+s-2)}\right)$	upper bound; hidden factor $a \cdot 2^{r+s}$, a the largest $\lvert M_{ij}\rvert$
Karlin (1959)	conjecture: $O(t^{-1/2})$	the rate no-regret methods get
Daskalakis and Pan (FOCS 2014)	on the $n \times n$ identity, FP *may* converge at $\Theta(t^{-1/n})$	only with ties broken *arbitrarily*: refutes the *strong* conjecture
Abernethy, Lai, Wibisono (2021)	$O(t^{-1/2})$ for diagonal M, and tight	lexicographic ties: the *weak* conjecture, for this class only

Status. The weak form (fixed lexicographic or random tie-breaking) is open for general 2p0s matrices.

Source: Shapiro's bound as stated by Brandt, Fischer, and Harrenstein (2013 preprint) §1; Shapiro (1958) not opened. Daskalakis and Pan (2014) Thm 1, Remark 1; Abernethy et al. (2021) abstract.

In Practice, FP Often Looks Like $t^{-1/2}$

Part I ▷ Abernethy, Lai, and Wibisono (2021), Fig. 2a

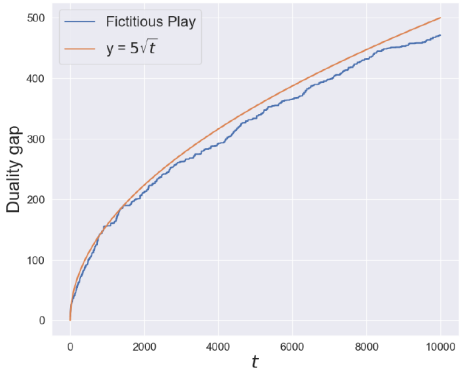


Figure: Abernethy et al., SIAM 2021, Fig. 2a.

- The plot uses action *counts*, so the gap grows like $5\sqrt{t}$; divided by t, the gap of the averages falls like $t^{-1/2}$.
- Maximum over 100 runs on random 10×10 Gaussian games.
- Evidence only: Shapiro's bound for $r = s = 10$ is $t^{-1/18}$.

Pause and predict. Why can an adversary who breaks ties slow FP down, when every individual step is still a best response?

Source: Abernethy et al. (2021) §7 and Fig. 2 caption.

Part II: From Fictitious Play to PFSP

Approximate best responses, a learned average, and a new way to pick opponents

Fictitious Self-Play (Heinrich, Lanctot, and Silver, ICML 2015)

- **XFP** (full width): realization-equivalent to normal-form FP (M4), so it "inherits its convergence guarantees." Cost per iteration is linear in the number of game states.
- **FSP** (sampled): best responses by reinforcement learning, average strategies by supervised learning.
- FSP implements *generalized weakened FP*: with step α_t and best-response error ϵ_t, it needs

$$\alpha_t \to 0, \qquad \sum_t \alpha_t = \infty, \qquad \epsilon_t \to 0,$$

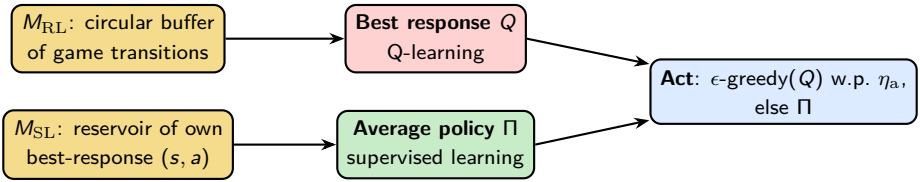
  plus vanishing perturbations.

"It remains an open question whether guaranteed convergence can be achieved with a finite computational budget per iteration." (§6). No rate is given.

Source: Heinrich et al. (2015), abstract, §§2.4, 6, Definition 5.

Neural Fictitious Self-Play (NFSP)

Part II ▷ Heinrich and Silver (2016), arXiv preprint



- Π learns the *average* of past best responses, by maximizing the log-probability of stored actions.
- η_{a} is the anticipatory parameter (the paper's η; 0.1 in Leduc). With $\eta_{\mathrm{a}} = 1$ the agent is plain DQN, whose average "does not approach a Nash equilibrium either."
- Leduc exploitability 0.06; no convergence theorem.

Source: Heinrich and Silver (2016), arXiv:1603.01121v2, §§1, 3, 4.1 to 4.2, Alg. 1.

PFSP, Recalled: Weighting Opponents by How Hard They Are

Learner A samples a frozen opponent B from the candidate set $\mathcal{C}$ with

$$\Pr[B] = \frac{f(\omega_B)}{\sum_{C \in \mathcal{C}} f(\omega_C)}, \qquad \omega_B = \Pr[A \text{ beats } B] \quad (\text{L20 wrote } x_B),$$

assuming some candidate has $f(\omega_C) > 0$.

- $f \equiv 1$: uniform over the past, which is FSP.
- $f_{\mathrm{hard}}(\omega) = (1-\omega)^p$, $p > 0$: the default. Opponents you always beat get weight 0.
- $f_{\mathrm{var}}(\omega) = \omega(1-\omega)$: opponents near your level; used "for main exploiters and struggling main agents."

AlphaStar: one can view f_{hard} "as a form of smooth approximation of max-min optimisation, as opposed to max-avg that FSP imposes."

Source: Vinyals et al. (2019), author MS pp. 15 to 16.

Why "Max-Min"? The $p \to \infty$ Limit

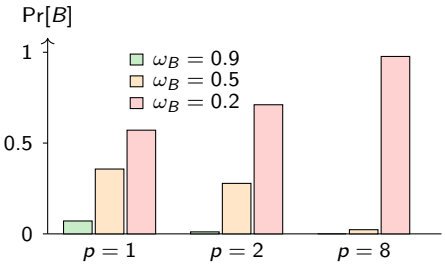
Part II ▷ Derived for this lecture; full derivation in the notes

Let $\omega_{\min} = \min_C \omega_C < 1$. For $\omega_B > \omega_{\min}$,

$$\frac{\Pr[B]}{\Pr[B']} = \left(\frac{1-\omega_B}{1-\omega_{\min}}\right)^p \to 0$$

for any hardest B'. So all weight goes to the hardest opponents.

As an objective. The PFSP gradient $\sum_B (1-\omega_B)^p \nabla \omega_B$ is, up to a positive factor, the descent direction of $\|\mathbf{1} - \boldsymbol{\omega}\|_{p+1}$, where $\boldsymbol{\omega} = (\omega_B)_{B \in \mathcal{C}}$. As $p \to \infty$ this norm tends to $\max_B(1-\omega_B)$: *maximize the worst win rate* (max-min). At $p = 0$ it is the sum: *maximize the average* (max-avg).



Pause and predict. If A beats every candidate with probability 1, what does f_{hard} do?

Source: Weights computed for this lecture from L20's example; the norm reading is our derivation, not a result in Vinyals et al. (2019).

Evidence for PFSP: AlphaStar's Ablation

Part II ▷ Vinyals et al. 2019, Extended Data Fig. 5

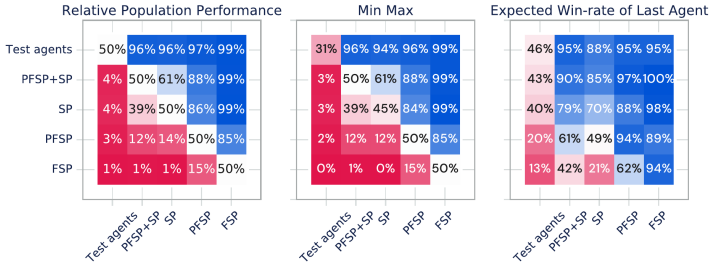


Figure: Vinyals et al., Nature 2019, Extended Data Fig. 5 (author manuscript p. 26).

- Row against column: PFSP beats FSP 85%, 85%, 89%; it "outperforms FSP under all measures considered."
- Ablation runs of 10^{10} steps. **No published convergence theory for PFSP** was found.

Source: Vinyals et al. (2019), Extended Data Fig. 5 caption and Fig. 3 caption.

Part III: Can You Average a Neural Network?

Yes. Here is how, and what it costs

Correction: Deep Methods *Can* Approximate the Average

Wrong: "you can't average neural networks." **Right:** the guarantees for MWU, FP, and CFR are about the *average* strategy, and there are three standard ways to approximate it.

1. **A second network (NFSP).** Train Π by supervised learning on a reservoir buffer of past best-response actions.
2. **A second network (Deep CFR, ICML 2019).** Store each sampled strategy $\sigma^t(I)$ at an infoset I with weight t; train a strategy network on that reservoir. Or keep every iteration's network and sample one at random during play.
3. **A population.** A league's snapshots, with weights, are a mixture. Play it by sampling one snapshot per game; under perfect recall, Kuhn's theorem (M4) gives an equivalent behavioral strategy.

Source: Heinrich and Silver (2016) §3; Brown, Lerer, Gross, and Sandholm (2019) §4 and Alg. 1; Kuhn's theorem: M4.

What Averaging Costs, and What It Leaves Out

- **Costs.** A second network and buffer; supervised approximation error; a lag (the average of t iterates moves $O(1/t)$ per step).
- **The bounds do not cover the approximation.** Deep CFR's Theorem 1 keeps a term growing with the value network's excess loss, so average regret need not reach 0; the strategy network adds its own error.
- **AlphaStar was not running no-regret learning anyway.** Its learners use actor-critic updates (V-trace, UPGO) on sampled returns, not Hedge on exact per-action losses. Part V shows that gap.
- **Last-iterate methods** instead try to make the current network itself converge (Parts VI and VII).

Ataraxos (2026) controls its updates so that it can avoid "more onerous techniques, such as trajectory importance reweighting and policy averaging."

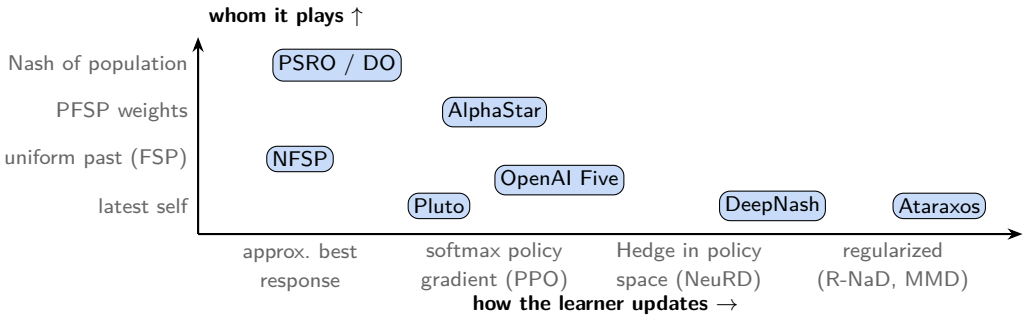
Source: Brown et al. (2019) Thm 1, Cor. 1, §4; Vinyals et al. (2019) MS pp. 2 to 3; Sokota et al. (2026) arXiv v2 p. 13. The $O(1/t)$ lag is our arithmetic.

Part IV: Whom to Play, and How to Update

Two design choices that the systems make separately

Two Axes: Opponent Selection and Learner Update

Part IV ▷ The course's framing, not a taxonomy from the literature

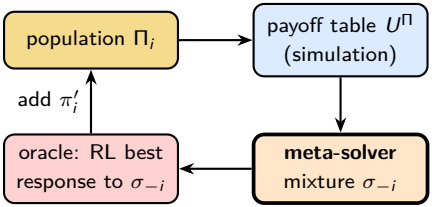


OpenAI Five sits between rows (80% latest, 20% past). DeepNash uses NeuRD plus regularization, so it straddles two columns.

Source: Placement from the cited papers (Parts 0 to VII); the two-axis framing is ours (INT).

PSRO: The Opponent Axis Is a Pluggable Meta-Solver

Part IV ▷ Recall Module 7: PSRO and double oracle



Choices for σ_{-i}
latest policy ⇒ self-play
uniform ⇒ fictitious play
Nash of U^{Π} ⇒ double oracle
regret matching, Hedge, PRD (tried by Lanctot et al.)
PFSP weights ⇒ league-like

- PRD: projected replicator dynamics. Any meta-solver can be plugged in.
- The oracle's *learning rule* is a separate choice: the other axis.

Source: Lanctot et al. (2017) §3, Alg. 1, p. 4; Vinyals et al. (2019) MS p. 13.

Guarantees on the Opponent Axis

- **Online Double Oracle** (Dinh et al., TMLR 2022): no-regret learning inside DO. Its regret bound depends on k, the size of the "effective strategy set," not on the total number of pure strategies.
- **Anytime DO and Anytime PSRO** (McAleer et al., ICLR 2024): exploitability "does not increase between iterations."
- **SP-PSRO** (same paper): also adds an approximately optimal *stochastic* policy each iteration; often needs few iterations.

So the opponent axis *does* have no-regret options with guarantees. AlphaStar's PFSP is not one of them; it is a heuristic with an intuitive max-min reading.

Pause and predict. (Optional if time is short.) Is plain PSRO's exploitability guaranteed to decrease each iteration? What did McAleer et al. observe?

Source: Dinh et al. (2022), abstract; McAleer et al. (2024), abstract.

Part V: NeuRD, Hedge in Policy Space

Softmax policy gradient is not Hedge; a one-line change makes it so

Softmax Policy Gradient Versus Hedge, in One State

Part V ▷ Hennes et al., AAMAS 2020, §3.1

One state, all actions observed. Logits $y \in \mathbb{R}^{|\mathcal{A}|}$, policy $\pi = \mathrm{softmax}(y)$, utilities $u_t(a)$ (in M3's loss notation, $u = -\ell$), mean $\bar{u}_t = \sum_a \pi_t(a) u_t(a)$.

Softmax PG (SPG): ascend $\pi \cdot u_t$ in y. Using $\partial \pi(a')/\partial y(a) = \pi(a')[\mathbf{1}_{a'=a} - \pi(a)]$,

$$y_t(a) = y_{t-1}(a) + \eta\, \pi_t(a) \left[u_t(a) - \bar{u}_t\right].$$

Hedge = NeuRD (tabular): add the advantage to the logit,

$$y_t(a) = y_{t-1}(a) + \eta \left[u_t(a) - \bar{u}_t\right].$$

Subtracting $\bar{u}_t$ changes nothing, since softmax is shift-invariant.

The only difference is the factor $\pi_t(a)$. An action with low probability learns slowly under SPG, even if it has become the best response.

Source: Hennes et al. (2020), §§2.3 to 3.1, eqs. (3), (6).

What the Factor $\pi_t(a)$ Costs: Regret

Part V ▷ Hennes et al. 2020, Fig. 1

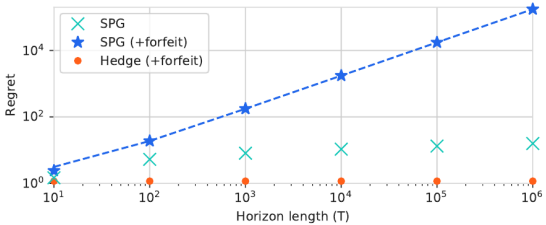


Figure: Hennes et al., AAMAS 2020, Fig. 1.

- Matching pennies against an opponent that switches from H to T after 40% of the rounds. A "forfeit" action always pays -1.
- SPG's regret then grows about linearly; Hedge's stays near constant.

Pause and predict. The forfeit action is useless. Why does adding it hurt SPG and not Hedge? (Hint: look at the logit gap.)

Source: Hennes et al. (2020), §1.1 and Fig. 1 caption.

NeuRD: The One-Line Fix, and What It Guarantees

Part V ▷ Hennes et al. 2020, eq. (10), Statement 1

With function approximation, logits $y(s, a; \theta)$, a critic q, and baseline v:

$$\theta_t = \theta_{t-1} + \eta \sum_a \nabla_\theta \, y(s, a; \theta_{t-1}) \left[q(s, a) - v(s) \right].$$

SPG has $\nabla_\theta \pi$ where NeuRD has $\nabla_\theta y$: the update "bypasses the gradient step through the softmax."

- **Statement 1:** Hedge $\equiv$ discrete-time replicator dynamics $\equiv$ single-state, all-actions, tabular NeuRD.
- In the sequential tabular case, CFR with NeuRD at each infoset is CFR with Hedge, so CFR's regret bound applies (M5).
- Convergence is in the *time average*, like Hedge.

NeuRD fixes the update rule. It still needs an average, and in zero-sum games the last iterate of replicator dynamics cycles (next part).

Source: Hennes et al. (2020), abstract, eq. (10), Alg. 1, Statement 1, §3.3; Perolat et al. (2022), p. 24.

Part VI: Regularize to Converge

R-NaD and DeepNash: making the last iterate settle

The Problem: FTRL Cycles; Regularization Stops It

Part VI ▷ Perolat et al., ICML 2021, Fig. 1

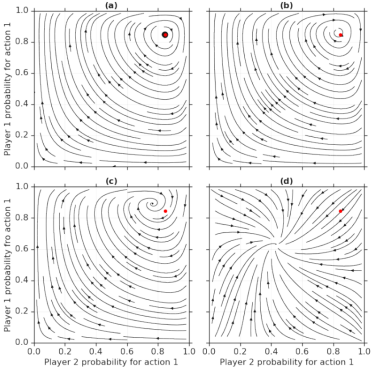


Figure: Perolat et al., ICML 2021, Fig. 1.

Biased matching pennies, payoffs $\left[\begin{smallmatrix} 1 & -1 \\ -1 & 10 \end{smallmatrix}\right]$. FoReL is "follow the regularized leader," the continuous-time FTRL of M3.

- (a) FoReL: orbits around the equilibrium (red); no convergence.
- (b) to (d): a regularized reward, at strengths 0.5, 1, 10. Trajectories converge, to a point that drifts from the red dot.
- Average-iterate guarantees "are not conducive to learning in large games, which rely on ... deep neural networks that are inherently difficult to time-average."

Source: Perolat et al. (2021), §1 and Fig. 1 caption.

R-NaD: Transform, Converge, Move the Reference

Part VI ▷ Perolat et al., Science 2022, §2.2 and Fig. 2

1. Reward transformation. With reference policy π_{reg} and strength $\alpha > 0$ (the paper's η), player i receives

$$r_i(a^i, a^{-i}) - \alpha \log \frac{\pi^i(a^i)}{\pi^i_{\text{reg}}(a^i)} + \alpha \log \frac{\pi^{-i}(a^{-i})}{\pi^{-i}_{\text{reg}}(a^{-i})}.$$

The two extra terms cancel across players, so the game stays zero-sum.

2. Dynamics. Run replicator dynamics on this game; it reaches a unique fixed point π_{fix} (Lyapunov argument).

3. Update. Set $\pi_{\text{reg}} \leftarrow \pi_{\text{fix}}$; repeat.

The fixed points "can be proven to converge to a Nash equilibrium of the original (unmodified) game."

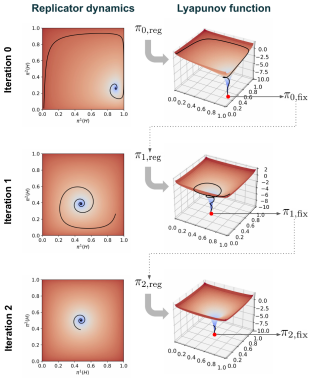


Figure: Perolat et al., Science 2022, Fig. 2c (arXiv p. 8).

Source: Perolat et al. (2022), arXiv:2206.15378, §2.2, pp. 7 to 9.

What Is Proven, and What DeepNash Achieved

Part VI ▷ Theory: ICML 2021, Thm 6.1. Practice: Science 2022

Thm 6.1 (Perolat et al. 2021). In a *monotone* game whose Nash equilibria are all *interior*, the sequence of reference policies converges to a Nash equilibrium. Inner steps are analyzed in *continuous time*. Nothing is proven for the deep implementation.

- *Monotone*: a condition on paired deviations (their Definition 3.2) that 2p0s games satisfy. *Interior*: every action has positive probability. The authors only conjecture the non-interior case.
- DeepNash = R-NaD with NeuRD updates in a deep network: "NeuRD combined with the regularization idea ... to converge in the last iterate." Each reference update uses an *approximate* fixed point after a fixed number of steps.
- Stratego: over 97% against existing bots; 84% of 50 ranked Gravon games (April 2022), top-3 yearly and all-time.

Source: Perolat et al. (2021) §§3.1, 5, 6, Thm 6.1 and remark; Perolat et al. (2022) pp. 5, 11 to 12, 24.

Part VII: Magnetic Mirror Descent

M3's mirror descent, plus a pull toward a magnet

Magnetic Mirror Descent (Sokota et al., ICLR 2023)

Part VII ▷ Co-authors include Noam Brown and Christian Kroer

Policy π_t, action values q_t, stepsize η, regularization temperature α, and a *magnet* policy ρ:

$$\pi_{t+1} = \arg\max_{\pi} \; \mathbb{E}_{a\sim\pi}\, q_t(a) - \alpha\, D_{\mathrm{KL}}(\pi\|\rho) - \tfrac{1}{\eta} D_{\mathrm{KL}}(\pi\|\pi_t) \quad\Longrightarrow\quad \pi_{t+1} \propto \left[\pi_t\, \rho^{\alpha\eta}\, e^{\eta q_t}\right]^{1/(1+\alpha\eta)}.$$

- $\alpha = 0$: the closed form is $\pi_{t+1} \propto \pi_t e^{\eta q_t}$, which is Hedge, that is, entropic mirror descent (M3).
- Fixed point: $\pi \propto \rho\, e^{q/\alpha}$. With ρ uniform, this is the *logit quantal response equilibrium* (QRE): each player plays a softmax of its action values at temperature α.
- **Thm 3.4:** let L be the game's smoothness constant ($\max_{ij} |M_{ij}|$ for a matrix game). If $\eta \le \alpha/L^2$, the Bregman distance to the solution (here the KL divergence) shrinks by a factor $1/(1+\eta\alpha)$ per step: linear, *last-iterate* convergence to the QRE.
- The same update with the Euclidean distance in place of KL and $\alpha = 0$ is projected gradient descent ascent, which "is known to diverge or cycle." So $\alpha > 0$ is what buys convergence.

Source: Sokota et al. (2023), §§2.3, 3.1 to 3.3, Prop. 2.4, eqs. (10), (12), Thm 3.4.

MMD on a Perturbed Rock-Paper-Scissors

Part VII ▷ Sokota et al. 2023, Fig. 4

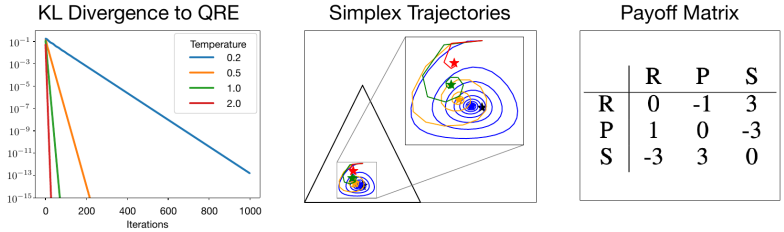


Figure: Sokota et al., ICLR 2023, Fig. 4 (appendix D.4).

Straight lines on a log scale: linear convergence of the last iterate. Higher temperature converges faster, to a QRE farther from Nash; Ataraxos manages this trade-off by annealing.

Pause and predict. If α is driven to 0 while η stays fixed, which guarantee from the previous slide do you lose first?

Ataraxos (Nature, 2026): MMD-Style Self-Play, Plus Search

- "The core innovation ... is a coordination between regularization strength and policy update size" (citing MMD). Strong regularization and large updates early; weak and small late, annealed by power laws.
- Regularization: a reverse-KL penalty toward a policy that picks a movable piece, then a legal move, uniformly at random. Update size: a reverse-KL penalty to the data-collection policy, ratio clipping, gradient clipping, and the Adam learning rate.
- A belief network samples the hidden pieces.
- At test time: rollouts, then "a tabular step of magnetic mirror descent" at the current decision only.

So Ataraxos is *not* plain MMD. It is MMD-style damped self-play with a belief model and a one-step MMD search.

Source: Sokota et al. (2026), Nature, doi:10.1038/s41586-026-11036-y; arXiv:2511.07312v2 pp. 4, 13, 15.

Ataraxos Against the Most Decorated Stratego Player

Part VII ▷ Sokota et al. 2026, Fig. 3

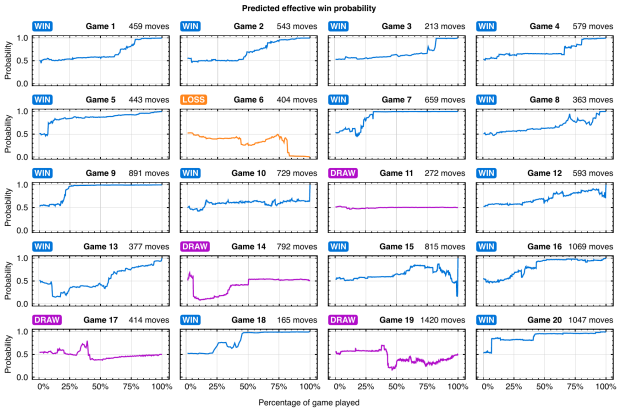


Figure: Sokota et al., Nature 2026, Fig. 3 (arXiv v2 p. 6).

Source: Sokota et al. (2026) arXiv v2 pp. 4 to 6, 15, 25 (cost section).

- 20 games against Pim Niemeijer: 15 wins, 1 loss, 4 draws.
- Training: 16 H100s for 1 week, plus 4 H100s for 4 days. Under $8,000 at 2025 prices.
- DeepNash, by the authors' estimate: $3M to $4.5M; about 5.5 billion games, versus about 160 million.

Part VIII: The 2026 Verdict

Exact exploitability, 7,000 runs, and an answer to the question

Rudolph et al. (ICLR 2026): Generic Policy Gradient Holds Up

Part VIII ▷ Exact exploitability on five large games

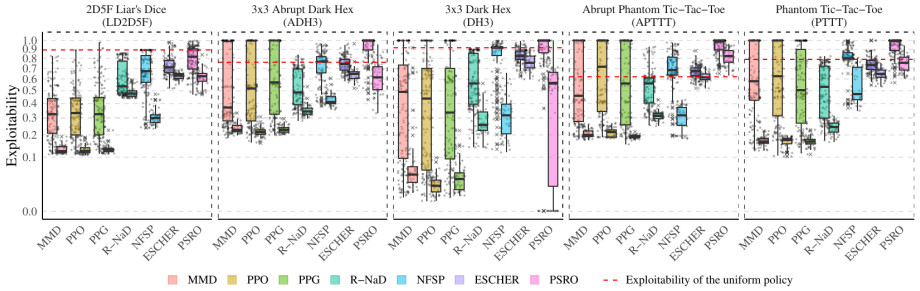


Figure: Rudolph et al., ICLR 2026, Fig. 2.

7,000 runs: "NFSP, PSRO, ESCHER, and R-NaD fail to outperform the generic PG methods (MMD, PPO, PPG)." Limits they state: five games, one variant per family, their tuning, 10 million steps.

Source: Rudolph et al. (2026), abstract, §1, §5.3, §6, Fig. 2 caption; venue on the ICLR 2026 proceedings page.

Theory Is Catching Up: Last-Iterate Policy Gradient

Part VIII ▷ Kalogiannis and Farina, NeurIPS 2025 (abstract)

- Setting: imperfect-information, perfect-recall, two-player zero-sum extensive-form games.
- *Regularized alternating* policy gradient, with direct, softmax, or natural-gradient softmax parametrization, converges to an approximate Nash equilibrium in the *last iterate*.
- Iterations and samples polynomial in $1/\epsilon$ and the game's parameters.
- The key fact: utilities are concave in the sequence form (M4). That gives *gradient dominance* in the behavioral policy: a small gradient implies a near-optimal policy, so gradient steps cannot stall at a bad point.

Same ingredients as MMD and R-NaD: regularize, control the step, and the current policy converges. These are tabular results; no theorem yet covers a deep network.

Source: Kalogiannis and Farina (2025), NeurIPS proceedings abstract (full text not read).

Closing Table: What Converges, and Under What Assumptions

Part VIII ▷ Today's methods side by side; BR = best response

Method	Averages or regularizes	Converging iterate	Rate	Assumptions	Deep-scale evidence
FP, XFP	averages best responses	average	bound $t^{-1/(r+s-2)}$; $t^{-1/n}$ on I_n	exact BR; ties adversarial for $t^{-1/n}$	none directly
FSP, NFSP	learned average Π	average	none	$\epsilon_t \to 0$	NFSP: Leduc, LHE
PFSP	opponent weights	none claimed	none	none stated	AlphaStar ablation
Hedge, NeuRD	averages	average	$O(\sqrt{\ln n / T})$	tabular, all-actions	NeuRD in DeepNash
R-NaD	KL to π_{reg}, then moved	last	exp. inner, cont. time	monotone; interior NE	DeepNash, Stratego
MMD	KL to magnet ρ	last	linear, to QRE	monotone; $\eta \leq \alpha / L^2$	Ataraxos (plus search)
Reg. PG	regularized utility	last	poly$(1/\epsilon)$	perfect-recall 2p0s	PPO, MMD (Rudolph)

Source: Previous slides; "Reg. PG" row: Kalogiannis and Farina (2025) and Rudolph et al. (2026).

So: Why PFSP, and Why Not MWU?

Part VIII ▷ The answer, in three parts

1. **PFSP picks opponents; it is not the learner.** Its $p \to \infty$ limit trains against the hardest opponents in the pool, a max-min target. It has strong ablation evidence and no published convergence theory.

2. **In 2019 the deep learner could not cheaply be MWU.** Softmax policy gradient is not Hedge (NeuRD), and Hedge's guarantee is for an average that costs a second network to approximate.

3. **The field did move to M3's toolkit,** in a form that makes the *current* policy converge: regularize toward a reference or magnet and control the step (R-NaD, MMD, Ataraxos). In 2026 tuned generic PG matches or beats the FP-, PSRO-, and CFR-based deep methods.

Discussion Questions

Use the 5-minute buffer for one; the others make good project or reading-response prompts.

1. **So why did PFSP work for AlphaStar, and pure self-play for Pluto?** Use both axes. What does each system's update rule do that the other's opponent rule might not need to?

2. **Is the league a meta-solver?** AlphaStar mixes 35% self-play with 50% PFSP and exploiters. Write it as a PSRO meta-solver. Which guarantee from Part IV could you borrow, and what would it cost?

3. **Magnets for LLM agents.** MMD regularizes toward a magnet ρ. In RL fine-tuning of an LLM, what plays the role of ρ, and what does annealing α correspond to?

Where This Is Heading (One Slide, Not Examined)

- **Gravel** (Mohammadpour, Sokota, Kaplowitz, Kolter, Brown, Farina; NeurIPS 2026, *abstract only*): public-belief-state search that "achieves game-theoretic soundness via regularization"; reported to beat ReBeL against Slumbot and in endgame solving.
- **Test-time RL** (Kubíček, Lisý, Sandholm; arXiv:2608.30635, *preprint*): subgame solving implemented "implicitly by modified sampling and neural policy," with regularized policy gradient.
- **Refinements** (same authors; arXiv:2601.17131, *preprint*): gadget games "typically contain infinitely many Nash equilibria"; choosing refined ones lowers exploitability.

What comes next. Per the syllabus, this is the last new-material lecture; Weeks 13 and 14 are for PA5 and projects. Reading: Hennes et al. (2020) §3; Perolat et al. (2022) §2; Rudolph et al. (2026) §§1 and 6.

Source: NeurIPS 2026 poster page (abstract); Kubíček et al. (2026a, 2026b), abstracts; syllabus-f26, Weeks 13 to 14.

References i

- Abernethy, J., Lai, K. A. & Wibisono, A. (2021). *Fast Convergence of Fictitious Play for Diagonal Payoff Matrices.* SIAM proceedings, doi:10.1137/1.9781611976465.84; arXiv:1911.08418. [peer-reviewed; specific SIAM venue not confirmed]
- Berner, C., Brockman, G., et al. (2019). *Dota 2 with Large Scale Deep Reinforcement Learning.* arXiv:1912.06680. [preprint]
- Brandt, F., Fischer, F. & Harrenstein, P. (2013). *On the Rate of Convergence of Fictitious Play.* Preprint submitted to Theory of Computing Systems (version of Feb. 19, 2013). [preprint as opened]
- Brown, N., Lerer, A., Gross, S. & Sandholm, T. (2019). *Deep Counterfactual Regret Minimization.* ICML 2019, PMLR 97; arXiv:1811.00164v3. [peer-reviewed]
- Churchill, D. (2026). *2026 CoG StarCraft AI Competition Report.* davechurchill.ca/starcraft/cog/results/2026/report/. [competition report; figure source]
- Daskalakis, C. & Pan, Q. (2014). *A Counter-Example to Karlin's Strong Conjecture for Fictitious Play.* FOCS 2014, pp. 11-20; arXiv:1412.4840. [peer-reviewed]
- Dinh, L. C., et al. (2022). *Online Double Oracle.* Transactions on Machine Learning Research, 10/2022; arXiv:2103.07780. [peer-reviewed]
- Heinrich, J., Lanctot, M. & Silver, D. (2015). *Fictitious Self-Play in Extensive-Form Games.* ICML 2015, PMLR 37. [peer-reviewed]
- Heinrich, J. & Silver, D. (2016). *Deep Reinforcement Learning from Self-Play in Imperfect-Information Games.* arXiv:1603.01121v2. [preprint] (NFSP)
- Hennes, D., Morrill, D., Omidshafiei, S., et al. (2020). *Neural Replicator Dynamics: Multiagent Learning via Hedging Policy Gradients.* AAMAS 2020, p. 492; arXiv:1906.00190. [peer-reviewed]
- Kalogiannis, F. & Farina, G. (2025). *Policy Gradient Methods Converge Globally in Imperfect-Information Extensive-Form Games.* NeurIPS 2025. [peer-reviewed; abstract read]
- Kubíček, O., Lisý, V. & Sandholm, T. (2026a). *Equilibrium Refinements Improve Subgame Solving in Imperfect-Information Games.* arXiv:2601.17131. [preprint]

References ii

- Kubíček, O., Lisý, V. & Sandholm, T. (2026b). *Test-time Reinforcement Learning in Imperfect Information Games*. arXiv:2608.30635. [preprint]
- Lanctot, M., Zambaldi, V., et al. (2017). *A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning*. NIPS 2017; arXiv:1711.00832. [peer-reviewed] (PSRO)
- McAleer, S., Lanier, J. B., Wang, K. A., Baldi, P., Sandholm, T. & Fox, R. (2024). *Toward Optimal Policy Population Growth in Two-Player Zero-Sum Games*. ICLR 2024. [peer-reviewed]
- Mella, V. (2026). *Pluto: StarCraft: Brood War AI, CoG 2026 entry* (PLUTO_README.txt, linked from the competition report). [author documentation, not peer-reviewed]
- Mohammadpour, S., Sokota, S., Kaplowitz, B., Kolter, J. Z., Brown, N. & Farina, G. (2026). *A Regularization-Based Approach to Public Belief State Search for Adversarial Games* (Gravel). NeurIPS 2026, to appear. [abstract only]
- Perolat, J., Munos, R., Lespiau, J.-B., et al. (2021). *From Poincaré Recurrence to Convergence in Imperfect Information Games: Finding Equilibrium via Regularization*. ICML 2021, PMLR 139: 8525-8535; arXiv:2002.08456. [peer-reviewed]
- Perolat, J., De Vylder, B., et al. (2022). *Mastering the game of Stratego with model-free multiagent reinforcement learning* (DeepNash). Science 378(6623): 990-996; arXiv:2206.15378 used. [peer-reviewed]
- Rudolph, M., Lichtlé, N., Mohammadpour, S., Bayen, A., Kolter, J. Z., Zhang, A., Farina, G., Vinitsky, E. & Sokota, S. (2026). *Reevaluating Policy Gradient Methods for Imperfect-Information Games*. ICLR 2026; arXiv:2502.08938v4. [peer-reviewed]
- Shapiro, H. N. (1958). *Note on a computation method in the theory of games*. Comm. Pure Appl. Math. 11: 587-593. [peer-reviewed; not opened, cited via Brandt et al.]
- Sokota, S., D'Orazio, R., Kolter, J. Z., Loizou, N., Lanctot, M., Mitliagkas, I., Brown, N. & Kroer, C. (2023). *A Unified Approach to Reinforcement Learning, Quantal Response Equilibria, and Two-Player Zero-Sum Games* (MMD). ICLR 2023; arXiv:2206.05825. [peer-reviewed]
- Sokota, S., Vinitsky, E., Hu, H., Fan, Z., Kolter, J. Z. & Farina, G. (2026). *Scalable decision-making for games of imperfect information* (Ataraxos). Nature, published Sept. 30, 2026, doi:10.1038/s41586-026-11036-y; arXiv:2511.07312v2 used. [peer-reviewed]

References iii

- Vinyals, O., Babuschkin, I., Czarnecki, W. M., et al. (2019). *Grandmaster level in StarCraft II using multi-agent reinforcement learning.* Nature 575: 350-354. Author manuscript used. [peer-reviewed]