Learning Equilibria via Regret Minimization in Normal-Form and Extensive-Form Games

Amy Greenwald

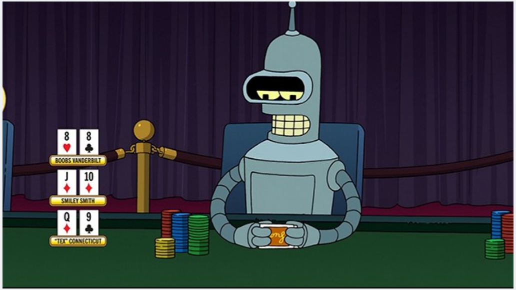
Joint work with: Michael Bowling, Ryan D'Orazio, Marc Lanctot, **Dustin Morrill**, Reca Sarfati, and James R. Wright

May 10, 2022

AI has been learning to play games (well!) for decades

Recent AI Poker Successes

1. Cepheus (heads-up limit)
2. DeepStack (heads-up no-limit)
3. Libratus (heads-up no-limit)
4. Pluribus (six-player no-limit)

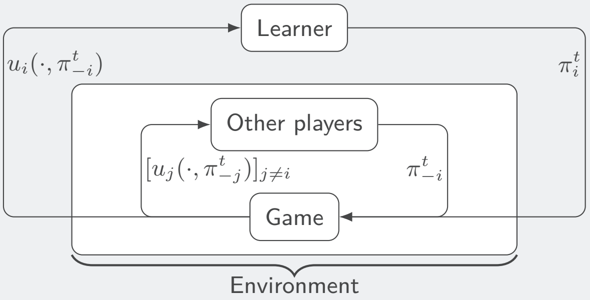


Counterfactual Regret Minimization (CFR) is key to all these poker successes!

[1] [2] [3] [4] [5]

[1] Bowling et al., “Heads-Up Limit Hold’em Poker is Solved”.
[2] Moravčík et al., “DeepStack: Expert-Level Artificial Intelligence in Heads-Up No-Limit Poker”.
[3] Brown and Sandholm, “Superhuman AI for Heads-Up No-Limit Poker: Libratus Beats Top Professionals”.
[4] Brown and Sandholm, “Superhuman AI for Multiplayer Poker”.
[5] Zinkevich et al., “Regret Minimization in Games with Incomplete Information”.

Learning Model



Hindsight Rationality (Regret Minimization)

$$\begin{array}{lllllll} \text{Learner} & : \pi_i^1 & \pi_i^2 & \cdots & \pi_i^T & \rightarrow \frac{1}{T}\sum_{t=1}^T u_i(\pi_i^t, \pi_{-i}^t) \\ \text{Deviation} & : \phi(\pi_i^1) & \phi(\pi_i^2) & \cdots & \phi(\pi_i^T) & \rightarrow \frac{1}{T}\sum_{t=1}^T u_i\big(\phi(\pi_i^t), \pi_{-i}^t\big) \end{array}$$

$$\text{Objective:} \quad \underbrace{\frac{1}{T}\sum_{t=1}^{T} u_i(\pi_i^t, \pi_{-i}^t)}_{\text{The learner's average reward.}} \geq \max_{\phi \in \Phi} \underbrace{\frac{1}{T}\sum_{t=1}^{T} u_i(\phi(\pi_i^t), \pi_{-i}^t)}_{\text{Deviation } \phi\text{'s average reward.}} - \underbrace{o(1)}_{\text{Leeway.}} .$$

Hannan, "Approximation to Bayes risk in repeated play", 1957.

Talk Outline: From Normal-Form Games to Extensive-Form Games

Normal-Form Game

Wikipedia	Chicken	Dare
Chicken	6,6	2,7
Dare	7,2	0,0

Extensive-Form Game

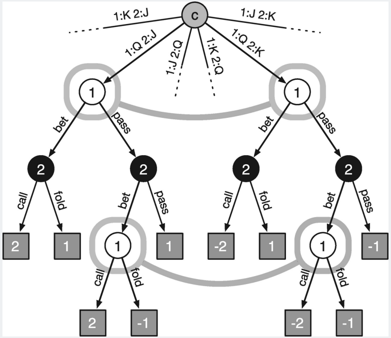


Image Source

Talk Outline: From Normal-Form Games to Extensive-Form Games

- Define internal and external regret, with examples of correlated and coarse correlated equilibria.
- Describe regret matching, a popular Φ-regret minimizing algorithm for normal-form games. Brief Interlude: Regret minimization with time-selection functions (e.g., sleeping experts).
- Define behavioral deviations, and a few notable subclasses (e.g., counterfactual, causal, action, etc.), with distinguishing examples of corresponding correlated equilibria.
- Describe EFR, a local regret-minimizing algorithm, enhanced with time-selection, where the time-selection weights depend on earlier recommendations.

No-Regret Learning in Normal-Form Games (NFGs)

Freund and Schapire, “Game Theory, Online learning, and Boosting”, 1996

- Developed an efficient no-external-regret learning algorithm.
- No-external-regret learning converges to minimax equilibrium in zero-sum NFGs (which corresponds to coarse correlated equilibrium in non-zero sum NFGs).

Foster and Vohra, “Calibrated Learning and Correlated Equilibrium”, 1997
(SIGEcom Test of Time Award)

- Developed an efficient no-internal-regret learning algorithm.
- No-internal-regret learning converges to correlated equilibrium in (non-zero-sum) NFGs.

Qualifiers: Convergence of the empirical distribution in self-play to an equilibrium set.

Deviations in Normal-Form Games

Definition

An action transformation Φ is a function $\phi : A \to A$.

Examples

$\phi_{\mathsf{EXT}}^{(a)} : x \mapsto a, \quad$ for all $x \in A$

$$\phi_{\mathsf{INT}}^{(a,b)} : x \mapsto \begin{cases} b & \text{if } x = a \\ x & \text{otherwise} \end{cases}$$

Φ_{SWAP} is the set of all n^n action transformations, where n is the number of actions.

Deviations as (Column) Stochastic Matrices

External Regret

$$[\phi]^{(2)}_{\mathsf{EXT}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \in \Phi_{\mathsf{EXT}}$$

$$[\phi]^{(2)}_{\mathsf{EXT}} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix} = \langle 0, 1, 0, 0 \rangle, \text{ for all } \boldsymbol{\pi}$$

Internal Regret

$$[\phi]^{(2,3)}_{\mathsf{INT}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \in \Phi_{\mathsf{INT}}$$

$$[\phi]^{(2,3)}_{\mathsf{INT}} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix} = \begin{pmatrix} \pi_1 \\ 0 \\ \pi_2 + \pi_3 \\ \pi_4 \end{pmatrix}, \text{ for all } \boldsymbol{\pi}$$

Deviations as (Column) Stochastic Matrices

External Regret

$$[\phi]^{(2)}_{\mathsf{EXT}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \in \Phi_{\mathsf{EXT}} \qquad [\phi]^{(2)}_{\mathsf{EXT}} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix} = \langle 0, 1, 0, 0 \rangle, \text{ for all } \boldsymbol{\pi}$$

Internal Regret

$$[\phi]^{(2,3)}_{\mathsf{INT}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \in \Phi_{\mathsf{INT}} \qquad [\phi]^{(2,3)}_{\mathsf{INT}} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix} = \begin{pmatrix} \pi_1 \\ 0 \\ \pi_2 + \pi_3 \\ \pi_4 \end{pmatrix}, \text{ for all } \boldsymbol{\pi}$$

Correlated Equilibrium [6]

Wikipedia	Chicken	Dare
Chicken	6,6	2,7
Dare	7,2	0,0

[6] Aumann, "Subjectivity and correlation in randomized strategies".

Correlated Equilibrium [6]

Wikipedia	Chicken	Dare
Chicken	6,6	2,7
Dare	7,2	0,0

A correlated equilibrium (CE) is a joint probability distribution D over the set of action profiles A s.t. for all players i, for all actions $a_i, a_i' \in A_i$,

$$\mathbb{E}_{(a_i, \boldsymbol{a}_{-i}) \sim D|_{a_i}} \left[u_i\left(a_i, \boldsymbol{a}_{-i}\right)\right] \geq \mathbb{E}_{(a_i, \boldsymbol{a}_{-i}) \sim D|_{a_i}} \left[u_i\left(a_i', \boldsymbol{a}_{-i}\right)\right]$$

[6] Aumann, "Subjectivity and correlation in randomized strategies".

Correlated Equilibrium [6]

Wikipedia	Chicken	Dare
Chicken	6,6	2,7
Dare	7,2	0,0

A correlated equilibrium (CE) is a joint probability distribution D over the set of action profiles A s.t. for all players i, for all actions $a_i, a_i' \in A_i$,

$$\mathbb{E}_{(a_i, \boldsymbol{a}_{-i}) \sim D|_{a_i}} \left[u_i \left(a_i, \boldsymbol{a}_{-i} \right) \right] \geq \mathbb{E}_{(a_i, \boldsymbol{a}_{-i}) \sim D|_{a_i}} \left[u_i \left(a_i', \boldsymbol{a}_{-i} \right) \right]$$

${}^{1}/{}_{3}$ probability on all cells with non-zero payoffs is a CE in Chicken.

Conditioned on the recommendation Chicken:

- $\mathbb{E}(\text{Chicken}) = ({}^{1}/{}_{2})(6) + ({}^{1}/{}_{2})(2) = 4$
- $\mathbb{E}(\text{Dare}) = ({}^{1}/{}_{2})(7) + ({}^{1}/{}_{2})(0) = 3.5$

Conditioned on the recommendation Dare:

- $\mathbb{E}(\text{Chicken}) = (1)(6) + (0)(2) = 6$
- $\mathbb{E}(\text{Dare}) = (1)(7) + (0)(0) = 7$

[6] Aumann, "Subjectivity and correlation in randomized strategies".

Coarse Correlated Equilibrium [7]

	a	b	c
a	1,1	-1,-1	0,0
b	-1,-1	1,1	0,0
c	0,0	0,0	-1.1, -1.1

[7] Moulin and Vial, "Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon".

Coarse Correlated Equilibrium [7]

	a	b	c
a	1,1	-1,-1	0,0
b	-1,-1	1,1	0,0
c	0,0	0,0	-1.1, -1.1

A coarse correlated equilibrium (CCE) is a joint probability distribution D over the set of action profiles A s.t. for all players i, for all actions $a_i' \in A_i$,

$$\mathbb{E}_{(a_i,\boldsymbol{a}_{-i})\sim D}\left[u_i\left(a_i,\boldsymbol{a}_{-i}\right)\right] \geq \mathbb{E}_{(a_i,\boldsymbol{a}_{-i})\sim D}\left[u_i\left(a_i',\boldsymbol{a}_{-i}\right)\right]$$

[7] Moulin and Vial, "Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon".

Coarse Correlated Equilibrium [7]

	a	b	c
a	1,1	-1,-1	0,0
b	-1,-1	1,1	0,0
c	0,0	0,0	-1.1, -1.1

A coarse correlated equilibrium (CCE) is a joint probability distribution D over the set of action profiles A s.t. for all players i, for all actions $a_i' \in A_i$,

$$\mathbb{E}_{(a_i, \boldsymbol{a}_{-i}) \sim D}\left[u_i\left(a_i, \boldsymbol{a}_{-i}\right)\right] \geq \mathbb{E}_{(a_i, \boldsymbol{a}_{-i}) \sim D}\left[u_i\left(a_i', \boldsymbol{a}_{-i}\right)\right]$$

${}^1/_3$ probability on all diagonal cells is a CCE in this game.

The expected rewards at this equilibrium are $({}^1/_3)(1) + ({}^1/_3)(1) - ({}^1/_3)(1.1) = 0.3$.

The expected rewards of playing a or b are $({}^1/_3)(1) - ({}^1/_3)(1) + ({}^1/_3)(0) = 0$.

The expected rewards of playing c are negative.

[Example borrowed from Aaron Roth's 2017 lecture notes on Correlated Equilibrium]

[7] Moulin and Vial, "Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon".

No-Regret Learning in Normal-Form Games (NFGs)

Greenwald, Jafari, and Marks, “A general class of no-regret learning algorithms and game-theoretic equilibria”, 2003

- No-internal- and no-external-regret can be defined along one continuum, no-Φ-regret.
- Efficient no-Φ-regret learning algorithms exist for NFGs, $\forall\Phi$.
- No-Φ-regret learning converges to the set of Φ-equilibria, $\forall\Phi$, with two interesting special cases:
  - No-internal-regret learning converges to correlated equilibrium.
  - No-external-regret learning converges to coarse correlated equilibrium.
- Swap regret harnesses no additional strategic power beyond internal regret.

Regret Matching Algorithm

Given Φ
Given $Y \in \mathbb{R}^{\Phi}$

Consider $Y^{+} \in \mathbb{R}^{\Phi}$

If $\sum_{\phi \in \Phi} Y_{\phi}^{+} = 0$, play arbitrarily
If $\sum_{\phi \in \Phi} Y_{\phi}^{+} > 0$, define stochastic matrix

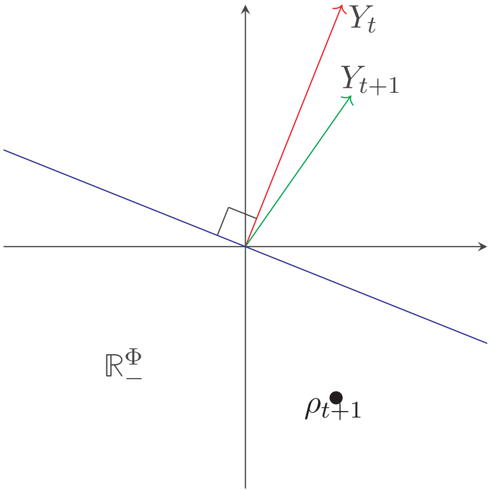
$$A \equiv A(\Phi, Y^{+}) = \frac{\sum_{\phi \in \Phi} [\phi] Y_{\phi}^{+}}{\sum_{\phi \in \Phi} Y_{\phi}^{+}}$$

play mixed strategy $A\pi = \pi$

Regret Matching Theorem

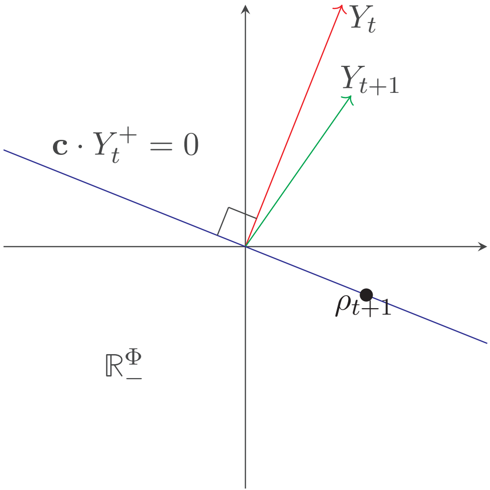
Regret matching satisfies Blackwell's approachability condition: $\rho(r, \pi) \cdot Y^{+} = 0$

Blackwell's Approachability Theorem



Blackwell, "An analog of the minimax theorem for vector payoffs", 1956.

Blackwell's Approachability Theorem



Blackwell, "An analog of the minimax theorem for vector payoffs", 1956.

$$
\begin{aligned}
\rho(r,\pi)\cdot Y^{+} &= \sum_{\phi\in\Phi} \rho_{\phi}(r,\pi) Y_{\phi}^{+} \\
&= \sum_{\phi\in\Phi} (r\cdot[\phi]\pi - r\cdot\pi) Y_{\phi}^{+} \\
&= \sum_{\phi\in\Phi} r\cdot([\phi]\pi Y_{\phi}^{+} - \pi Y_{\phi}^{+}) \\
&= r\cdot\left(\left(\sum_{\phi\in\Phi}[\phi]Y_{\phi}^{+}\right)\pi - \left(\sum_{\phi\in\Phi} Y_{\phi}^{+}\right)\pi\right) \\
&= \left(\sum_{\phi\in\Phi} Y_{\phi}^{+}\right) r\cdot\left(\left(\frac{\sum_{\phi\in\Phi}[\phi]Y_{\phi}^{+}}{\sum_{\phi\in\Phi} Y_{\phi}^{+}}\right)\pi - \pi\right) \\
&= \left(\sum_{\phi\in\Phi} Y_{\phi}^{+}\right) r\cdot(A\pi - \pi) \\
&= \left(\sum_{\phi\in\Phi} Y_{\phi}^{+}\right) r\cdot(\pi - \pi) \\
&= 0
\end{aligned}
$$

Time-Selection Regret Minimization

$$
W \left\{
\begin{array}{lllllll l}
w_1 & : 1 & 1 & 1 & 1 & \cdots & 1 & \to \frac{1}{T}\sum_{t=1}^{T} u_i\left(\cdot, \pi_{-i}^{t}\right) \\
w_2 & : 0 & 1 & 0 & 1 & \cdots & 0 & \to \frac{1}{T}\sum_{t=1}^{T/2} u_i\left(\cdot, \pi_{-i}^{2t}\right) \\
w_3 & : 1 & 1/2 & 1/3 & 1/4 & \cdots & 1/T & \to \frac{1}{T}\sum_{t=1}^{T} \frac{1}{t} u_i\left(\cdot, \pi_{-i}^{t}\right) \\
\cdots & & & & & & & \\
w_m & : w_m^1 & w_m^2 & w_m^3 & w_m^4 & \cdots & w_m^T & \to \frac{1}{T}\sum_{t=1}^{T} w_m^t u_i\left(\cdot, \pi_{-i}^{t}\right)
\end{array}
\right.
$$

Objective: $\forall w \in W,\ \underbrace{\frac{1}{T}\sum_{t=1}^{T} w^t u_i(\pi_i^t, \pi_{-i}^t)}_{\text{The learner's average reward.}} \geq \max_{\phi \in \Phi} \underbrace{\frac{1}{T}\sum_{t=1}^{T} w^t u_i(\phi(\pi_i^t), \pi_{-i}^t)}_{\text{Deviation } \phi\text{'s average reward.}} - \underbrace{o(1)}_{\text{Leeway.}}$.

Freund et al., "Using and combining predictors that specialize", 1997.

Blum and Mansour, "From external to internal regret", 2007.

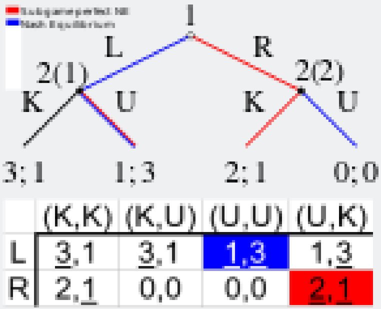
$$
\begin{aligned}
\sum_{w\in W} w\rho(r,\pi)\cdot Y^{+}(w) &= \sum_{w\in W} w \sum_{\phi\in\Phi} \rho_\phi(r,\pi) Y^{+}_{\phi}(w) \\
&= \sum_{\phi\in\Phi} (r\cdot[\phi]\pi - r\cdot\pi) \sum_{w\in W} wY^{+}_{\phi}(w) \\
&= \sum_{\phi\in\Phi} r\cdot\left([\phi]\pi \sum_{w\in W} wY^{+}_{\phi}(w) - \pi \sum_{w\in W} wY^{+}_{\phi}(w)\right) \\
&= r\cdot\left(\left(\sum_{\phi\in\Phi}[\phi] \sum_{w\in W} wY^{+}_{\phi}(w)\right)\pi - \left(\sum_{\phi\in\Phi}\sum_{w\in W} wY^{+}_{\phi}(w)\right)\pi\right) \\
&= \left(\sum_{\phi\in\Phi}\sum_{w\in W} wY^{+}_{\phi}(w)\right) r\cdot\left(\left(\frac{\sum_{\phi\in\Phi}[\phi]\sum_{w\in W} wY^{+}_{\phi}(w)}{\sum_{\phi\in\Phi}\sum_{w\in W} wY^{+}_{\phi}(w)}\right)\pi - \pi\right) \\
&= \left(\sum_{\phi\in\Phi}\sum_{w\in W} wY^{+}_{\phi}(w)\right) r\cdot(A\pi - \pi) \\
&= \left(\sum_{\phi\in\Phi}\sum_{w\in W} wY^{+}_{\phi}(w)\right) r\cdot(\pi - \pi) \\
&= 0
\end{aligned}
$$

Talk Outline: From Normal-Form Games to Extensive-Form Games

Normal-Form Game

Wikipedia	Chicken	Dare
Chicken	6,6	2,7
Dare	7,2	0,0

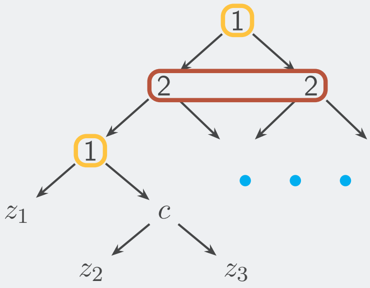
Extensive-Form Game



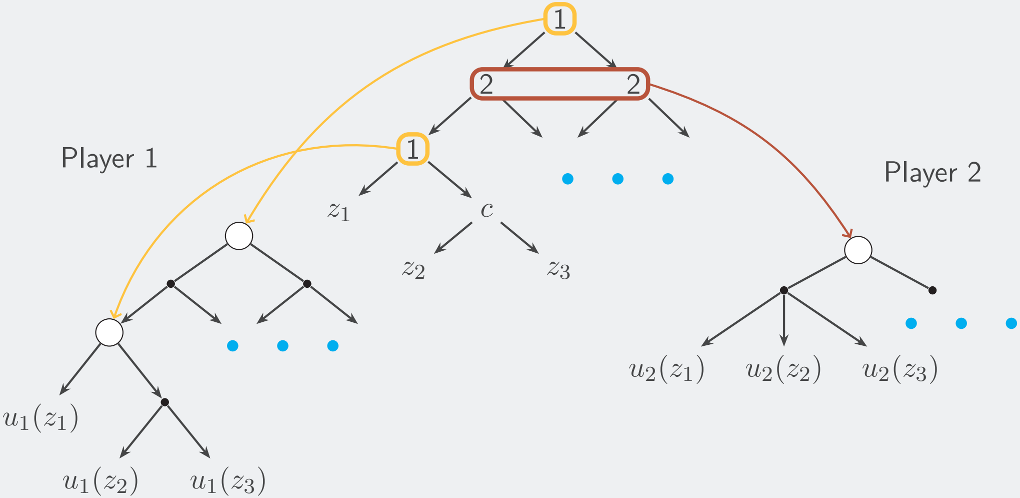
	(K,K)	(K,U)	(U,U)	(U,K)
L	3,1	3,1	1,3	1,3
R	2,1	0,0	0,0	2,1

Image Source

Extensive-Form Game Trees

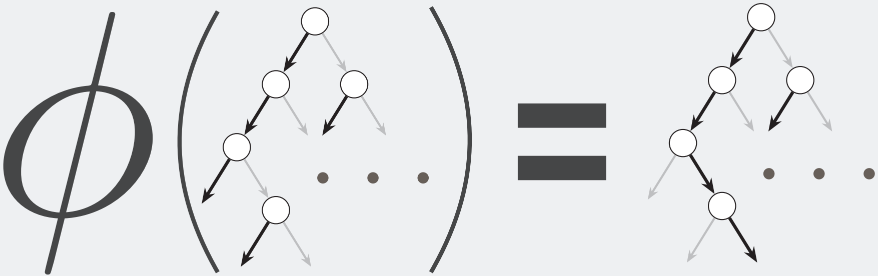


Partially Observable History Processes (POHPs)



[7] Morrill, Greenwald, and Bowling, "The Partially Observable History Process".

Deviations in Extensive-Form Games



32^{32} swap deviations!

32^{2} internal (32 external) deviations.

No-Regret Learning in Zero-sum Extensive-Form Games (EFGs)

Zinkevich et al., "Regret Minimization in Games with Incomplete Information", 2007

- Efficient no-external-regret learning algorithms exist for EFGs, namely counterfactual regret minimization (CFR).
- Play no-external-regret algorithms locally: i.e., at all agent states (information sets), using long-run counterfactual values (calculated with particular weights).
- No-external-regret learning converges to minimax equilibrium in zero-sum EFGs.

Counterfactual Regret Minimization (CFR) is key to all the poker successes!

Deviations in Extensive-Form Games

von Stengel & Forges [8] (2008) proposed two restricted deviation classes for EFGs

1. Behavioral deviations: Recommendations at an information set can only depend on observations up to and including that information set. They cannot depend on recommendations off the recommended path of play, or at later information sets.
2. Reduced strategies: No recommmendations are made at information sets off the recommended path of play, so after deviating there are no further recommendations.

Behavioral deviations define behavioral correlated equilibrium (BCE).
Behavioral deviations with reduced strategies define (causal) EFCE.

[8] von Stengel and Forges, "Extensive-form correlated equilibrium: Definition and computational complexity".

Deviations in Extensive-Form Games

von Stengel & Forges [8] (2008) proposed two restricted deviation classes for EFGs

1. Behavioral deviations: Recommendations at an information set can only depend on observations up to and including that information set. They cannot depend on recommendations off the recommended path of play, or at later information sets.
2. Reduced strategies: No recommmendations are made at information sets off the recommended path of play, so after deviating there are no further recommendations.

Celli, et al. [9] (2020) developed a learning algorithm called ICFR that minimizes causal regret, and hence converges to the set of (causal) EFCE. (NeurIPS best paper award, 2020)

[8] von Stengel and Forges, "Extensive-form correlated equilibrium: Definition and computational complexity".
[9] Celli et al., "No-regret learning dynamics for extensive-form correlated equilibrium".

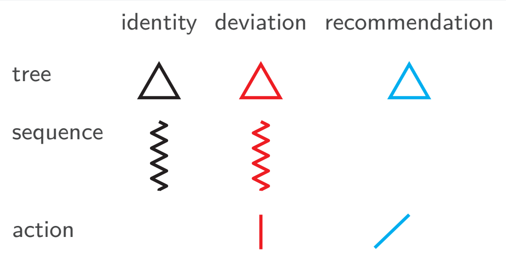
No-Regret Learning in Non-zero-sum Extensive-Form Games (EFGs)

Morrill, D'Orazio, Sarfati, et al., "Hindsight and sequential rationality of correlated play"
Morrill, D'Orazio, Lanctot, et al., "Efficient Deviation Types and Learning for Hindsight Rationality in Extensive-Form Games", 2021

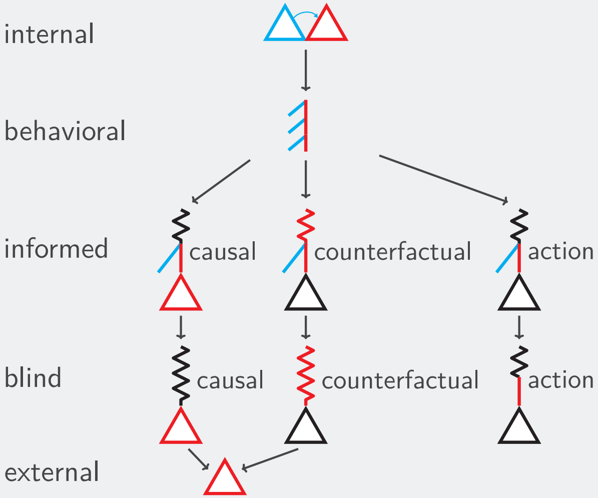
- No-internal-regret learning converges to correlated equilibrium.
  No-external-regret learning converges to coarse correlated equilibrium.
- No-internal- and no-external-regret can be defined along one continuum, no-Φ-regret.
- Efficient no-Φ-regret learning algorithms exist for EFGs, namely extensive-form regret minimization (EFR), for certain choices of Φ in the class of behavioral deviations.
- EFR generalizes CFR: choose Φ to be the set of counterfactual deviations.
- EFR generalizes ICFR: choose Φ to be the set of causal deviations.

EFR opens the door to efficient no-regret learning in non-zero-sum EFGs.

Basic Behavioral Deviations in EFGs



type	blind	informed
internal	-	$\mathcal{O}(n^{2\lvert\mathcal{I}\rvert})$
behavioral	-	$\mathcal{O}(n^{d+2}\lvert\mathcal{I}\rvert)$
causal	$\mathcal{O}(n^{\lvert\mathcal{I}\rvert}\lvert\mathcal{I}\rvert)$	$\mathcal{O}(n^{\lvert\mathcal{I}\rvert+1}\lvert\mathcal{I}\rvert)$
CF	$\mathcal{O}(n\lvert\mathcal{I}\rvert)$	$\mathcal{O}(n^2\lvert\mathcal{I}\rvert)$
action	$\mathcal{O}(n\lvert\mathcal{I}\rvert)$	$\mathcal{O}(n^2\lvert\mathcal{I}\rvert)$
external	$\mathcal{O}(n^{\lvert\mathcal{I}\rvert})$	-



Equilibrium Relationships

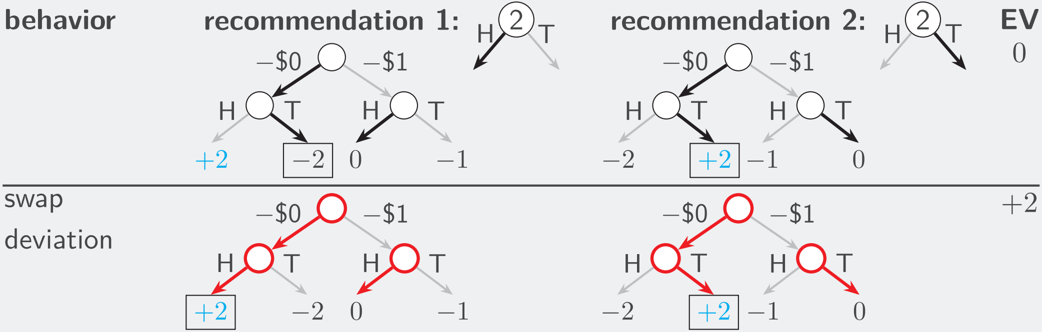
$\subseteq$	CE	CCE	EF	AF	CF
CE	$=$	$\doteq$	$\doteq$	$\doteq$	$\doteq$
EF	B	$\doteq$	$=$	*B*	**B**
AF	I	I	I	$=$	**I**
CF	**M**	$\doteq$	**M**	**M**	$=$
CCE	M	$=$	M	*B*	**B**

- Cyan cells show where the row concept implies the column concept. *E.g.*, an EF(C)CE is also a CCE.
- Red cells indicate that the subset relationship does not hold. *E.g.*, an EF(C)CE may not be an AFCCE.
- Letters refer to game examples.

Example: BCE that is not a CE

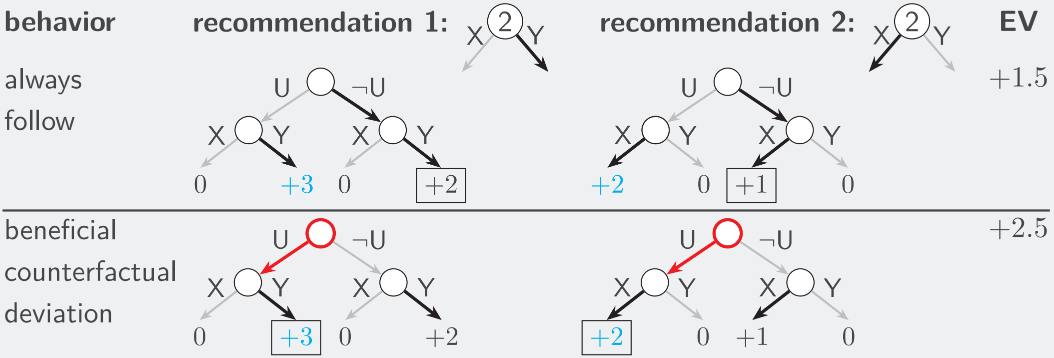
Matching Pennies (−\$0)	H	T
H	2	-2
T	-2	2

Matching Pennies (−\$1)	H	T
H	0	-1
T	-1	0



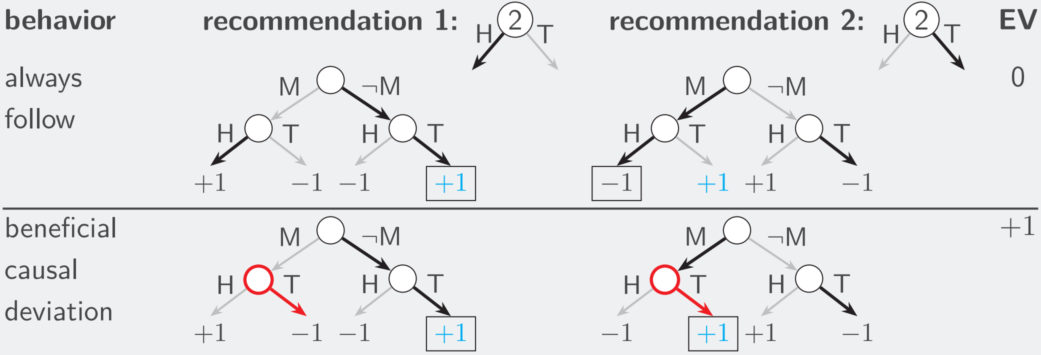
Example: Causal CE, but not a counterfactual, action, or behavioral CCE

Battle of the Sexes	X	Y
X	1,2	0,0
Y	0,0	2,1



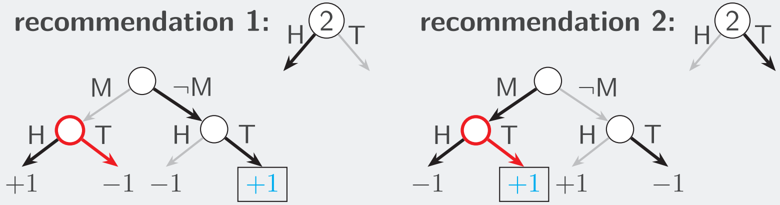
Example: Counterfactual CE, but not a causal, action, or behavioral CCE

Matching Pennies	H	T
H	1	-1
T	-1	1



Matching Pennies (CFCE, but not EFCCE)

beneficial causal deviation



	$*,\phi^{\to \mathrm{M}}$	M,ϕ^{id}	¬M,$\phi^{\neg\mathrm{M}\to\mathrm{M}}$
recommendation 1	1	0	1
recommendation 2	1	1	0

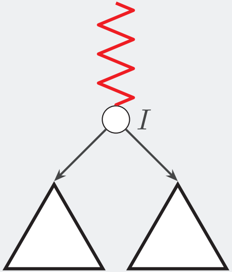
recommendation 1

	ρ_{CF}	M,ϕ^{id}	¬M,$\phi^{\neg\mathrm{M}\to\mathrm{M}}$
H $\to$ T	-2	0	-2
T $\to$ H	0	0	0

recommendation 2

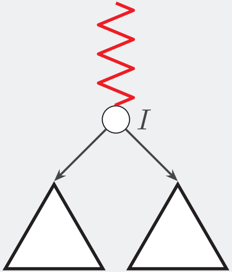
	ρ_{CF}	M,ϕ^{id}	¬M,$\phi^{\neg\mathrm{M}\to\mathrm{M}}$
H $\to$ T	2	2	0
T $\to$ H	0	0	0

Counterfactual Regret Minimization (CFR)



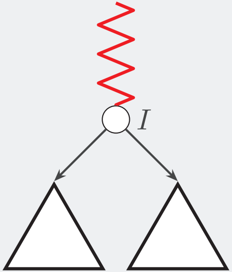
CFR works by learning $\pi_i^t(I) \in \Delta^{|\mathcal{A}(I)|}$.

Counterfactual Regret Minimization (CFR)



What is the regret at I?

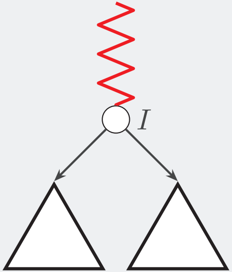
Counterfactual Regret Minimization (CFR)



What is the regret at I?

What is the value of each action a under i's current strategy, π_i^t?

Counterfactual Regret Minimization (CFR)

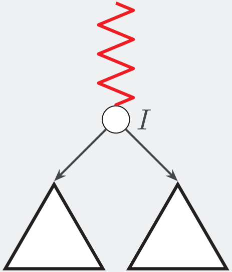


What is the regret at I?

$\forall a,\ \underbrace{v_I(a;\pi^t)}$

Counterfactual value, meaning the expected payoff of a under policy π, weighted by the probability of reaching I, assuming i deviates to I: *i.e.*, weighted by the probability players other than i play to I.

Counterfactual Regret Minimization (CFR)

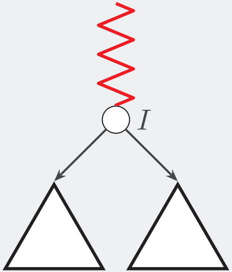


$$\underbrace{v_I([\phi_I \pi_i^t](I); \pi^t) - v_I(\pi_i^t(I); \pi^t)}_{\text{Counterfactual regret, } \rho_I^{\mathrm{CF}}(\phi_I; \pi^t).}$$

$$\forall a,\ \underbrace{v_I(a; \pi^t)}$$

Counterfactual value, meaning the expected payoff of a under policy π, weighted by the probability of reaching I, assuming i deviates to I: *i.e.*, weighted by the probability players other than i play to I.

Counterfactual Regret Minimization (CFR)



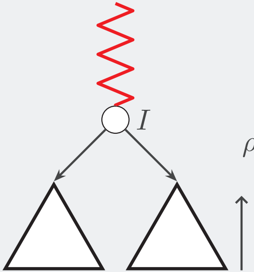
$\rho_I^{\mathrm{CF}}(\phi_I; \pi^t)$ This is a regret minimization problem! [11]

$\forall a,\ \underbrace{v_I(a; \pi^t)}$

Counterfactual value, meaning the expected payoff of a under policy π, weighted by the probability of reaching I, assuming i deviates to I: *i.e.*, weighted by the probability players other than i play to I.

[11]Zinkevich et al., "Regret Minimization in Games with Incomplete Information".

Counterfactual Regret Minimization (CFR)



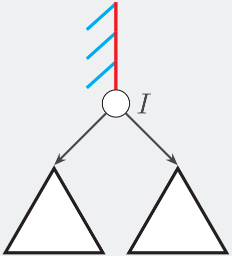
$\rho_I^{\mathrm{CF}}(\phi_I; \pi^t)$ **One solution: regret matching.** [11]

$\forall a, \underbrace{v_I(a; \pi^t)}$

Counterfactual value, meaning the expected payoff of a under policy π, weighted by the probability of reaching I, assuming i deviates to I: *i.e.*, weighted by the probability players other than i play to I.

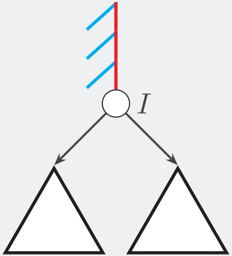
[11] Zinkevich et al., "Regret Minimization in Games with Incomplete Information".

Extensive-Form Regret Minimization (EFR)



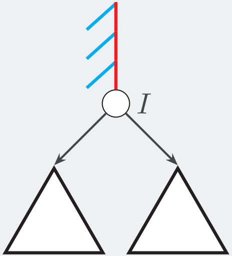
EFR works by learning $\pi_i^t(I) \in \Delta^{|\mathcal{A}(I)|}$.

Extensive-Form Regret Minimization (EFR)



What is the regret at I?

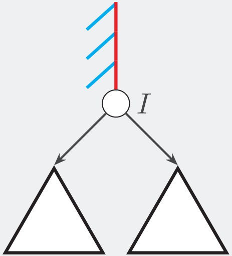
Extensive-Form Regret Minimization (EFR)



What is the regret at I?

What is the value of each action a under i's current strategy, π_i^t?

Extensive-Form Regret Minimization (EFR)

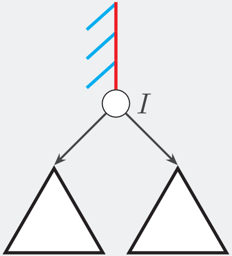


How often does each $\phi \in \Phi_{\mathcal{I}_i}^{\mathrm{BHV}}$ reach I in memory state g?

What is the regret at I?

What is the value of each action a under i's current strategy, π_i^t?

Extensive-Form Regret Minimization (EFR)



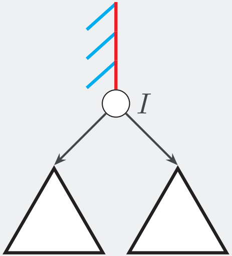
How often does each $\phi \in \Phi^{\mathrm{BHV}}_{\mathcal{I}_i}$ reach I in memory state g?

What is the regret at I?

$\forall a, \underbrace{v_I(a; \pi^t)}$

Counterfactual value, meaning the expected payoff of a under policy π, weighted by the probability of reaching I, assuming i deviates to I: *i.e.*, weighted by the probability players other than i play to I.

Extensive-Form Regret Minimization (EFR)



Memory probability,

i.e., the chance that $\phi(\pi_i^t)$ reaches I in memory state g.

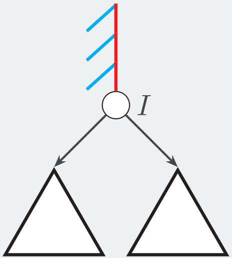
$$\forall \phi \in \Phi^{\mathrm{BHV}}_{\mathcal{I}_i}, g, \overbrace{w_\phi(I, g; \pi_i^t)} \in [0, 1]$$

What is the regret at I?

$$\forall a, \underbrace{v_I(a; \pi^t)}$$

Counterfactual value, meaning the expected payoff of a under policy π, weighted by the probability of reaching I, assuming i deviates to I: *i.e.*, weighted by the probability players other than i play to I.

Extensive-Form Regret Minimization (EFR)



Memory probability,

i.e., the chance that $\phi(\pi_i^t)$ reaches I in memory state g.

$$\forall \phi \in \Phi_{\mathcal{I}_i}^{\mathrm{BHV}}, g, \overbrace{w_\phi(I, g; \pi_i^t)} \in [0, 1]$$

$$w_\phi(I, g; \pi_i^t)\rho_I^{\mathrm{CF}}(\phi_I; \pi^t)$$

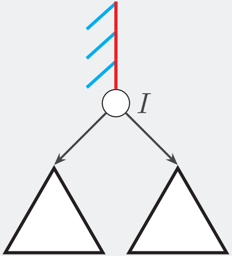
This is a time selection problem! [13]

$$\forall a, \underbrace{v_I(a; \pi^t)}$$

Counterfactual value, meaning the expected payoff of a under policy π, weighted by the probability of reaching I, assuming i deviates to I: *i.e.*, weighted by the probability players other than i play to I.

[13] Blum and Mansour, "From external to internal regret".

Extensive-Form Regret Minimization (EFR)



Memory probability,

i.e., the chance that $\phi(\pi_i^t)$ reaches I in memory state g.

$$\forall \phi \in \Phi_{\mathcal{I}_i}^{\mathrm{BHV}}, g, \overbrace{w_\phi(I, g; \pi_i^t)} \in [0, 1]$$

$$w_\phi(I, g; \pi_i^t)\rho_I^{\mathrm{CF}}(\phi_I; \pi^t)$$

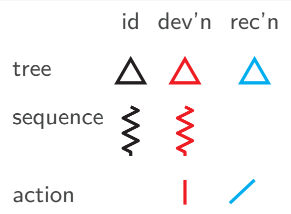
Our solution: time selection regret matching. [13]

$$\forall a, \underbrace{v_I(a; \pi^t)}$$

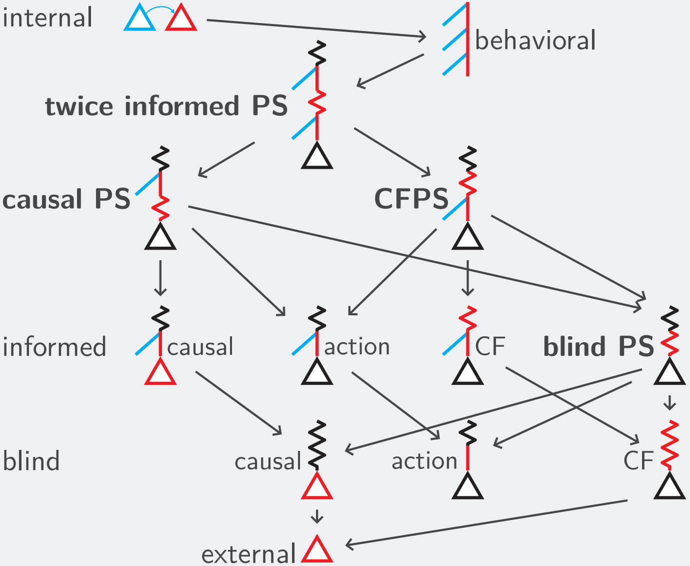
Counterfactual value, meaning the expected payoff of a under policy π, weighted by the probability of reaching I, assuming i deviates to I: *i.e.*, weighted by the probability players other than i play to I.

[13] Blum and Mansour, "From external to internal regret".

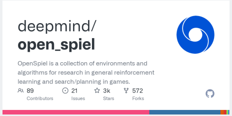
The EFR Deviation Landscape



type	# deviations
TIPS	$\mathcal{O}(dn^3\lvert\mathcal{I}\rvert)$
CSPS	$\mathcal{O}(dn^2\lvert\mathcal{I}\rvert)$
CFPS	$\mathcal{O}(dn^2\lvert\mathcal{I}\rvert)$
BPS	$\mathcal{O}(dn\lvert\mathcal{I}\rvert)$



Learning Curves



Summary

Past work:

- Defined a deviation landscape for NFGs that encompasses all known deviations.
- Identified those deviations with their corresponding correlated equilibria.
- Proved the existence of no-regret learning algorithms for all deviations, thus an algorithm that converges to all correlated equilibria in NFGs.

Summary

AAAI Paper:

- Defined a deviation landscape for EFGs that encompasses all known deviations.
- Identified those deviations with their corresponding correlated equilibria.
- Future work: prove the existence of a no-regret learning algorithm for all deviations, thus an algorithm that converges to all correlated equilibria in EFGs.

Summary

AAAI Paper:

- Defined a deviation landscape for EFGs that encompasses all known deviations.
- Identified those deviations with their corresponding correlated equilibria.
- Future work: devise a no-regret learning algorithm for all behavioral deviations, thus an algorithm that converges to all correlated EFCE in EFGs.

ICML Paper:

- Devised a no-regret learning algorithm for all behavioral deviations, thus an algorithm that converges to all correlated EFCE in EFGs.

Takeaways

- Behavioral deviations are natural and expressive.
- EFR generalizes CFR and ICFR to all behavioral deviations.
- There is an inherent tradeoff within EFR: strategic power increases with larger, more inclusive deviation classes, but so does computational complexity.
- We believe the partial sequence deviations manage this tradeoff well. They are both efficient and powerful.

Remaining Challenges

- **In practice:** Can EFR help us solve Stratego, Hanabi, Diplomacy, etc.? Perhaps, once we achieve performance at scale: enhance EFR with function approximation, Monte carlo sampling, variance reduction, etc.?
- **In theory:** What is the largest class of EFG deviations for which we can efficiently learn the corresponding correlated equilibrium concept? (Internal? Behavioral? A smaller subset?)

References

- Greenwald, Jafari, and Marks, “A general class of no-regret learning algorithms and game-theoretic equilibria”
- Gordon, Greenwald, and Marks, “No-regret learning in convex games”
- Morrill, D’Orazio, Sarfati, et al., “Hindsight and sequential rationality of correlated play”
- Morrill, D’Orazio, Lanctot, et al., “Efficient Deviation Types and Learning for Hindsight Rationality in Extensive-Form Games”
- Bowling et al., “Heads-Up Limit Hold’em Poker is Solved”
- Moravčík et al., “DeepStack: Expert-Level Artificial Intelligence in Heads-Up No-Limit Poker”
- Brown and Sandholm, “Superhuman AI for Heads-Up No-Limit Poker: Libratus Beats Top Professionals”
- Brown and Sandholm, “Superhuman AI for Multiplayer Poker”
- Zinkevich et al., “Regret Minimization in Games with Incomplete Information”
- Morrill, Greenwald, and Bowling, “The Partially Observable History Process”

References [cont'd]

- Aumann, "Subjectivity and correlation in randomized strategies"
- Moulin and Vial, "Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon"
- von Stengel and Forges, "Extensive-form correlated equilibrium: Definition and computational complexity"
- Dudík and Gordon, "A sampling-based approach to computing equilibria in succinct extensive-form games"
- Celli et al., "No-regret learning dynamics for extensive-form correlated equilibrium"
- Lanctot et al., "OpenSpiel: A Framework for Reinforcement Learning in Games"
- Blum and Mansour, "From external to internal regret"
- Freund et al., "Using and combining predictors that specialize"
- Hannan, "Approximation to Bayes risk in repeated play"
- Blackwell, "An analog of the minimax theorem for vector payoffs"

Image Source