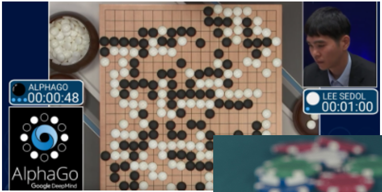
Learning in Zero-Sum Games

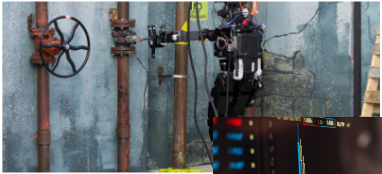
Alessandro LAZARIC (*Facebook AI Research* / on leave *Inria Lille*)

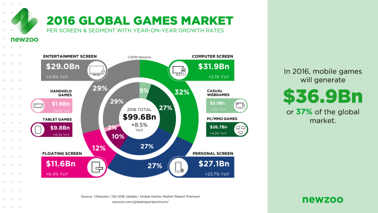
Institut d'Automne en Intelligence Artificielle, Lyon

Motivation: a Long-Standing Goal of AI...



...with Potential Applications in Real-World Environments





Outline

Learning in Two-Player Zero-Sum Games
- Regret Minimization and Nash Equilibria
- The Exp3 Algorithm

From Normal Form to Extensive Form Imperfect Information Games
- Regret Minimization and Nash Equilibria
- Counterfactual Regret Minimization

Outline

Learning in Two-Player Zero-Sum Games
Regret Minimization and Nash Equilibria
The Exp3 Algorithm

From Normal Form to Extensive Form Imperfect Information Games

Outline

Learning in Two-Player Zero-Sum Games

Regret Minimization and Nash Equilibria

The Exp3 Algorithm

From Normal Form to Extensive Form Imperfect Information Games

Normal Form Games

The game

- Set of players $N = \{1, \ldots, n\}$
- Action sets A_i, joint action set $A = A_1 \times \cdots \times A_n$
- Joint action $a \in A$, player i's action a_i, all other players a_{-i}
- Utility (payoff/reward) function $u : A \to \mathbb{R}^n$, player i's utility $u_i : A \to \mathbb{R}$

Mixed strategies

- Joint strategy $\sigma \in \mathcal{D}(A)$ such that $\sigma(a) = \prod_{i=1}^n \sigma_i(a_i)$
- Utility of a strategy $u_i(\sigma) = \sum_{a_i} \sum_{a_{-i}} \sigma_i(a_i)\sigma_{-i}(a_{-i})u_i(a_i, a_{-i})$

Two-Player Zero-Sum Games

The game

- Set of players $N = \{1, 2\} = \{i, j\}$
- Action sets A_i, joint action set $A = A_1 \times A_2$
- Joint action $a \in A$, player i's action a_i, other player's a_j
- Utility (payoff/reward) function $u : A \to \mathbb{R}^n$, player i's utility $u_i : A \to \mathbb{R}$

$$\forall a \in A, \quad u_1(a) = -u_2(a)$$

Solution concept

- *Nash equilibrium* $(\sigma_1^*, \sigma_2^*) = \arg \max_{\sigma_1} \min_{\sigma_2} u_1(\sigma_1, \sigma_2)$
- *Value of the game* $V = \max_{\sigma_1} \min_{\sigma_2} u_1(\sigma_1, \sigma_2)$

Rock-Paper-Scissors – The Game

Action set $A_1 = A_2 = \{(\text{R})\text{ock}, (\text{P})\text{aper}, (\text{S})\text{cissor}\}$

	R	*P*	*S*
R	*0*, *0*	*-1*, *1*	*1*, *-1*
P	*1*, *-1*	*0*, *0*	*-1*, *1*
S	*-1*, *1*	*1*, *-1*	*0*, *0*

Rock-Paper-Scissors – The Solution *(sketch)*

	R	*P*	*S*
R	*0*, *0*	*-1*, *1*	*1*, *-1*
P	*1*, *-1*	*0*, *0*	*-1*, *1*
S	*-1*, *1*	*1*, *-1*	*0*, *0*

- If (σ_1^*, σ_2^*) is a Nash equilibrium, then

$$\sigma_1^* = \mathrm{BR}(\sigma_2^*) = \arg\max_{\sigma_1} u_1(\sigma_1, \sigma_2^*) = \arg\max_{\sigma_1} \sum_{a_1 \in A_1} \sigma_1(a_1) u_1(a_1, \sigma_2^*)$$

Rock-Paper-Scissors – The Solution *(sketch)*

	R	*P*	*S*
R	*0*, *0*	*-1*, *1*	*1*, *-1*
P	*1*, *-1*	*0*, *0*	*-1*, *1*
S	*-1*, *1*	*1*, *-1*	*0*, *0*

- If (σ_1^*, σ_2^*) is a Nash equilibrium, then

$$\sigma_1^* = \mathrm{BR}(\sigma_2^*) = \arg\max_{\sigma_1} u_1(\sigma_1, \sigma_2^*) = \arg\max_{\sigma_1} \sum_{a_1 \in A_1} \sigma_1(a_1) u_1(a_1, \sigma_2^*)$$

$$\Rightarrow \forall a_1 \in A, \quad u_1 = u_1(a_1, \sigma_2^*)$$

Rock-Paper-Scissors – The Solution *(sketch)*

	R	P	S
R	0, 0	-1, 1	1, -1
P	1, -1	0, 0	-1, 1
S	-1, 1	1, -1	0, 0

- Let $\sigma_2 = \big(\sigma_2(R), \sigma_2(P), \sigma_2(S)\big)$ the strategy of player *column* then

$$
\begin{aligned}
u_1 &= u_1(R, \sigma_2) = 0\sigma_2(R) - 1\sigma_2(P) + 1\sigma_2(S) \\
u_1 &= u_1(P, \sigma_2) = 1\sigma_2(R) + 0\sigma_2(P) - 1\sigma_2(S) \\
u_1 &= u_1(S, \sigma_2) = -1\sigma_2(R) + 1\sigma_2(P) + 0\sigma_2(S) \\
1 &= \sigma_2(R) + \sigma_2(P) + \sigma_2(S)
\end{aligned}
$$

Rock-Paper-Scissors – The Solution *(sketch)*

	R	P	S
R	0, 0	-1, 1	1, -1
P	1, -1	0, 0	-1, 1
S	-1, 1	1, -1	0, 0

- Let $\sigma_2 = \big(\sigma_2(R), \sigma_2(P), \sigma_2(S)\big)$ the strategy of player *column* then

$$\begin{aligned} u_1 = u_1(R, \sigma_2) &= 0\sigma_2(R) - 1\sigma_2(P) + 1\sigma_2(S) \\ u_1 = u_1(P, \sigma_2) &= 1\sigma_2(R) + 0\sigma_2(P) - 1\sigma_2(S) \\ u_1 = u_1(S, \sigma_2) &= -1\sigma_2(R) + 1\sigma_2(P) + 0\sigma_2(S) \\ 1 &= \sigma_2(R) + \sigma_2(P) + \sigma_2(S) \end{aligned}$$

- Solving for all variables gives $\sigma_2^* = (1/3, 1/3, 1/3)$ and $u_1 = 0$

Rock-Paper-Scissors – The Solution *(sketch)*

	R	P	S
R	0, 0	-1, 1	1, -1
P	1, -1	0, 0	-1, 1
S	-1, 1	1, -1	0, 0

- Let $\sigma_2 = \big(\sigma_2(R), \sigma_2(P), \sigma_2(S)\big)$ the strategy of player *column* then

$$\begin{aligned} u_1 = u_1(R, \sigma_2) &= 0\sigma_2(R) - 1\sigma_2(P) + 1\sigma_2(S) \\ u_1 = u_1(P, \sigma_2) &= 1\sigma_2(R) + 0\sigma_2(P) - 1\sigma_2(S) \\ u_1 = u_1(S, \sigma_2) &= -1\sigma_2(R) + 1\sigma_2(P) + 0\sigma_2(S) \\ 1 &= \sigma_2(R) + \sigma_2(P) + \sigma_2(S) \end{aligned}$$

- Solving for all variables gives $\sigma_2^* = (1/3, 1/3, 1/3)$ and $u_1 = 0$
- Repeating for player *row* gives $\sigma_1^* = (1/3, 1/3, 1/3)$ and $u_2 = 0$

Rock-Paper-Scissors – The Solution *(sketch)*

	R	P	S
R	0, 0	-1, 1	1, -1
P	1, -1	0, 0	-1, 1
S	-1, 1	1, -1	0, 0

- Let $\sigma_2 = \big(\sigma_2(R), \sigma_2(P), \sigma_2(S)\big)$ the strategy of player *column* then

$$u_1 = u_1(R, \sigma_2) = 0\sigma_2(R) - 1\sigma_2(P) + 1\sigma_2(S)$$
$$u_1 = u_1(P, \sigma_2) = 1\sigma_2(R) + 0\sigma_2(P) - 1\sigma_2(S)$$
$$u_1 = u_1(S, \sigma_2) = -1\sigma_2(R) + 1\sigma_2(P) + 0\sigma_2(S)$$
$$1 = \sigma_2(R) + \sigma_2(P) + \sigma_2(S)$$

- Solving for all variables gives $\sigma_2^* = (1/3, 1/3, 1/3)$ and $u_1 = 0$
- Repeating for player *row* gives $\sigma_1^* = (1/3, 1/3, 1/3)$ and $u_2 = 0$
- (σ_1^*, σ_2^*) is a Nash equilibrium and the value of the game is $V = 0$

A Single-Player Perspective

Sequential game

- For $t = 1, \ldots, n$
  - Player 1 chooses $\sigma_{1,t}$
  - Player 2 chooses $\sigma_{2,t}$
  - Players play actions $a_{1,t} \sim \sigma_{1,t}$ and $a_{2,t} \sim \sigma_{2,t}$
  - Players receive payoffs $u_1(a_{1,t}, a_{2,t})$ and $u_2(a_{1,t}, a_{2,t})$

Solution: Nash equilibrium

$$(\sigma_1^*, \sigma_2^*) = \arg\max_{\sigma_1} \min_{\sigma_2} u_1(\sigma_1, \sigma_2)$$

A Single-Player Perspective

Sequential game $\Rightarrow$ *Single-player game*

- For $t = 1, \ldots, n$
  - Player 1 chooses $\sigma_{1,t}$
  - ~~*Player 2 chooses* $\sigma_{2,t}$~~
  - Players play actions $a_{1,t} \sim \sigma_{1,t}$ and $\cancel{a_{2,t} \sim \sigma_{2,t}}$
  - Players receive payoffs $u_1(a_{1,t}, a_{2,t})$ and $\cancel{u_2(a_{1,t}, a_{2,t})}$

Solution: Nash equilibrium $\Rightarrow$ *Maximize the (average) utility*

$$\cancel{(\sigma_1^*, \sigma_2^*) = \arg\max_{\sigma_1} \min_{\sigma_2} u_1(\sigma_1, \sigma_2)}$$

$$\begin{aligned}(a_{1,1}^*, \ldots, a_{1,n}^*) &= \arg \max_{(a_{1,1},\ldots,a_{1,n})} \frac{1}{n} \sum_{t=1}^{n} u_1(a_{1,t}, a_{2,t}) \\ &= \arg \max_{(a_{1,1},\ldots,a_{1,n})} \frac{1}{n} \sum_{t=1}^{n} u_{1,t}(a_{1,t})\end{aligned}$$

The (Multi-Armed Bandit) Problem

A learning problem

- For $t = 1, \dots, n$
  - Player 1 chooses $\sigma_{1,t}$
  - Player 1 plays action $a_{1,t} \sim \sigma_{1,t}$
  - Player 1 receives payoff $u_{1,t}(a_{1,t})$

Remarks

- No information about $a_{2,t}$ and utility u_2
- Utility function $u_{1,t}$ is only observed for $a_{1,t}$ (i.e., $u_{1,t}(a_{1,t})$)

The (Multi-Armed Bandit) Problem

- *Regret in hindisight* w.r.t. any fixed action a_1

$$R_n(a_1) = \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_1) - \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_{1,t})$$

The (Multi-Armed Bandit) Problem

- *Regret in hindisight* w.r.t. any fixed action a_1

$$R_n(a_1) = \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_1) - \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_{1,t})$$

- *Objective:* find actions $(a_{1,1}, \ldots, a_{1,n})$ that maximize average utility $\approx$ *minimize the regret* w.r.t. the best action a_1

Utility: $$\frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_{1,t})$$

Regret: $$R_n = \max_{a_1} \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_1) - \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_{1,t})$$

Regret Minimization and Nash Equilibria

Theorem

A learning algorithm is Hannan's consistent if

$$\lim_{n\to\infty} R_n = 0 \quad a.s.$$

Given a two-player zero-sum game with value V, if players choose strategies $\sigma_{1,t}$ and $\sigma_{2,t}$ using a Hannan's consistent algorithm, then

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} u_1(a_{1,t}, a_{2,t}) = V$$

Furthermore, let empirical frequency strategies be

$$\widehat{\sigma}_{1,n}(a_1) = \frac{1}{n}\sum_{t=1} \mathbb{I}\{a_{1,t} = a_1\} \quad \text{and} \quad \widehat{\sigma}_{2,n}(a_2) = \frac{1}{n}\sum_{t=1} \mathbb{I}\{a_{2,t} = a_2\}$$

then the joint empirical strategy

$$\widehat{\sigma}_{1,n} \times \widehat{\sigma}_{2,n} \stackrel{n\to\infty}{\longrightarrow} \big\{(\sigma_1^*, \sigma_2^*)\big\}_{Nash}$$

Regret Minimization and Nash Equilibria *[proof]*

- *[Hannan's consistency]*

$$\lim_{n\to\infty} R_n = 0 \quad \Longleftrightarrow \quad \lim_{n\to\infty}\Big(\max_{a_1} \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_1) - \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_{1,t})\Big) = 0$$

Regret Minimization and Nash Equilibria *[proof]*

- *[Hannan's consistency]*

$$\lim_{n\to\infty} R_n = 0 \quad \Longleftrightarrow \quad \lim_{n\to\infty} \Big(\max_{a_1} \frac{1}{n} \sum_{t=1}^{n} u_{1,t}(a_1) - \frac{1}{n} \sum_{t=1}^{n} u_{1,t}(a_{1,t}) \Big) = 0$$

- *[linearity of utility function]*

$$\max_{\sigma_1} \frac{1}{n} \sum_{t=1}^{n} u_{1,t}(\sigma_1) = \max_{\sigma_1} \frac{1}{n} \sum_{t=1}^{n} \sum_{a_1\in A_1} \sigma_1(a_1) u_{1,t}(a_1) = \max_{a_1} \frac{1}{n} \sum_{t=1}^{n} u_{1,t}(a_1)$$

Regret Minimization and Nash Equilibria *[proof]*

- *[Hannan's consistency]*

$$\lim_{n\to\infty} R_n = 0 \iff \lim_{n\to\infty}\Big(\max_{a_1}\frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_1) - \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_{1,t})\Big) = 0$$

- *[linearity of utility function]*

$$\max_{\sigma_1}\frac{1}{n}\sum_{t=1}^{n} u_{1,t}(\sigma_1) = \max_{\sigma_1}\frac{1}{n}\sum_{t=1}^{n}\sum_{a_1\in A_1}\sigma_1(a_1)u_{1,t}(a_1) = \max_{a_1}\frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_1)$$

- *[definition]* $u_{1,t}(\sigma_1) = u_1(\sigma_1, a_{2,t})$

$$\Rightarrow \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(\sigma_1) = \frac{1}{n}\sum_{t=1}^{n}\sum_{a_2\in A_2}\mathbb{I}\{a_{2,t}=a_2\}u_1(\sigma_1,a_2) = \sum_{a_2\in A_2} u_1(\sigma_1,a_2)\underbrace{\frac{1}{n}\sum_{t=1}^{n}\mathbb{I}\{a_{2,t}=a_2\}}_{\widehat{\sigma}_{2,n}(a_2)}$$

Regret Minimization and Nash Equilibria *[proof]*

- *[Hannan's consistency]*

$$\lim_{n\to\infty} R_n = 0 \iff \lim_{n\to\infty}\Big(\max_{a_1}\frac{1}{n}\sum_{t=1}^n u_{1,t}(a_1) - \frac{1}{n}\sum_{t=1}^n u_{1,t}(a_{1,t})\Big) = 0$$

- *[linearity of utility function]*

$$\max_{\sigma_1}\frac{1}{n}\sum_{t=1}^n u_{1,t}(\sigma_1) = \max_{\sigma_1}\frac{1}{n}\sum_{t=1}^n\sum_{a_1\in A_1}\sigma_1(a_1)u_{1,t}(a_1) = \max_{a_1}\frac{1}{n}\sum_{t=1}^n u_{1,t}(a_1)$$

- *[definition]* $u_{1,t}(\sigma_1) = u_1(\sigma_1, a_{2,t})$

$$\Rightarrow \frac{1}{n}\sum_{t=1}^n u_{1,t}(\sigma_1) = \frac{1}{n}\sum_{t=1}^n\sum_{a_2\in A_2}\mathbb{I}\{a_{2,t}=a_2\}u_1(\sigma_1,a_2) = \sum_{a_2\in A_2} u_1(\sigma_1,a_2)\underbrace{\frac{1}{n}\sum_{t=1}^n\mathbb{I}\{a_{2,t}=a_2\}}_{\widehat{\sigma}_{2,n}(a_2)}$$

- *[one-side of the result]*

$$\max_{\sigma_1}\frac{1}{n}\sum_{t=1}^n u_{1,t}(\sigma_1) = \max_{\sigma_1}\frac{1}{n}\sum_{t=1}^n u_1(\sigma_1,\widehat{\sigma}_{2,n}) \geq \max_{\sigma_1}\min_{\sigma_2}\frac{1}{n}\sum_{t=1}^n u_1(\sigma_1,\sigma_2) = V$$

- *[for player 2]* $\Rightarrow$ desired result.

Regret Minimization and Nash Equilibria

Corollary

If

$$R_n \leq \epsilon$$

then the joint empirical strategy is ϵ-Nash, i.e.,

$$u_1(\widehat{\sigma}_{1,n} \times \widehat{\sigma}_{2,n}) \geq V - \epsilon$$

Outline

Learning in Two-Player Zero-Sum Games

Regret Minimization and Nash Equilibria

The Exp3 Algorithm

From Normal Form to Extensive Form Imperfect Information Games

Hannan's Consistent Algorithms

A learning problem

- For $t = 1, \dots, n$
  - Player 1 chooses $\sigma_{1,t}$
  - Player 1 plays action $a_{1,t} \sim \sigma_{1,t}$
  - Player 1 receives payoff $u_{1,t}(a_{1,t})$

Objective

- Regret

$$R_n = \max_{a_1} \frac{1}{n} \sum_{t=1}^{n} u_{1,t}(a_1) - \frac{1}{n} \sum_{t=1}^{n} u_{1,t}(a_{1,t})$$

- Hannan's consistent algorithm

$$\lim_{n \to \infty} R_n = 0 \quad \text{a.s.}$$

Learning the Nash Equilibrium

Version 1: fictitious play full information (aka follow-the-leader)

- For $t = 1, \ldots, n$
  - Compute greedy action

$$a_t^* = \arg\max_{a \in A_1} \sum_{s=1}^{t-1} u_{1,t}(a)$$

  - Player chooses $\sigma_{1,t} = \delta(a_t^*)$
  - Player plays action $a_{1,t} \sim \sigma_{1,t}$
  - Player receives payoff $u_{1,t}(a_{1,t})$

Remarks

- This strategy is easily exploitable $R_n = O(1)$
- Self play *does not converge* in general

Learning the Nash Equilibrium

Version 2: exponentially weighted forcaster (EWF)

- *Initialize weights* $w_0(a) = 0$ for all $a \in A_1$
- For $t = 1, \ldots, n$
  - Player chooses

$$\sigma_{1,t}(a) = \frac{w_{t-1}(a)}{\sum_{b \in A_1} w_{t-1}(b)}$$

  - Player plays action $a_{1,t} \sim \sigma_{1,t}$
  - Player receives payoff $u_{1,t}(a_{1,t})$ and $u_{1,t}(a)$ for all a
  - Update weights

$$w_t(a) = w_{t-1}(a) \exp\big(\eta_t u_{1,t}(a)\big)$$

Learning the Nash Equilibrium

Version 2: exponentially weighted forcaster (EWF)

- *Initialize weights* $w_0(a) = 0$ for all $a \in A_1$
- For $t = 1, \ldots, n$
  - Player chooses

$$\sigma_{1,t}(a) = \frac{w_{t-1}(a)}{\sum_{b \in A_1} w_{t-1}(b)} \quad \textit{[prop. to weights]}$$

  - Player plays action $a_{1,t} \sim \sigma_{1,t}$
  - Player receives payoff $u_{1,t}(a_{1,t})$ and $u_{1,t}(a)$ for all a
  - Update weights

$$w_t(a) = w_{t-1}(a) \exp\big(\eta_t u_{1,t}(a)\big)$$

Learning the Nash Equilibrium

Version 2: exponentially weighted forcaster (EWF)

- *Initialize weights* $w_0(a) = 0$ for all $a \in A_1$
- For $t = 1, \ldots, n$
  - Player chooses

$$\sigma_{1,t}(a) = \frac{w_{t-1}(a)}{\sum_{b \in A_1} w_{t-1}(b)} \quad \textit{[prop. to weights]}$$

  - Player plays action $a_{1,t} \sim \sigma_{1,t}$
  - Player receives payoff $u_{1,t}(a_{1,t})$ and $u_{1,t}(a)$ for all a *[full info]*
  - Update weights

$$w_t(a) = w_{t-1}(a) \exp\big(\eta_t u_{1,t}(a)\big)$$

Learning the Nash Equilibrium

Version 2: exponentially weighted forcaster (EWF)

- *Initialize weights* $w_0(a) = 0$ for all $a \in A_1$
- For $t = 1, \ldots, n$
  - Player chooses

$$\sigma_{1,t}(a) = \frac{w_{t-1}(a)}{\sum_{b \in A_1} w_{t-1}(b)} \quad \textit{[prop. to weights]}$$

  - Player plays action $a_{1,t} \sim \sigma_{1,t}$
  - Player receives payoff $u_{1,t}(a_{1,t})$ and $u_{1,t}(a)$ for all a *[full info]*
  - Update weights

$$w_t(a) = w_{t-1}(a) \exp\big(\eta_t u_{1,t}(a)\big) \quad \textit{[exponentiated utility]}$$

Learning the Nash Equilibrium

Theorem

If EWF *is run over n steps with* $\eta_t = \eta$, *then with probability* $1 - \delta$

$$R_n = \max_{a_1} \frac{1}{n} \sum_{t=1}^{n} u_{1,t}(a_1) - \frac{1}{n} \sum_{t=1}^{n} u_{1,t}(a_{1,t}) \leq \frac{\log(A_1)}{n\eta} + \frac{\eta}{8} + \sqrt{\frac{1}{2n} \log(1/\delta)}$$

Setting $\eta = \sqrt{8 \log(A_1)/n}$ *we obtain*

$$R_n \leq \sqrt{\frac{\log(A_1)}{2n}} + \sqrt{\frac{1}{2n} \log(1/\delta)}$$

Learning the Nash Equilibrium

Theorem

If EWF *is run over n steps with* $\eta_t = \eta$, *then with probability* $1 - \delta$

$$R_n = \max_{a_1} \frac{1}{n} \sum_{t=1}^{n} u_{1,t}(a_1) - \frac{1}{n} \sum_{t=1}^{n} u_{1,t}(a_{1,t}) \leq \frac{\log(A_1)}{n\eta} + \frac{\eta}{8} + \sqrt{\frac{1}{2n} \log(1/\delta)}$$

Setting $\eta = \sqrt{8 \log(A_1)/n}$ *we obtain*

$$R_n \leq \sqrt{\frac{\log(A_1)}{2n}} + \sqrt{\frac{1}{2n} \log(1/\delta)}$$

Remarks

- $\lim_{n\to\infty} R_n \leq 0 \quad \Rightarrow$ *Hannan's consistency*
- Rate of convergence $O(1/\sqrt{n})$
- In self-play EWF "converges" to the Nash equilibrium

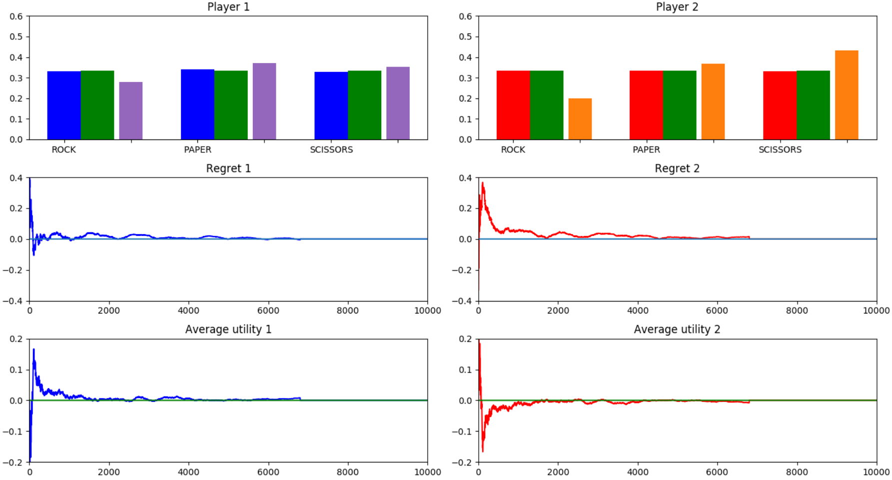
Rock-Paper-Scissors – The Simulation

Action set $A_1 = A_2 = \{(\text{R})\text{ock}, (\text{P})\text{aper}, (\text{S})\text{cissor}\}$

	R	*P*	*S*
R	*0*, *0*	*-1*, *1*	*1*, *-1*
P	*1*, *-1*	*0*, *0*	*-1*, *1*
S	*-1*, *1*	*1*, *-1*	*0*, *0*

- Equilibrium $\sigma_1^* = \sigma_2^* = (1/3,\ 1/3,\ 1/3)$
- Value of the game $V = 0.0$

Rock-Paper-Scissors – The Simulation



Rock-Paper-Scissors – The Simulation *Mod*

Action set $A_1 = A_2 = \{\text{(R)ock}, \text{(P)aper}, \text{(S)cissor}\}$

	R	*P*	*S*
R	*0*, *0*	*-1*, *1*	**2**, **-2**
P	*1*, *-1*	*0*, *0*	*-1*, *1*
S	*-1*, *1*	*1*, *-1*	*0*, *0*

- Equilibrium $\sigma_1^* = (1/4,\ 5/12,\ 1/3)$
- Value of the game $V = 1/12 (\approx 0.833)$

Learning the Nash Equilibrium

Version 2: exponentially weighted forcaster (EWF)

- *Initialize weights* $w_0(a) = 0$ for all $a \in A_1$
- For $t = 1, \ldots, n$
  - Player chooses

$$\sigma_{1,t}(a) = \frac{w_{t-1}(a)}{\sum_{b \in A_1} w_{t-1}(b)} \quad \textit{[prop. to weights]}$$

  - Player plays action $a_{1,t} \sim \sigma_{1,t}$
  - Player receives payoff $u_{1,t}(a_{1,t})$ and $u_{1,t}(a)$ for all a *[full info]*
  - Update weights

$$w_t(a) = w_{t-1}(a) \exp\big(\eta_t u_{1,t}(a)\big) \quad \textit{[exponentiated utility]}$$

Learning the Nash Equilibrium

Version 2: exponentially weighted forcaster (EWF)

- *Initialize weights* $w_0(a) = 0$ for all $a \in A_1$
- For $t = 1, \dots, n$
  - Player chooses

$$\sigma_{1,t}(a) = \frac{w_{t-1}(a)}{\sum_{b \in A_1} w_{t-1}(b)} \quad \textit{[prop. to weights]}$$

  - Player plays action $a_{1,t} \sim \sigma_{1,t}$
  - Player receives payoff $u_{1,t}(a_{1,t})$ ~~and $u_{1,t}(a)$ for all a *[full info]*~~
  - Update weights

$$w_t(a) = w_{t-1}(a) \exp\big(\eta_t u_{1,t}(a)\big) \quad \textit{[exponentiated utility]}$$

Learning the Nash Equilibrium

Problem:

- Player plays action $a_{1,t} \sim \sigma_{1,t}$
- Player receives payoff $u_{1,t}(a_{1,t})$
- Update weights

$$w_t(a) = w_{t-1}(a)\exp\big(\eta_t u_{1,t}(a)\big) \quad \textit{[exponentiated utility]}$$

Learning the Nash Equilibrium

Problem:

- Player plays action $a_{1,t} \sim \sigma_{1,t}$
- Player receives payoff $u_{1,t}(a_{1,t})$
- Update weights

$$w_t(a) = w_{t-1}(a)\exp\big(\eta_t u_{1,t}(a)\big) \quad \textit{[exponentiated utility]}$$

Solution:

- Importance sampling

$$\widetilde{u}_{1,t}(a) = \begin{cases} \dfrac{u_{1,t}(a_{1,t})}{\sigma_{1,t}(a_{1,t})} & \text{if } a = a_{1,t} \\ 0 & \text{otherwise} \end{cases}$$

- Unbiased estimator

$$\forall a \in A_1 \quad \mathbb{E}_{a\sim\sigma_{1,t}}\big[\widetilde{u}_{1,t}(a)\big] = \sigma_{1,t}(a)\frac{u_{1,t}(a)}{\sigma_{1,t}} + (1-\sigma_{1,t}(a))\times 0 = u_{1,t}(a)$$

Learning the Nash Equilibrium

Version 3: EWF for Exploration-Exploitation (EXP3)

- *Initialize weights* $w_0(a) = 0$ for all $a \in A_1$
- For $t = 1, \ldots, n$
  - Player chooses
$$\sigma_{1,t}(a) = \frac{w_{t-1}(a)}{\sum_{b \in A_1} w_{t-1}(b)} \quad \textit{[prop. to weights]}$$
  - Player plays action $a_{1,t} \sim \sigma_{1,t}$
  - Player receives payoff $u_{1,t}(a_{1,t})$
  - Compute *pseudo-payoffs*
$$\widetilde{u}_{1,t}(a) = \begin{cases} \frac{u_{1,t}(a_{1,t})}{\sigma_{1,t}(a_{1,t})} & \text{if } a = a_{1,t} \\ 0 & \text{otherwise} \end{cases}$$
  - Update weights
$$w_t(a) = w_{t-1}(a) \exp\big(\eta_t \widetilde{u}_{1,t}(a)\big)$$

Learning the Nash Equilibrium

Theorem

If $\mathrm{EXP3}$ *is run over* n *steps with* $\eta_t = \sqrt{2\log(A_1)/(nA_1)}$, *then its psuedo-regret is bounded as*

$$\overline{R}_n = \max_{a_1} \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\big[u_{1,t}(a_1)\big] - \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\big[u_{1,t}(a_{1,t})\big] \leq \sqrt{\frac{2A_1 \log(A_1)}{n}}$$

Learning the Nash Equilibrium

Theorem

If $\mathrm{EXP3}$ *is run over n steps with* $\eta_t = \sqrt{2\log(A_1)/(nA_1)}$, *then its psuedo-regret is bounded as*

$$\overline{R}_n = \max_{a_1} \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\big[u_{1,t}(a_1)\big] - \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\big[u_{1,t}(a_{1,t})\big] \leq \sqrt{\frac{2A_1\log(A_1)}{n}}$$

Remarks

- $\lim_{n\to\infty} \overline{R}_n \leq 0 \quad \Rightarrow$ *Hannan's consistency?*
- Rate of convergence $O(1/\sqrt{n})$
- Regret larger by a factor $\sqrt{A_1}$ (observing 1 *vs* A_1 payoffs)

Rock-Paper-Scissors – The Simulation *Mod2*

Action set $A_1 = A_2 = \{(\text{R})\text{ock}, (\text{P})\text{aper}, (\text{S})\text{cissor}\}$

	R	*P*	*S*
R	*0*, *0*	*-1*, *1*	**5**, **-5**
P	*1*, *-1*	*0*, *0*	*-1*, *1*
S	*-1*, *1*	*1*, *-1*	*0*, *0*

- Equilibrium $\sigma_1^* = (1/7,\ 11/21,\ 1/3)$
- Value of the game $V = 4/21 (\approx 0.1904)$

Learning the Nash Equilibrium

Problem:

- Importance sampling is unbiased

$$\widetilde{u}_{1,t}(a) = \begin{cases} \frac{u_{1,t}(a_{1,t})}{\sigma_{1,t}(a_{1,t})} & \text{if } a = a_{1,t} \\ 0 & \text{otherwise} \end{cases} \quad ; \quad \mathbb{E}_{a\sim\sigma_{1,t}}\left[\widetilde{u}_{1,t}(a)\right] = u_{1,t}(a)$$

- Variance

$$\mathbb{V}_{a\sim\sigma_{1,t}}\left[\widetilde{u}_{1,t}(a)\right] \xrightarrow{\sigma_{1,t}(a)\to 0} \infty$$

Learning the Nash Equilibrium

Problem:

- Importance sampling is unbiased

$$\widetilde{u}_{1,t}(a) = \begin{cases} \frac{u_{1,t}(a_{1,t})}{\sigma_{1,t}(a_{1,t})} & \text{if } a = a_{1,t} \\ 0 & \text{otherwise} \end{cases} \quad ; \quad \mathbb{E}_{a\sim\sigma_{1,t}}\big[\widetilde{u}_{1,t}(a)\big] = u_{1,t}(a)$$

- Variance

$$\mathbb{V}_{a\sim\sigma_{1,t}}\big[\widetilde{u}_{1,t}(a)\big] \xrightarrow{\sigma_{1,t}(a)\to 0} \infty$$

Solution:

- *Bias* both pseudo-payoff

$$\widetilde{u}_{1,t}(a) = \frac{u_{1,t}(a_{1,t})\mathbb{I}\{a = a_{1,t}\} + \beta_t}{\sigma_{1,t}(a_{1,t})}$$

- Mix strategy with *uniform* exploration

$$\sigma_{1,t}(a) = (1-\gamma_t)\frac{w_{1,t}(a)}{\sum b \in A_1 w_{1,t}(b)} + \frac{\gamma_t}{A_1}$$

Learning the Nash Equilibrium

Version 3: EWF for Exploration-Exploitation w.h.p. (EXP3.P)

- *Initialize weights* $w_0(a) = 0$ for all $a \in A_1$
- For $t = 1, \ldots, n$
  - Player chooses

$$\sigma_{1,t}(a) = (1 - \gamma_t) \frac{w_{1,t}(a)}{\sum b \in A_1 w_{1,t}(b)} + \frac{\gamma_t}{A_1}$$

  - Player plays action $a_{1,t} \sim \sigma_{1,t}$
  - Player receives payoff $u_{1,t}(a_{1,t})$
  - Compute *pseudo-payoffs*

$$\widetilde{u}_{1,t}(a) = \frac{u_{1,t}(a_{1,t})\mathbb{I}\{a = a_{1,t}\} + \beta_t}{\sigma_{1,t}(a_{1,t})}$$

  - Update weights

$$w_t(a) = w_{t-1}(a)\exp\big(\eta_t \widetilde{u}_{1,t}(a)\big)$$

Learning the Nash Equilibrium

Theorem

If EXP3.P *is run over n steps with* $\beta_t \approx \eta_t = \sqrt{2\log(A_1)/(nA_1)}$, $\gamma_t = \sqrt{A_1 \log(A_1)/n}$, *then with probability* $1-\delta$ *its regret is bounded as*

$$R_n = \max_{a_1} \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_1) - \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_{1,t}) \leq 6\sqrt{\frac{A_1\log(A_1/\delta)}{n}}$$

Learning the Nash Equilibrium

Theorem

If $\mathrm{EXP3.P}$ *is run over* n *steps with* $\beta_t \approx \eta_t = \sqrt{2\log(A_1)/(nA_1)}$, $\gamma_t = \sqrt{A_1 \log(A_1)/n}$, *then with probability* $1-\delta$ *its regret is bounded as*

$$R_n = \max_{a_1} \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_1) - \frac{1}{n}\sum_{t=1}^{n} u_{1,t}(a_{1,t}) \leq 6\sqrt{\frac{A_1 \log(A_1/\delta)}{n}}$$

Remarks

- $\lim_{n\to\infty} R_n \leq 0 \quad \Rightarrow$ *Hannan's consistency!*
- $\mathrm{EXP3.P}$ in self-play converges to Nash equilibrium

Rock-Paper-Scissors – The Simulation *Mod2*

Action set $A_1 = A_2 = \{(\mathrm{R})\mathrm{ock}, (\mathrm{P})\mathrm{aper}, (\mathrm{S})\mathrm{cissor}\}$

	R	*P*	*S*
R	*0*, *0*	*-1*, *1*	***5***, ***-5***
P	*1*, *-1*	*0*, *0*	*-1*, *1*
S	*-1*, *1*	*1*, *-1*	*0*, *0*

- Equilibrium $\sigma_1^* = (1/7,\ 11/21,\ 1/3)$
- Value of the game $V = 4/21 (\approx 0.1904)$

Summary

+ $\mathrm{EXP3.P}$ minimizes regret in adversarial environments
+ $\mathrm{EXP3.P}$ converges to Nash equilibria in self-play
+ No need to know
  - Utility function (i.e., the rules of the game)
  - Actions performed by the adversary

Summary

+ $\mathrm{EXP3.P}$ minimizes regret in adversarial environments
+ $\mathrm{EXP3.P}$ converges to Nash equilibria in self-play
+ No need to know
  - Utility function (i.e., the rules of the game)
  - Actions performed by the adversary

≈ Some of this can be extended to learn correlated equilibria

Summary

+ EXP3.P minimizes regret in adversarial environments
+ EXP3.P converges to Nash equilibria in self-play
+ No need to know
    - Utility function (i.e., the rules of the game)
    - Actions performed by the adversary

≈ Some of this can be extended to learn correlated equilibria

– Exponential may be tricky to manage
– Convergence is only in the empirical frequency
– Convergence is relatively slow

Outline

Learning in Two-Player Zero-Sum Games

From Normal Form to Extensive Form Imperfect Information Games

- Regret Minimization and Nash Equilibria
- Counterfactual Regret Minimization

Outline

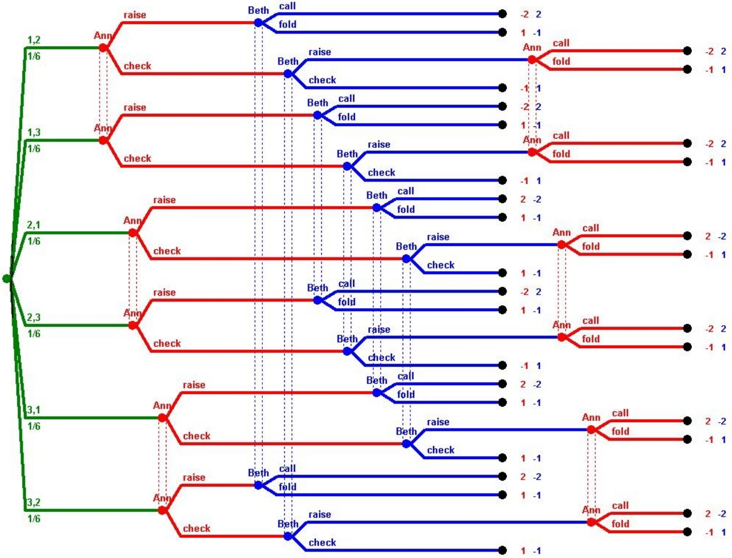
Learning in Two-Player Zero-Sum Games

From Normal Form to Extensive Form Imperfect Information Games

Regret Minimization and Nash Equilibria

Counterfactual Regret Minimization

Kuhn Poker – The Game



Imperfect Information Extensive Form Games

The game

- Set of players $N = \{1, \ldots, n\}$ and c chance player (e.g., deck)
- Set of possible sequences of actions H, $Z \subseteq H$ set of terminal histories
- Player function $P : H \to N \cup \{c\}$
- Set of information sets $\mathcal{I} = \{I\}$ (i.e., I is a subset of histories that are not "distinguishable")
- Utility of a terminal history $u_i : Z \to \mathbb{R}$
- Strategy $\sigma_i : \mathcal{I} \to \mathcal{D}(A)$ (in all $h \in I$ such that $P(h) = i$)

Imperfect Information Extensive Form Games

The game

- Set of players $N = \{1, \ldots, n\}$ and c chance player (e.g., deck)
- Set of possible sequences of actions H, $Z \subseteq H$ set of terminal histories
- Player function $P : H \to N \cup \{c\}$
- Set of information sets $\mathcal{I} = \{I\}$ (i.e., I is a subset of histories that are not "distinguishable")
- Utility of a terminal history $u_i : Z \to \mathbb{R}$
- Strategy $\sigma_i : \mathcal{I} \to \mathcal{D}(A)$ (in all $h \in I$ such that $P(h) = i$)

Two-Player Zero-Sum Extensive Form Game

- $N = \{1, 2\}$
- $u_1 = -u_2$

Extensive Form Games

Histories

- Prob. of reaching history $h \in H$ following joint strategy σ, $\pi^\sigma(h)$
- Prob. of reaching information set $I \in \mathcal{I}$ following joint strategy σ, $\pi^\sigma(I) = \sum_{h \in I} \pi^\sigma(h)$
- Prob. of reaching history $h \in H$ following joint strategy σ_{-i}, except player i following actions in h w.p. 1, $\pi^\sigma_{-i}(h)$
- Prob. of reaching history $h \in H$ following player i's actions, except others, $\pi^\sigma_i(h)$
- Replacement of $\sigma(I)$ to $\delta(a)$, $\sigma_{I \rightarrow a}$

Solution concept

- *Nash equilibrium* $(\sigma_1^*, \sigma_2^*) = \arg\max_{\sigma_1} \min_{\sigma_2} u_1(\sigma_1, \sigma_2)$
- *Value of the game* $V = \max_{\sigma_1} \min_{\sigma_2} u_1(\sigma_1, \sigma_2)$
- *Remark:* other concepts exist in this case, NE

The Regret View

- *Regret in hindsight* w.r.t. any fixed strategy σ_1

$$R_n(\sigma_1) = \frac{1}{n}\sum_{t=1}^{n} u_1(\sigma_1, \sigma_{2,t}) - \frac{1}{n}\sum_{t=1}^{n} u_1(\sigma_{1,t}, \sigma_{2,t})$$

- Regret against the best strategy in hindsight

$$R_n = \max_{\sigma_1} R_n(\sigma_1)$$

The Regret View

- *Regret in hindsight* w.r.t. any fixed strategy σ_1

$$R_n(\sigma_1) = \frac{1}{n}\sum_{t=1}^{n} u_1(\sigma_1, \sigma_{2,t}) - \frac{1}{n}\sum_{t=1}^{n} u_1(\sigma_{1,t}, \sigma_{2,t})$$

- Regret against the best strategy in hindsight

$$R_n = \max_{\sigma_1} R_n(\sigma_1)$$

- *Empirical strategy:*

$$\widehat{\sigma}_{1,n}(I,a) = \frac{\sum_{t=1}^{n} \pi_i^{\sigma_t}(I)\sigma_t(I,a)}{\sum_{t=1}^{n} \pi_i^{\sigma_t}(I)}$$

Regret Minimization and Nash Equilibria

Theorem

A learning algorithm is Hannan's consistent if

$$\lim_{n\to\infty} R_n = 0 \quad a.s.$$

Given a two-player zero-sum extensive-form game with value V, if players choose strategies $\sigma_{1,t}$ and $\sigma_{2,t}$ using a Hannan's consistent algorithm, then

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} u_1(\sigma_{1,t}, \sigma_{2,t}) = V$$

Furthermore, the joint empirical strategy

$$\widehat{\sigma}_{1,n} \times \widehat{\sigma}_{2,n} \stackrel{n\to\infty}{\longrightarrow} \big\{(\sigma_1^*, \sigma_2^*)\big\}_{Nash}$$

Outline

Learning in Two-Player Zero-Sum Games

From Normal Form to Extensive Form Imperfect Information Games

Regret Minimization and Nash Equilibria

Counterfactual Regret Minimization

Regret Matching Algorithm

- Back to Rock-Paper-Scissors
- Let $a_1 = rock$ and $a_2 = paper$
- Then the ***counterfactual*** regret

$$r(a_1 \rightarrow rock) = u_1(rock, a_{2,t}) - u_1(a_{1,t}, a_{2,t}) = -1 - (-1) = 0$$
$$r(a_1 \rightarrow paper) = u_1(paper, a_{2,t}) - u_1(a_{1,t}, a_{2,t}) = 0 - (-1) = 1$$
$$r(a_1 \rightarrow scissors) = u_1(scissors, a_{2,t}) - u_1(a_{1,t}, a_{2,t}) = 1 - (-1) = 2$$

- Regret matching idea

$$\sigma(a) = \frac{r(a_1 \rightarrow a)}{\sum_{b \in A_1} r(a_1 \rightarrow b)}$$

Sequential Problem

A learning problem

- For $t = 1, \ldots, n$
  - Player 1 chooses $\sigma_{1,t}$
  - Player 1 executes actions prescribed by $\sigma_{1,t}$ through a full game
  - Player 1 receives payoff $u_{1,t}$

Counterfactual Regret

- Counterfactual value of a history

$$v_i(\sigma, h) = \sum_{z \in Z, h \sqsubset z} \pi^{\sigma}_{-i}(h) \pi^{\sigma}(h, z) u_i(z)$$

- Counterfactual regret of not taking a in h

$$r_i^{\sigma}(h, a) = v_i(\sigma_{I \to a}, h) - v_i(\sigma, h), \quad I \ni h$$

- Counterfactual regret of not taking a in an information set I

$$r_i^{\sigma}(I, a) = \sum_{h \in I} r_i^{\sigma}(h, a)$$

- Cumulative counterfactual regret

$$R_{i,t}(I, a) = \sum_{s=1}^{t} r_i^{\sigma_t}(I, a)$$

Learning the Nash Equilibrium

Version 1: Counterfactual Regret Minimization (CFR)

- For $t = 1, \ldots, n$
  - Player 1 chooses strategy

$$\sigma_{1,t}(I,a) = \begin{cases} \frac{R^+_{1,t}(I,a)}{\sum_{b \in A_1} R^+_{1,t}(I,b)} & \text{if } \sum_{b \in A_1} R^+_{1,t}(I,b) > 0 \\ \frac{1}{A_1} & \text{otherwise} \end{cases}$$

  - Player 1 executes actions prescribed by $\sigma_{1,t}$ through a *full game*
  - Player 1 receives payoff $u_{1,t}$
  - Player 1 computes instantaneous regret $r_i^{\sigma_t}$ over information sets *observed over the game*

$R^+ = \max\{0, R\}$

Learning the Nash Equilibrium

Theorem

If CFR *is run over n steps, then the regret is bounded as*

$$R_n = \max_{\sigma_1} \frac{1}{n} \sum_{t=1}^{n} u_1(\sigma_1, \sigma_{2,t}) - \frac{1}{n} \sum_{t=1}^{n} u_1(\sigma_{1,t}, \sigma_{2,t}) \leq |\mathcal{I}_i| \sqrt{\frac{A_1}{n}}$$

Learning the Nash Equilibrium

Theorem

If CFR *is run over n steps, then the regret is bounded as*

$$R_n = \max_{\sigma_1} \frac{1}{n} \sum_{t=1}^{n} u_1(\sigma_1, \sigma_{2,t}) - \frac{1}{n} \sum_{t=1}^{n} u_1(\sigma_{1,t}, \sigma_{2,t}) \leq |\mathcal{I}_i| \sqrt{\frac{A_1}{n}}$$

Remarks

- $\lim_{n\to\infty} R_n \leq 0 \quad \Rightarrow$ *Hannan's consistency*
- Rate of convergence $O(1/\sqrt{n})$
- Linear dependence on the number of information sets
- In self-play EWF "converges" to the Nash equilibrium

Learning the Nash Equilibrium

Version 2: Counterfactual Regret Minimization+ (CFR^+)

- For $t = 1, \ldots, n$
  - *At t even* player 1 chooses strategy

$$\sigma_{1,t}(I,a) = \begin{cases} \frac{Q_{1,t}(I,a)}{\sum_{b \in A_1} Q_{1,t}(I,b)} & \text{if } \sum_{b \in A_1} Q_{1,t}(I,b) > 0 \\ \frac{1}{A_1} & \text{otherwise} \end{cases}$$

  - *At t odd* player 1 chooses strategy $\sigma_{1,t} = \sigma_{1,t-1}$
  - Player 1 executes actions prescribed by $\sigma_{1,t}$ through a full game
  - Player 1 receives payoff $u_{1,t}$
  - Player 1 computes instantaneous regret $r_i^{\sigma_t}$ over information sets *observed over the game*
- Return

$$\widehat{\sigma}_{1,n} = \sum_{t=1}^{n} \frac{2t}{n^2 + n} \sigma_{1,t}$$

$Q_{1,t} = (Q_{1,t-1} + r_i^{\sigma_{t-1}})^+$ instead of $R^+_{1,t} = (\sum_{s=1}^{t-1} r_i^{\sigma_s})^+$

Learning the Nash Equilibrium

Theorem

If CFR^+ *is run over n steps, then the regret is bounded as*

$$R_n = \max_{\sigma_1} \frac{1}{n} \sum_{t=1}^{n} u_1(\sigma_1, \sigma_{2,t}) - \frac{1}{n} \sum_{t=1}^{n} u_1(\sigma_{1,t}, \sigma_{2,t}) \leq |\mathcal{I}_i| \sqrt{\frac{A_1}{n}}$$

Learning the Nash Equilibrium

Theorem

If CFR^{+} *is run over n steps, then the regret is bounded as*

$$R_n = \max_{\sigma_1} \frac{1}{n}\sum_{t=1}^{n} u_1(\sigma_1, \sigma_{2,t}) - \frac{1}{n}\sum_{t=1}^{n} u_1(\sigma_{1,t}, \sigma_{2,t}) \leq |\mathcal{I}_i|\sqrt{\frac{A_1}{n}}$$

Remarks

- Same performance as CFR
- Empirically is more "reactive"
- Empirically $\widehat{\sigma}_{1,t}$ tends to converge

CFR in Large Problems: Heads-up Limit Texas Hold'em

The problem

- Four rounds of cards, four rounds of betting, *discrete bets*
- About 10^{18} states, 3.2×10^{14} information sets

CFR in Large Problems: Heads-up Limit Texas Hold'em

The problem

- Four rounds of cards, four rounds of betting, *discrete bets*
- About 10^{18} states, 3.2×10^{14} information sets

Abstraction: cluster together "similar" histories

- Symmetries (reducing to 10^{13} information sets)
- Clustering
  - Buckets based on (roll-out) hand strength
  - "Hierarchical" buckets (e.g., second hand is indexed by the first bucket as well)
  - About 1.65×10^{12} states, 5.73×10^{7} information sets

CFR in Large Problems: Heads-up Limit Texas Hold'em

The problem

- Four rounds of cards, four rounds of betting, *discrete bets*
- About 10^{18} states, 3.2×10^{14} information sets

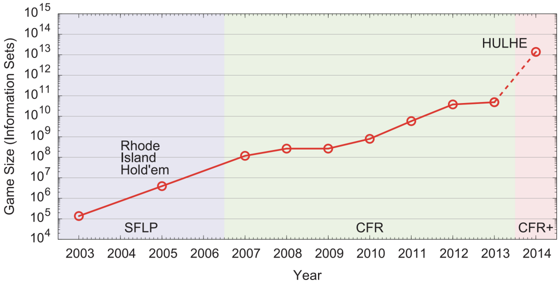
Abstraction: cluster together "similar" histories

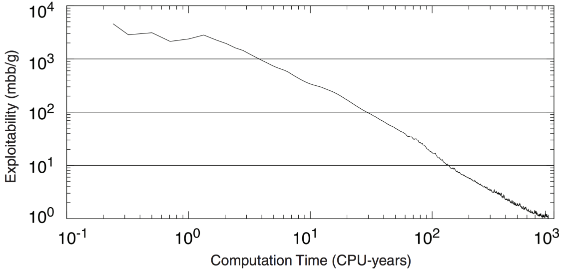
- Symmetries (reducing to 10^{13} information sets)
- Clustering
  - Buckets based on (roll-out) hand strength
  - "Hierarchical" buckets (e.g., second hand is indexed by the first bucket as well)
  - About 1.65×10^{12} states, 5.73×10^{7} information sets

Engineering:

- Rounding: $\sigma(a) = 0.0$ if smaller than threshold, fixed-point arithmetic
- Dynamic compression regret and strategy (from 262 TiB to 10.9 TiB)
- Distribute recursive computation of regret and strategy over rounds

CFR in Large Problems: Heads-up Limit Texas Hold'em





Heads-up No-Limit Texas Hold'em

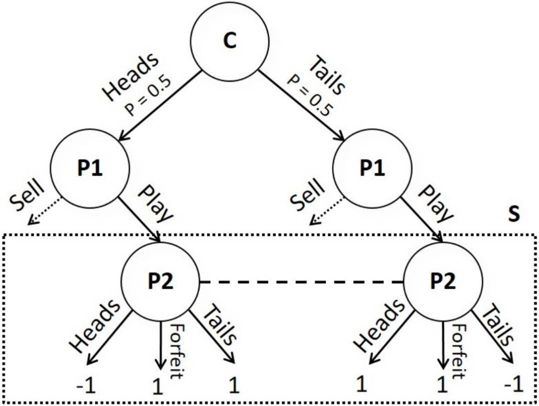
The problem

- In *no-limit* bets are arbitrary
- With standard discretized bets (1$ up to 20,000$) 10^{160} decision points!

The Learning problem

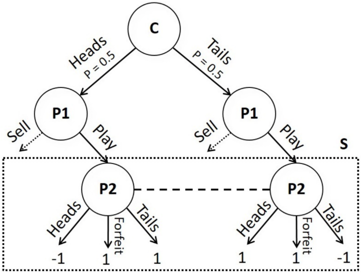
- "Simple" abstraction techniques no longer work
- Safe subgame solving

Subgame Solving in Imperfect Information Games



$$P1(\text{head}, \text{sell}) = 0.5\$, \; P1(\text{tail}, \text{sell}) = -0.5\$$$

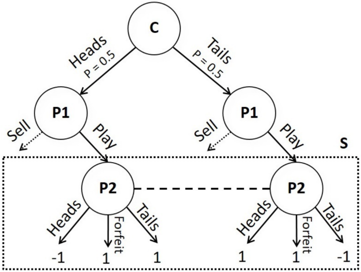
Subgame Solving in Imperfect Information Games



$$P1(\text{head}, \text{sell}) = 0.5\$, \; P1(\text{tail}, \text{sell}) = -0.5\$$$

- $\sigma_2 = \text{head} \Rightarrow \sigma_1(\text{head}) = \text{“Sell”}, \sigma_1(\text{tail}) = \text{“Play”}$
  $\Rightarrow u_1 = 0.5 \times 0.5 + 0.5 \times 1 = 0.75$
- $\sigma_2 = \text{tail} \Rightarrow \sigma_1(\text{head}) = \text{“Play”}, \sigma_1(\text{tail}) = \text{“Sell”}$
  $\Rightarrow u_1 = 0.5 \times 1 + 0.5 \times (-0.5) = 0.25$
- Optimal strategy $\sigma_2 = (0.25, 0.75)$

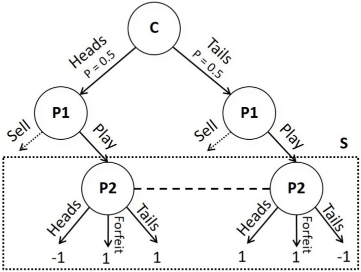
Subgame Solving in Imperfect Information Games



$$P1(\text{head}, \text{sell}) = -0.5\$, \; P1(\text{tail}, \text{sell}) = 0.5\$$$

- Optimal strategy $\sigma_2 = (0.25, 0.75)$

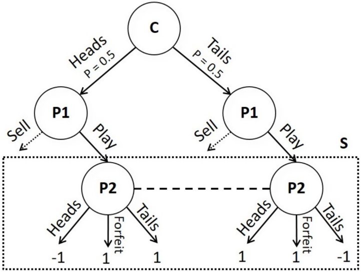
Subgame Solving in Imperfect Information Games



$$P1(\text{head}, \text{sell}) = -0.5\$, \; P1(\text{tail}, \text{sell}) = 0.5\$$$

- Optimal strategy $\sigma_2 = (0.75, 0.25)$

Subgame Solving in Imperfect Information Games



$$P1(\text{head}, \text{sell}) = -0.5\$,\ P1(\text{tail}, \text{sell}) = 0.5\$$$

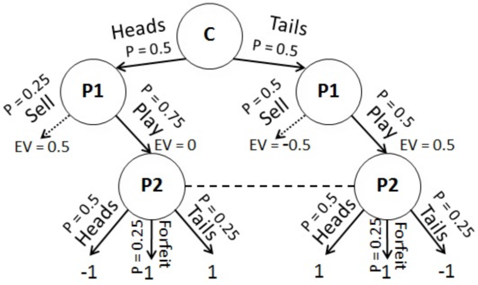
- Optimal strategy $\sigma_2 = (0.75, 0.25)$

$\Rightarrow$ the optimal solution of the subgame depends on "things" *outside* the subgame itself!

Subgame Solving in Imperfect Information Games

Version 1: unsafe subgame solving

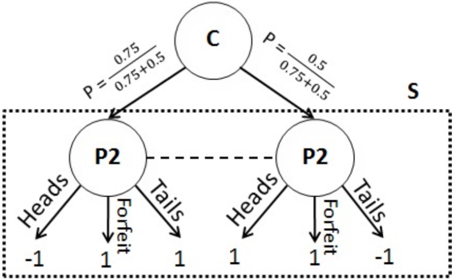
1. Start with a pre-computed solution (e.g., through abstraction)



Subgame Solving in Imperfect Information Games

Version 1: unsafe subgame solving

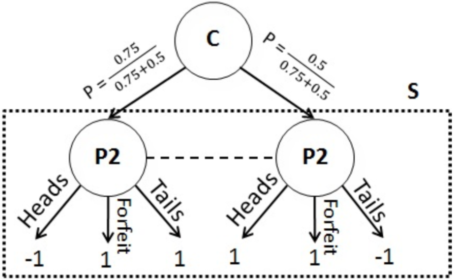
1. Start with a pre-computed solution (e.g., through abstraction) called *trunk*
2. Solve the subgame *as-if* everything else was as in the *trunk*



Subgame Solving in Imperfect Information Games

Version 1: unsafe subgame solving

1. Start with a pre-computed solution (e.g., through abstraction) called *trunk*
2. Solve the subgame *as-if* everything else was as in the *trunk*

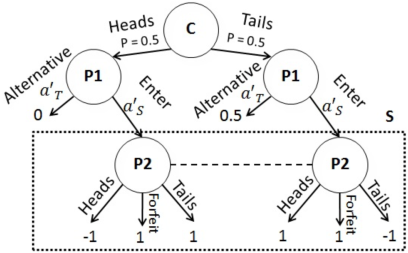


⇒ subgame strategy can be *arbitrarily bad*

Subgame Solving in Imperfect Information Games

Version 2: subgame re-solving

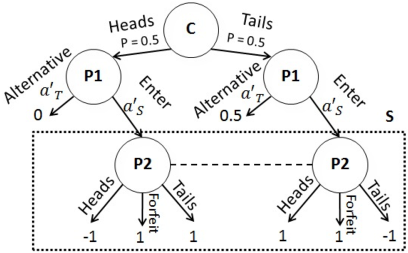
1. Start with a pre-computed solution (e.g., through abstraction) called *trunk*
2. Construct an augmented subgame giving $P1$ the chance to *opt-out* from the subgame and play in the trunk
3. Solve the augmented subgame with *maxmargin*



Subgame Solving in Imperfect Information Games

Version 2: subgame re-solving

1. Start with a pre-computed solution (e.g., through abstraction) called *trunk*
2. Construct an augmented subgame giving $P1$ the chance to *opt-out* from the subgame and play in the trunk
3. Solve the augmented subgame with *maxmargin*

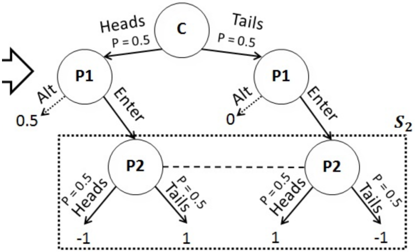


$\Rightarrow$ subgame strategy better but potentially far from optimal

Subgame Solving in Imperfect Information Games

Version 3: reach subgame solving

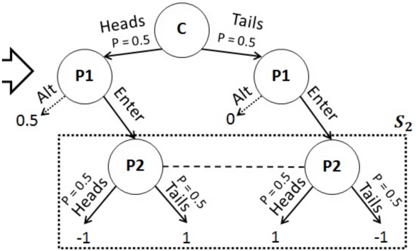
1. Start with a pre-computed solution (e.g., through abstraction) called *trunk*
2. Construct an augmented subgame considering the *gift* given to $P2$ (i.e., consider *any* possible action *not leading* to the subgame)
3. Solve the augmented subgame



Subgame Solving in Imperfect Information Games

Version 3: reach subgame solving

1. Start with a pre-computed solution (e.g., through abstraction) called *trunk*
2. Construct an augmented subgame considering the *gift* given to $P2$ (i.e., consider *any* possible action *not leading* to the subgame)
3. Solve the augmented subgame



⇒ provably reduce exploitability

Brains vs. AI

Libratus

- Monte-Carlo CFR + abstraction to compute the trunk
- Reach subgame solving with no abstraction (using CFR^+ to solve subgames) in-game

Brains vs. AI

Libratus

- Monte-Carlo CFR + abstraction to compute the trunk
- Reach subgame solving with no abstraction (using CFR^+ to solve subgames) in-game

Comptetition

- January 2017, over 20 days
- About 120,000 hands
- 4 top human players
- $200,000 prize

Brains vs. AI

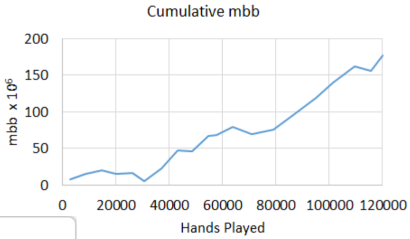
Libratus

- Monte-Carlo CFR + abstraction to compute the trunk
- Reach subgame solving with no abstraction (using CFR^+ to solve subgames) in-game

Comptetition

- January 2017, over 20 days
- About 120,000 hands
- 4 top human players
- $200,000 prize

Results



Summary

+ CFR^+ converges to Nash equilibria in self-play in imperfect-information extensive-form games
+ ReachSubgame provides a tool for safely decomposing the game
+ Efficient and (somehow) general purpose implementation
+ Beyond games: risk-averse planning

Summary

+ CFR^+ converges to Nash equilibria in self-play in imperfect-information extensive-form games
+ ReachSubgame provides a tool for safely decomposing the game
+ Efficient and (somehow) general purpose implementation
+ Beyond games: risk-averse planning

? Do we really care about (normal form) Nash?
? Beyond two-player games
? Opponent modeling
? Stochastic games (SG) / partially observable stochastic games (POSG)

Bibliography I

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire.
The nonstochastic multiarmed bandit problem.
SIAM J. Comput., 32(1):48–77, January 2003.

Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin.
Heads-up limit hold'em poker is solved.
Science, 2015.

Noam Brown and Tuomas Sandholm.
Safe and nested subgame solving for imperfect-information games.
CoRR, abs/1705.02955, 2017.

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

Nicolo Cesa-Bianchi and Gabor Lugosi.
Prediction, Learning, and Games.
Cambridge University Press, New York, NY, USA, 2006.

Bibliography II

Gergely Neu.
Explore no more: Improved high-probability regret bounds for non-stochastic bandits.
In *NIPS*, pages 3168–3176, 2015.

Wesley Tansey.
Counterfactual regret minimization for po.
https://github.com/tansey/pycfr, 2017.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione.
Regret minimization in games with incomplete information.
In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1729–1736. Curran Associates, Inc., 2008.

Learning in Zero-Sum Games

Alessandro Lazaric

lazaric@fb.com