Learning in Zero-sum games

Reinforcement Learning Seminar

Yingru Li

May 12, 2019

yingruli@link.cuhk.edu.cn
The Chinese University of Hong Kong, Shenzhen

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

Figure 1: Deep Blue

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

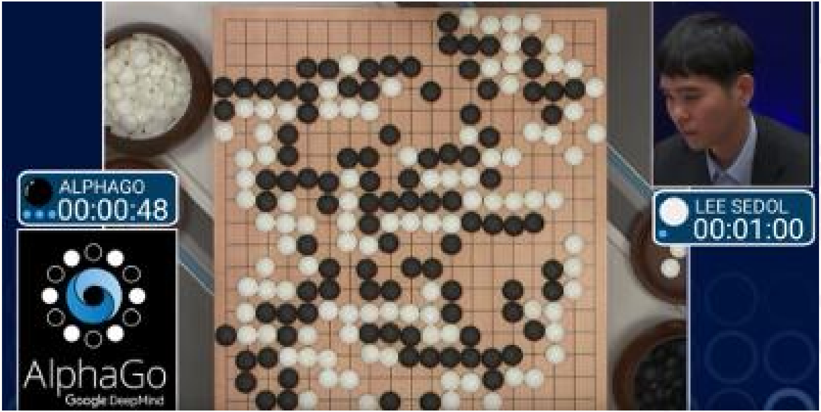

Figure 2: AlphaGo

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

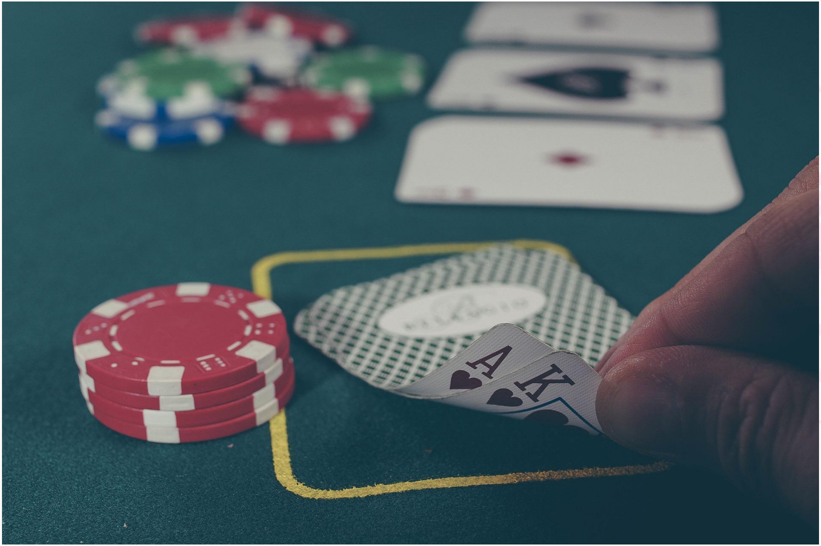

Figure 3: Libratus

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

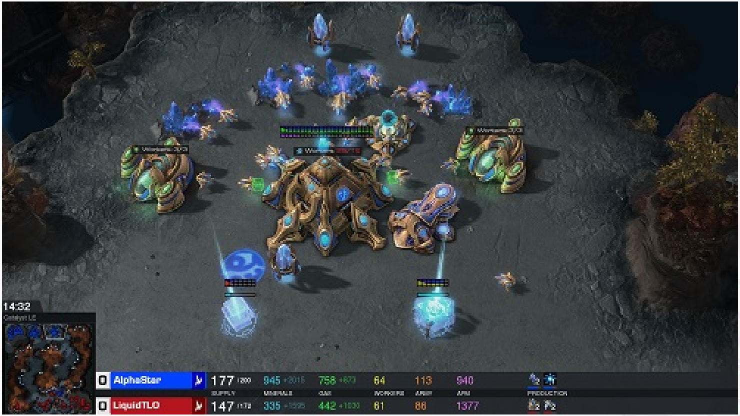

Figure 4: StarCraft II: A New Challenge for Reinforcement Learning. DeepMind AlphaStar Jan. 2019; Tencent AI Lab TStarBots Sep. 2018

...with Potential Applications in Real-World Environments

- Security
- Negotiation
- Diplomatic and Military Strategy
- Financial Market
- E-Commerce
- Distributed Cooperated and Competitive Robotics
- Game AI
- ......

Table of Contents

Table of Contents

Normal Form Games

The Game

- Set of players $N = \{1, \cdots, n\}$
- 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 $\sigma \in \mathcal{D}(A)$ is distribution over A, such that

$$\sigma(a) = \prod_{i=1}^{n} \sigma_i(a_i)$$

- Utility of a strategy for player i (expected utility):

$$u_i(\sigma) = \sum_{a_i} \sum_{a_{-i}} \sigma_i(a_i)\, \sigma_{-i}(a_{-i})\, u_i(a_i, a_{-i})$$

Normal Form Games

The Game

- Best response:

$$\sigma_i^* \in BR(\sigma_{-i}) \text{ iff } \forall \sigma_i \in \mathcal{D}(A_i), u_i(\sigma_i^*, \sigma_{-i}) \geq u_i(\sigma_i, \sigma_{-i})$$

- Nash equilibrium: σ is a Nash equilibrium iff $\forall i, \sigma_i \in BR(\sigma_{-i})$
- Every finite game has a Nash equilibrium! [Nash, 1950]

Finite 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, all other players' a_j
- Utility (payoff/reward) function $u : A \to \mathbb{R}^n$, player i's utility $u_i : A \to \mathbb{R}$

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

Mixed strategies

- Nash equilibrium [Minimax theorem (von Neumann, 1928)]

$$\begin{aligned}(\sigma_1^*, \sigma_2^*) &= \arg \max_{\sigma_1} \min_{\sigma_2} u_1(\sigma_1, \sigma_2) \\ &= \arg \min_{\sigma_1} \max_{\sigma_2} u_2(\sigma_1, \sigma_2)\end{aligned}$$

- Value of the game
  $V = \max_{\sigma_1} \min_{\sigma_2} u_1(\sigma_1, \sigma_2) = \min_{\sigma_2} \max_{\sigma_1} u_1(\sigma_1, \sigma_2)$

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

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

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

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

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

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

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

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

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

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

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

	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 $\sigma_2 = (\sigma_2(R), \sigma_2(P), \sigma_2(S))$ the strategy of player column,

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

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

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

$$1 = \sigma_2(R) + \sigma_2(P) + \sigma_2(S)$$

	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 $\sigma_2 = (\sigma_2(R), \sigma_2(P), \sigma_2(S))$ the strategy of player column,

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

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

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

$$1 = \sigma_2(R) + \sigma_2(P) + \sigma_2(S)$$

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

	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 $\sigma_2 = (\sigma_2(R), \sigma_2(P), \sigma_2(S))$ the strategy of player column,

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

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

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

$$1 = \sigma_2(R) + \sigma_2(P) + \sigma_2(S)$$

  - 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 $t = 1, \ldots, n$
  - 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., bandit feedback $u_{1,t}(a_{1,t})$)

The (Multi-Armed Bandit) Problem

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

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

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

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

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

Regret Minimization and Nash Equilibrium

Theorem

A learning algorithm is Hannan's consistent if

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

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

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

Furthermore, let empirical frequency strategies be

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

then the joint empirical strategy

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

Regret Minimization and Nash Equilibria [proof]

- Hannan's consistency

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

- linearity of utility function

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

Regret Minimization and Nash Equilibria [proof]

- Hannan's consistency

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

- linearity of utility function

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

- definition $u_{1,t}\left(\sigma_1\right) = u_1\left(\sigma_1, a_{2,t}\right) \Rightarrow$

$$\frac{1}{n} \sum_{t=1}^{n} u_{1,t}\left(\sigma_1\right) = \frac{1}{n} \sum_{t=1}^{n} \sum_{a_2 \in A_2} \mathbb{I}\left\{a_{2,t} = a_2\right\} u_1\left(\sigma_1, a_2\right) = \sum_{z_2 \in A_2} u_1\left(\sigma_1, a_2\right) \underbrace{\frac{1}{n} \sum_{t=1}^{n} \mathbb{I}\left\{a_{2,t} = a_2\right\}}_{\widehat{\sigma}_{2,n}(a_2)}$$

Regret Minimization and Nash Equilibria [proof]

- Hannan's consistency

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

- linearity of utility function

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

- definition $u_{1,t}\left(\sigma_1\right) = u_1\left(\sigma_1, a_{2,t}\right) \Rightarrow$

$$\frac{1}{n}\sum_{t=1}^{n} u_{1,t}\left(\sigma_1\right) = \frac{1}{n}\sum_{t=1}^{n}\sum_{a_2\in A_2} \mathbb{I}\left\{a_{2,t} = a_2\right\} u_1\left(\sigma_1, a_2\right) = \sum_{z_2\in A_2} u_1\left(\sigma_1, a_2\right) \underbrace{\frac{1}{n}\sum_{t=1}^{n}\mathbb{I}\left\{a_{2,t} = a_2\right\}}_{\widehat{\sigma}_{2,n}\left(a_2\right)}$$

- one-side of the result

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

- one-side of the result

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} u_{1,t}\left(a_{1,t}, a_{2,t}\right) \geq \max_{\sigma_1} \min_{\sigma_2} \frac{1}{n} \sum_{t=1}^{n} u_1\left(\sigma_1, \sigma_2\right) = V$$

- one-side of the result

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} u_{1,t}\left(a_{1,t}, a_{2,t}\right) \geq \max_{\sigma_1}\min_{\sigma_2} \frac{1}{n}\sum_{t=1}^{n} u_1\left(\sigma_1, \sigma_2\right) = V$$

- If player 2 also plays Hannan consistent strategies, then we get

$$\max_{\sigma_2} \frac{1}{n}\sum_{t=1}^{n} u_{2,t}\left(\sigma_2\right) \geq \max_{\sigma_2}\min_{\sigma_1} \frac{1}{n}\sum_{t=1}^{n} u_2\left(\sigma_1, \sigma_2\right) = V$$

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} u_{1,t}\left(a_{1,t}, a_{2,t}\right) \leq \min_{\sigma_2}\max_{\sigma_1} \frac{1}{n}\sum_{t=1}^{n} u_1\left(\sigma_1, \sigma_2\right) = V$$

-

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} u_{1,t}\left(a_{1,t}, a_{2,t}\right) = V \quad a.s.$$

Regret Minimization and Nash Equilibria

Remark

The joint empirical strategy converges to the set of correlated equilibrium almost surly as $n \to \infty$.

In particular, for any (finite) two-person zero-sum game, for each player, the empirical distribution of play converges to the set of optimal mixed actions.

$$\widehat{\sigma}_{1,n} \times \widehat{\sigma}_{2,n} \overset{n\to\infty}{\longrightarrow} \{(\sigma_1^*, \sigma_2^*)\}_{\text{Nash}} \quad a.s.$$

Note that approaching to a set does not imply convergence to particular point.

Corollary

If

$$R_n \leq \epsilon$$

then the joint empirical strategy is ϵ-Nash (more precisely, correlated ϵ-equilibrium), i.e.,

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

Hannan's Consistent Algorithms

A learning problem

- For $t = 1, \ldots, n$
  - 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

$$\limsup_{n \to \infty} R_n \leq 0 \quad a.s.$$

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

Remarks

- This strategy is easily exploitable $R_n = O(1)$
- E.g. Opponents set $u_{1,t}(a = a_{1,t}) = -1$ and $u_{1,t}(a \neq a_{1,t}) = 1$

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

Remarks

- This strategy is easily exploitable $R_n = O(1)$
- Self play does not converge in general [Recall Hannan's consistency]

Version 2: [Randomization]

Learning the Nash Equilibrium

Version 2: [Randomization] exponentially weighted forcaster (EWF)

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

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

  - 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 $u_{1,t}(a_{1,t})$

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

Learning the Nash Equilibrium

Theorem

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

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

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

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

Remarks

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

Learning the Nash Equilibrium

Version 2: [Randomization] exponentially weighted forcaster (EWF)

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

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

  - 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 $u_{1,t}(a_{1,t})$

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

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 $u_{1,t}\,(a_{1,t})$

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

Learning the Nash Equilibrium

Problem:

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

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

Solution:

- Importance sampling

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

- Unbiased estimator

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

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

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

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

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

  - Update weights $u_{1,t}\left(a_{1,t}\right)$

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

Theorem

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

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

Learning the Nash Equilibrium

Theorem

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

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

Remarks

- $\limsup_{n\to\infty} \overline{R}_n \leq 0 \quad \Rightarrow$ 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

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

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

- Variance

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

Solution:

- Bias both pseudo-payoff

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

- Mix strategy with uniform exploration (now bounded below)

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

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 $u_{1,t}\left(a_{1,t}\right)$
  - Compute pseudo-payoffs

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

  - Update weights $u_{1,t}\left(a_{1,t}\right)$

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

Lemma

For $\beta_t \leq 1$, let

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

Then, w.p. at least $1 - \delta$,

$$\sum_{t=1}^{n} u_{i,t}(a) \leq \sum_{t=1}^{n} \tilde{u}_{i,t}(a) + \frac{\log \delta^{-1}}{\beta_t}$$

Theorem

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

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

Remarks

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

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

Table of Contents

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 $P : H \rightarrow N \cup \{c\}$
- Set of information sets $\mathcal{I} = \{I\}$ (i.e., I is a subset of histories that are not distinguishable)
- Utility of a terminal history $u_i : Z \rightarrow \mathbb{R}$
- Strategy $\sigma_i : \mathcal{I} \rightarrow \mathcal{D}(A)$ (in all $h \in l$ such that $P(h) = i$)

Histories

- Prob. of reaching history $h \in H$ following joint strategy $\sigma, \pi^\sigma(h)$
- 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 $\delta(a), \sigma_{I \to a}$

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

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

- Empirical strategy:

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

Regret Minimization and Nash Equilibria

Theorem

A learning algorithm is Hannan's consistent if

$$\limsup_{n\to\infty} R_n \leq 0 \quad a.s.$$

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

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

Furthermore, the joint empirical strategy

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

Regret Matching Algorithm

- Back to Rock-Paper-Scissors
- Let $a_1 = \mathit{rock}$ and $a_2 = \text{paper}$
- Then the counterfactual regret

$$r\left(a_1 \rightarrow \mathit{rock}\right) = u_1\left(\mathit{rock}, a_{2,t}\right) - u_1\left(a_{1,t}, a_{2,t}\right) = -1 - (-1) = 0$$

$$r\left(a_1 \rightarrow \text{ paper }\right) = u_1\left(\text{ paper}, , a_{2,t}\right) - u_1\left(a_{1,t}, a_{2,t}\right) = 0 - (-1) = 1$$

$$r\left(a_1 \rightarrow \text{ scissors }\right) = u_1\left(\text{scissors}, a_{2,t}\right) - u_1\left(a_{1,t}, a_{2,t}\right) = 1 - (-1) = 2$$

- Regret matching idea

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

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

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

- Counterfactual regret of not taking a in h

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

- 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

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

  - 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

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

Remarks

- $\lim_{n\to\infty} R_n \leq 0 \quad \Rightarrow$ *Hannan's consistency*
- Rate of convergence $O(1/\sqrt{n})$
- Player 1 receives payoff $u_{1,t}$
- 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

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

  - 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

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

$$Q_{1,t} = \left(Q_{1,t-1} + r_i^{\sigma_{t-1}}\right)^+ \text{ instead of } R_{1,t}^+ = \left(\sum_{s=1}^{t-1} r_i^{\sigma_s}\right)^+$$

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

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

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

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