Solving Games

Combinatorial Algs for Matrix Games

Lemke - Howson (Last week)
PNS (Porter - Nudelman - Shoham)

Optimization

General Formulation → MIP
Zero - Sum Games .

Learning / Evolution

Fictitious Play ← Alpha Star
Regret Matching
└ Counterfactual Regret Minimization ← Deep Stack

Optimization

Alg 4 DM

$$\underset{\pi\ U}{\text{minimize}} \quad \sum_i \left(U^i - U^i(\pi)\right) \quad \text{0 when NE}$$

$$\text{subject to} \quad U^i \geq U^i(a^i, \pi^{-i}) \quad \text{Best Response} \quad \forall\, i, a^i$$

$$\sum \pi^i(a^i) = 1 \quad \text{''}$$

$$\pi^i(a^i) \geq 0 \quad \text{''}$$

Mixed - Integer Program

social optimum

$$\underset{\pi, u_a^i, u^i, b_a^i}{\text{maximize}} \quad 1 \leftarrow \sum_i u_i$$

subject to

$$\left.\begin{aligned} &\sum_a \pi^i(a) = 1 \quad \forall\, i, a \\ &\pi^i(a) \geq 0 \end{aligned}\right] \ \pi^i \text{ is prob}$$

linear equations that ~~enforce~~ encode indifference

$$u_a^i = \sum_{a^{-i}} \pi^{-i}(a^{-i})\, U^i(a, a^{-i})$$

$$u^i \geq u_a^i$$

regret →

$$r_a^i = u^i - u_a^i$$

defining regret

$$\pi^i(a) \leq 1 - b_a^i \quad \leftarrow \text{if inactive } \pi^i(a) = 0,$$

$$r_a^i \leq U_i\, b_a^i \quad \leftarrow \text{if active, regret is zero}$$

$$b_a^i \in \{0, 1\}$$

active ↑ ↑ inactive

defining which actions are active in NE

$U_i \equiv$ max difference between utilities for player i

Only NE are feasible solutions

Can change objective to find NE we want, e.g. social optimum, fairest

Outperforms PNS when supports are large

$$\min_{x,y,u,v} \; u - \cancel{x^T A}\, y + v - \cancel{x^T B}\, y$$

x y
A B

Subject to $u \geq \hat{x}^T A y$ $\forall$ one-hot vectors $\hat{x}$

$v \geq x^T B \hat{y}$ $\forall$ " " $\hat{y}$

$\sum_i x_i = 1$

$x \geq 0$

$\sum y_i = 1$

$y \geq 0$

Zero sum $B = -A$

Linear Program

$$\min_{x,v} \; v$$

such that $v \geq x^T B \hat{y}$

$\sum_i x_i = 1$

$x \geq 0$

Learning Methods

Fictitious Play

Initialize $N^i(a) \leftarrow 0$ $\forall i, a$

for $t \in 1..T$

$\pi^i_t \leftarrow \text{normalize}(N^i)$ $\forall i$

$a^i \leftarrow \text{best-response}(\pi^{-i})$

$N^i(a^i) \leftarrow N^i(a^i) + 1$

return π_T

Does not always converge

Guaranteed to converge in 2player games if

- Constant sum
- nondegenerate 2xn — 2005
- potential game
- solvable by iterated elimination of strictly-dominated strategies

0,0	2,1	1,2
1,2	0,0	2,1
2,1	1,2	0,0

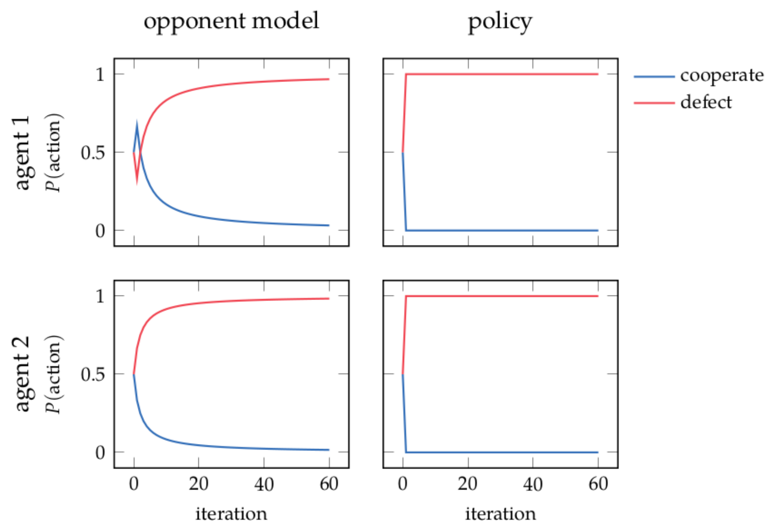


Figure 24.2. Two fictitious play agents learning and adapting to one another in a prisoner's dilemma game. The first row illustrates agent 1's learned model of 2 (left) and agent 1's policy (right) over iteration. The second row follows the same pattern, but for agent 2. To illustrate variation in learning behavior, the initial counts for each agent's model over the other agent's action were assigned to a random number between 1 and 10.

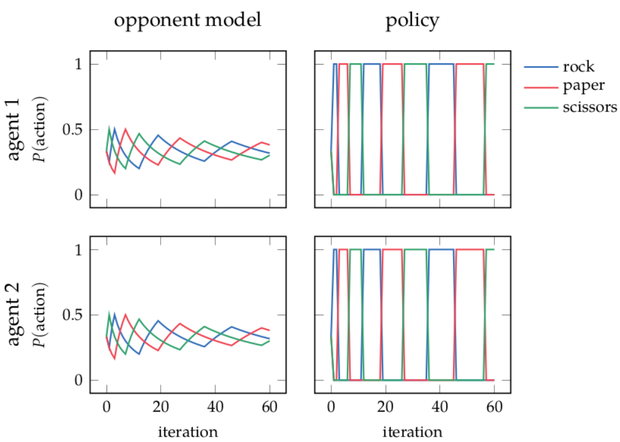
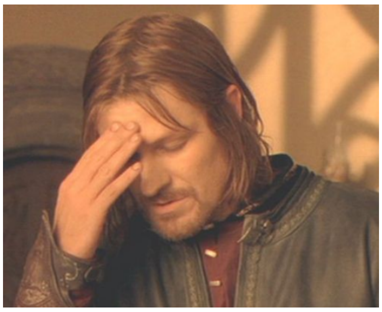


Figure 24.3. A visualization of two fictitious play agents learning and adapting to one another in a rock-paper-scissors game. The first row illustrates agent 1's learned model of 2 (left) and agent 1's policy (right) over time. The second row follows the same pattern, but for agent 2. To illustrate variation in learning behavior, the initial counts for each agent's model over the other agent's action were assigned to a random number between 1 and 10. In this zero-sum game, fictitious play agents approach convergence to their stochastic policy Nash equilibrium.

Regret

How much better one could have done by taking another action

$$R^i(a^i, \pi) = U^i(a^i, \pi^{-i}) - U^i(\pi)$$



		Odd	
		H	T
Even	H	1,-1	-1,1
	T	-1,1	1,-1

$\pi^1 = [0.5, 0.5]$
$\pi^2 = [0.5, 0.5]$

$$R^1(H,\pi) = (0.5\cdot 1 + 0.5\cdot -1) - (0.25\cdot 1 + 0.25\cdot -1 + 0.25\cdot 1 + 0.25\cdot -1) = 0$$
$$R^1(T,\pi) = 0$$
$$R^2(H,\pi) = 0$$
$$R^2(T,\pi) = 0$$

$\pi^1 = [1,0]$
$\pi^2 = [1,0]$

$$R^1(H,\pi) = (1) - (1) = 0$$
$$R^1(T,\pi) = (-1) - (1) = -2$$
$$R^2(H,\pi) = 0$$
$$R^2(T,\pi) = (1) - (-1) = 2$$

Regret Matching aka. Blackwell's algorithm

Initialize $\bar{\pi}^i_0 \leftarrow [0,0,0\ldots]$ (strat. prof. sum) $\quad \bar{R}^i_0 \leftarrow [0,0,0\ldots]$ (cumulative regret) $\quad \forall i$

for $t \in 1..T$

$\quad \pi^i_t \leftarrow \text{normalize}(\bar{R}^i_t) \quad \forall i$ — if argument is 0, return uniform

$\quad \bar{\pi}_t \leftarrow \bar{\pi}_{t-1} + \pi_t$ — does not converge to Nash

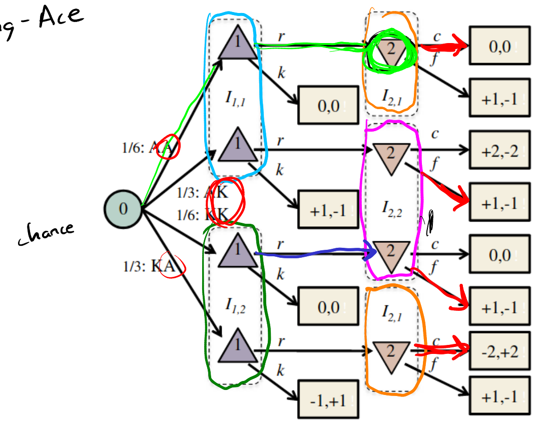
$\quad R^i_t(a) \leftarrow \max(U^i(a, \pi^{-i}_t) - U^i(\pi_t), 0) \quad \forall i, a$

$\quad \bar{R}^i_t \leftarrow \bar{R}^i_{t-1} + R^i_t \quad \forall i$

return normalize$(\bar{\pi}_t)$ — converges for some classes e.g. zero sum

Extensive - Form Games

King-Ace



h = history : sequence of actions
node in tree

I = information set

histories in the same I are indistinguishable to that player

policy: mapping from each information set to a distribution over actions

Evaluating strategies

$P_\pi(h) \equiv$ prob. of reaching h under π

$$U(\pi) = \sum_{h \in Z} U(h)\, P_\pi(h)$$

Z ← terminal

	cc	cf	ff	fc
rr	0	-1/6	1	7/6
kr	-1/3	-1/6	5/6	2/3
rk	1/3	0	1/6	1/2
kk	0	0	0	0

Counterfactual Regret Minimization (CFR)

Key Idea: break overall regret into terms for each I that can be added together to bound overall regret

Counterfactual Utility

$$U^i(\pi, I) = \frac{\sum_{h \in I, h' \in Z} P_\pi^{-i}(h) P_\pi(h,h') U^i(h')}{P_\pi^{i}(I)}$$

$P_\pi(h,h') \equiv$ probability of going from h to h' under π

$P_\pi^{-i}(h) \equiv$ probability of reaching h if i deliberately tries to reach h, everyone else plays according to π

Counterfactual Regret

$$R_\pi^i(I,a) = P_\pi^i(I)\left(U^i(\pi|_{I \to a}, I) - U^i(\pi, I)\right)$$

$\pi|_{I\to a} \equiv$ play w/ policy π except at I, take a

$$\to \bar{R}^i_{t,imm}(I) = \frac{1}{t} \max_{a \in A(I)} \sum_{\tau=1}^{t} R^i_{\pi_\tau}(I,a)$$

$$\bar{R}^{i+}_{t,imm} = \max\left(\bar{R}^i_{t,imm}(I), 0\right)$$

Theorem $\bar{R}^i_t \leq \sum_I \bar{R}^{i+}_{t,imm}(I)$ [Zinkevich et al. 07]

CFR algorithm: Apply regret matching at each I with regret $R_\pi^i(I,a)$

In 2007, abstracted limit Texas Hold em

10^{18} game states $\to$ 10^{12} in abstraction

Poker-specific optimization: 18.5k reachable states
6.5k reachable info sets

750 iterations / sec on single core