Algorithmic Game Theory

Learning in Games

Viliam Lisý

Artificial Intelligence Center
Department of Computer Science, Faculty of Electrical Engineering
Czech Technical University in Prague

(May 11, 2018)

Plan

Online learning and prediction

- single agent learns to select the best action

Learning in normal form games

- **the same algorithms used by multiple agents**

Learning in extensive form games

- generalizing these ideas to sequential games

DeepStack

Algorithmic Game Theory

Learning in Normal Form Games

Viliam Lisý

Artificial Intelligence Center
Department of Computer Science, Faculty of Electrical Engineering
Czech Technical University in Prague

(May 9, 2017)

Introduction

How may simple learning agents achieve equilibrium outcomes?

Best Response Dynamics (Fictitious play)

- best response to average empirical play
- needs to know the game

No-Regret Dynamics

- each player uses no-regret algorithm
- may now only their own actions and received rewards

Best response dynamics

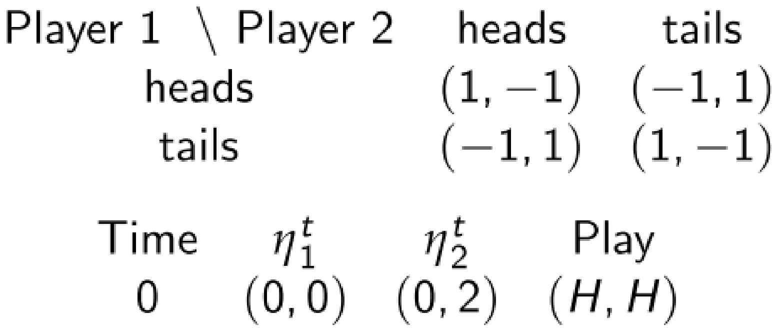
Fictitious play

Players maintain empirical distribution of past opponent’s actions

$$\bar{\sigma}^{T}_{-i} = \frac{1}{\mathrm{T}}\sum_{t=1}^{T}\sigma^{t}_{-i}$$ (often in form of frequencies η_i^T)

In each round, each player plays BR to these distributions

$$\sigma_i^t = \arg\max_{a_i \in A_i} U_i(a_i, \bar{\sigma}^{t}_{-i})$$



Player 1 \ Player 2	heads	tails
heads	$(1, -1)$	$(-1, 1)$
tails	$(-1, 1)$	$(1, -1)$

Time	η_1^t	η_2^t	Play
0	$(0, 0)$	$(0, 2)$	(H, H)

Result of FP in case of convergence

Theorem: If the empirical action frequencies of fictitious play converge ($\bar{\sigma}^t \ \rightarrow \sigma^*$) they converge to the Nash equilibrium of the game.

Proof: For contradiction assume σ^* is not a NE.

Then there exists player i and actions $a_i, a_i' \in A_i \ \ \sigma^*(a_i) > 0$, such that $U_i(a_i', \sigma^*_{-i}) > U_i(a_i, \sigma^*_{-i})$.

Choose ϵ, such that $0 < \epsilon < \frac{1}{2}(U_i(a_i', \sigma^*_{-i}) - U_i(a_i, \sigma^*_{-i}))$.

Since ($\bar{\sigma}^t \ \rightarrow \sigma^*$), there is T,such that for all $t > T$

$$\forall a_{-i} \in A_{-i} \ : \ |\bar{\sigma}^t_{-i}(a_{-i}) \ - \sigma^*_{-i}(a_{-i})| < \epsilon.$$

For all $t > T$, we have $U_i\left(a_i, \bar{\sigma}^t_{-i}\right) = \sum_{a_{-i}} U_i(a_i, a_{-i})\, \bar{\sigma}^t_{-i}(a_{-i}) \leq \sum_{a_{-i}} U_i(a_i, a_{-i})\, \sigma^*_{-i}(a_{-i}) + \epsilon < \sum_{a_{-i}} U_i(a'_i, a_{-i})\, \sigma^*_{-i}(a_{-i}) - \epsilon \leq \sum_{a_{-i}} U_i(a_i, a_{-i})\, \bar{\sigma}^t_{-i}(a_{-i}) = U_i\left(a'_i, \bar{\sigma}^t_{-i}\right)$.

Hence a_i is not played after T, which contradicts $\sigma^*(a_i) > 0$.

Convergence of FP

Theorem: The empirical frequencies of FP converge to NE in

- constant-sum games
- two player games where each player has up to two actions
- games solvable by iterated strict dominance
- identical interest games
- potential games

Why it may not converge?

Shapley’s example in a modified rock-paper-scissors:

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

Unique NE is the uniform strategy for both players.

Let $\eta_1^0 = (1,0,0)$ and $\eta_2^0 = (0,1,0)$.

Play may be (P,R),(P,R)… for k steps until column switches to S.

Then (P,S) follows until row switches to R (for βk steps, $\beta > 1$).

Then (R,S) follows until column switches to P (for $\beta^2 k$ steps).

The play cycles among all 6 non-diagonal profiles with periods of ever-increasing length, hence, the empirical frequencies cannot converge.

Convergence of FP

Theorem (Brandt, Fischer, Harrenstein, 2010): In symmetric two-player constant-sum games, FP may require exponentially many rounds (in the size of the representation of the game) before an equilibrium action is eventually played. This holds even for games solvable via iterated strict dominance.

Proof:

	a	b	c
a	0	-1	-ϵ
b	1	0	-ϵ
c	ϵ	ϵ	0

With $\epsilon = 2^{-k}$, FP may take 2^k rounds to play the equilibrium action c for the first time.

(a,a),(b,b),…,(b,b)

$2^k - 1$ times

No-Regret Learning Summary

Immediate regret at time t for not choosing action i

$$r^t(i) = u^t(i) - \sigma^t \cdot u^t$$

Cumulative external regret for playing $\sigma^0, \sigma^1 \dots \sigma^T$

$$R^T = max_{i \in A} \sum_{t=0}^{T} r^t(i) = max_{i \in A} \sum_{t=0}^{T} u^t(i) - \sum_{t=0}^{T} \sigma^t \cdot u^t$$

Average external regret for playing $\sigma^0, \sigma^1 \dots \sigma^T$

$$\bar{r}^T = \frac{1}{T} R^T$$

An algorithm has **no regret** if for any $u^0, u^1 \dots u^T$ produces $\sigma^0, \sigma^1 \dots \sigma^T$ such that $\bar{r}^T \rightarrow 0$ as $T \rightarrow \infty$.

From External to Swap Regret

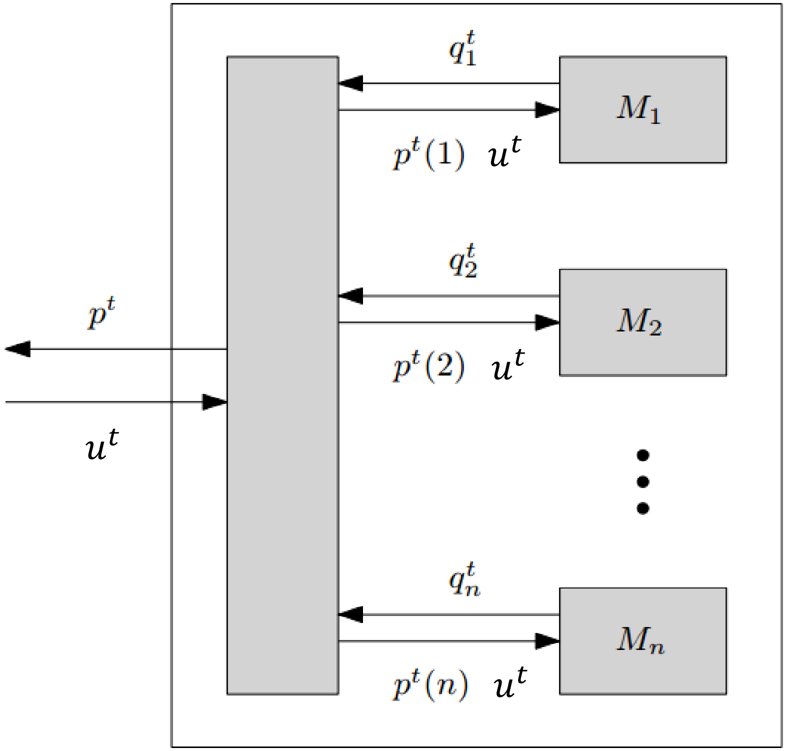
Cumulative swap regret for playing $\sigma^0, \sigma^1 \ldots \sigma^T$

$$R^T = max_{\delta:A\rightarrow A} \sum_{t=0}^{T} \sum_{i\in A} \sigma^t(i)(u^t(\delta(i)) - u^t(i))$$

From External to Swap Regret

Theorem (Blum & Mansour 2007):If there is a no-external-regret algorithm for a setting, there is also a no-swap-regret algorithm.

Proof: Polynomial black-box reduction.



From External to Swap Regret

Proof: Average expected reward of the overall algorithm

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} p^t(i)\, u^t(i)$$

Algorithm M_j choses $q_j^1, \dots, q_j^T$ and gets $p^1(j)u^1, \dots, p^T(j)u^T$. Thus

$$\forall k \in A\colon \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} q_j^t(i)\,(p^t(j)\,u^t(i)) \geq \frac{1}{T}\sum_{t=1}^{T} p^t(j)u^t(k) - \bar{r}_{\mathrm{j}}$$

Fix an arbitrary $\delta\colon A \to A$ and sum over all $j \in A$:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{j=1}^{n} q_j^t(i)(p^t(j)\,u^t(i)) \geq \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{n} p^t(j)u^t\big(\delta(j)\big) - \sum_{j=1}^{n}\bar{r}_j$$

From External to Swap Regret

We are done if we ensure

$$p^t(i) = \sum_{j=1}^{n} q_j^t(i) p^t(j)$$

This is true if p^t is the eigenvector of matrix given by q_j^t for $\lambda = 1$.

Equivalently, p^t are the stationary distribution of Markov chain.

Such vector p^t always exists and can be easily found!

From External to Swap Regret

Corollary: Let $\overline{r_M}(t) \to 0$ be the external regret convergence bound for a base algorithm used in the black-box reduction with $|A|$ actions. Than the swap regret of the overall algorithm is

$$\overline{r_{sw}}(T) \leq |A|\overline{r_M}(T).$$

Corollary: The black-box reduction with Hedge for all actions produces an algorithm with $\overline{r_{sw}}(T) \leq 2|A|\sqrt{\ln |A| /T}$.

No-Regret Dynamics – full information

Definition:

1) Each player i choses independently a mixed strategy σ_i^t using a no-regret algorithm.

2) Each player receives for all $a_i \in A_i$ rewards

$$u_i^t(a_i) = \mathbf{E}_{a_{-i} \sim \sigma_{-i}}[U(a_i, a_{-i})]$$

No-Regret Dynamics – full information

Theorem: If after T iterations of no-regret dynamics each player has external regret lower then ϵ than $\sigma = \frac{1}{T}\sum_t^T \sigma^t$, where $\sigma^t = \prod_{i=1}^k \sigma_i^t$, is an ϵ-coarse correlated equilibrium of the game. I.e., for any $a_i' \in A_i$

$$\mathbf{E}_{a\sim\sigma}[U_i(a)] \geq \mathbf{E}_{a\sim\sigma}[U_i(a_i', a_{-i})] - \epsilon$$

Corollary: If we run Hedge in a game with less than $|A|$ actions for each player for T iterations, the resulting average strategy is an ($\sqrt{ln(|A|)/T}$)-coarse correlated equilibrium of the game.

Corollary: If we run regret matching+ in a game with less than $|A|$ actions for each player for T iterations, the resulting average strategy is an ($\sqrt{|A|/T}$)-coarse correlated equilibrium of the game.

Minimax Theorem

Note: In zero-sum games, coarse correlated equilibria are Nash.

Theorem (Minimax Theorem): For any matrix game G

$$\max_x \min_y x^T G y = \min_y \max_x x^T G y$$

Proof: For contradiction assume that for some $\alpha > 0$

$$\max_x \min_y x^T G y < \min_y \max_x x^T G y - \alpha\,.$$

Set $\epsilon = \frac{\alpha}{2}$ and let both players run Hedge for time $\tau = 2 \ln n / \epsilon^2$.
Let $\hat{x}, \hat{y}$ be the empirical frequencies of their play and v the average reward of the maximizer.

$$\max_x \min_y x^T G y \geq \min_y \hat{x}^T G y \geq v - \epsilon \geq \max_x x^T G \hat{y} - 2\epsilon \geq \min_y \max_x x^T G y - \alpha$$

No-Regret Dynamics

Theorem: If after T iterations of no-regret dynamics each player has swap regret lower then ϵ than $\sigma = \frac{1}{T}\sum_t^T \sigma^t$, where $\sigma^t = \prod_{i=1}^k \sigma_i^t$, is an ϵ-correlated equilibrium of the game. I.e., for any player i and switching function $\delta : A \to A$

$$\mathbf{E}_{a\sim\sigma}[U_i(a)] \geq \mathbf{E}_{a\sim\sigma}[U_i(\delta(a_i), a_{-i})] - \epsilon$$

No-Regret Dynamics – bandit case

Definition:

1) Each player i choses independently a mixed strategy σ_i^t using a no-regret algorithm and independently samples $a_i \sim \sigma_i^t$.

2) Each player receives single reward $u_i^t(a_i) = U(a_i, a_{-i})$

No-Regret Dynamics – bandit case

Theorem: For any $p \in (0,1)$ there are parameters for Exp3.P, such that if both players use Exp3.P to choose their actions for T time steps then $\sigma = \frac{1}{T}\sum_t^T \sigma^t$, where $\sigma^t = \prod_{i=1}^k \sigma_i^t$, is an ϵ-coarse correlated equilibrium of the game with probability at least p and

$$\epsilon = 5.15\sqrt{\frac{|A|}{T}\ln\frac{|A|}{1-\sqrt{p}}}.$$

Proof sketch: It is enough to run Exp3.P for long enough so that both players have regret lower then ϵ at once with high probability. It can be achieved by using Exp3.P convergence bound with $\delta = 1 - \sqrt{p}$.

References

Asu Ozdaglar. 6.254 : Game Theory with Engineering Applications. Lecture 11: Learning in Games. March 11, 2010.

Brandt, Felix, Felix Fischer, and Paul Harrenstein. "On the rate of convergence of fictitious play." International Symposium on Algorithmic Game Theory. Springer Berlin Heidelberg, 2010.

T. Roughgarden, “Lecture Notes: Algorithmic Game Theory,” tech. rep., Stanford, 2013.