Playing Games without Observing Payoffs

Michal Feldman

Hebrew University &
Microsoft Israel R&D Center

Joint work with **Adam Kalai** and **Moshe Tennenholtz**

FLA-TAK-BONG-DING


FLA


TAK


BONG


DING

鲍步

爱丽丝


FLA

10 Y


TAK


TAK

10 Y


FLA


TAK

5 Y


BONG


DING

0 Y


DING

FLA-TAK-BONG-DING

爱丽丝

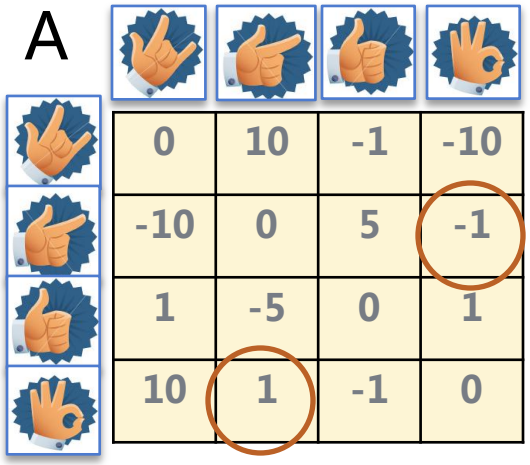



	FLA	TAK	BONG	DING
FLA	0	10	-1	-10
TAK	-10	0	5	-1
BONG	1	-5	0	1
DING	10	1	-1	0

鲍步

Properties of Fla-Tak-Bong-Ding

A



0	10	-1	-10
-10	0	5	-1
1	-5	0	1
10	1	-1	0

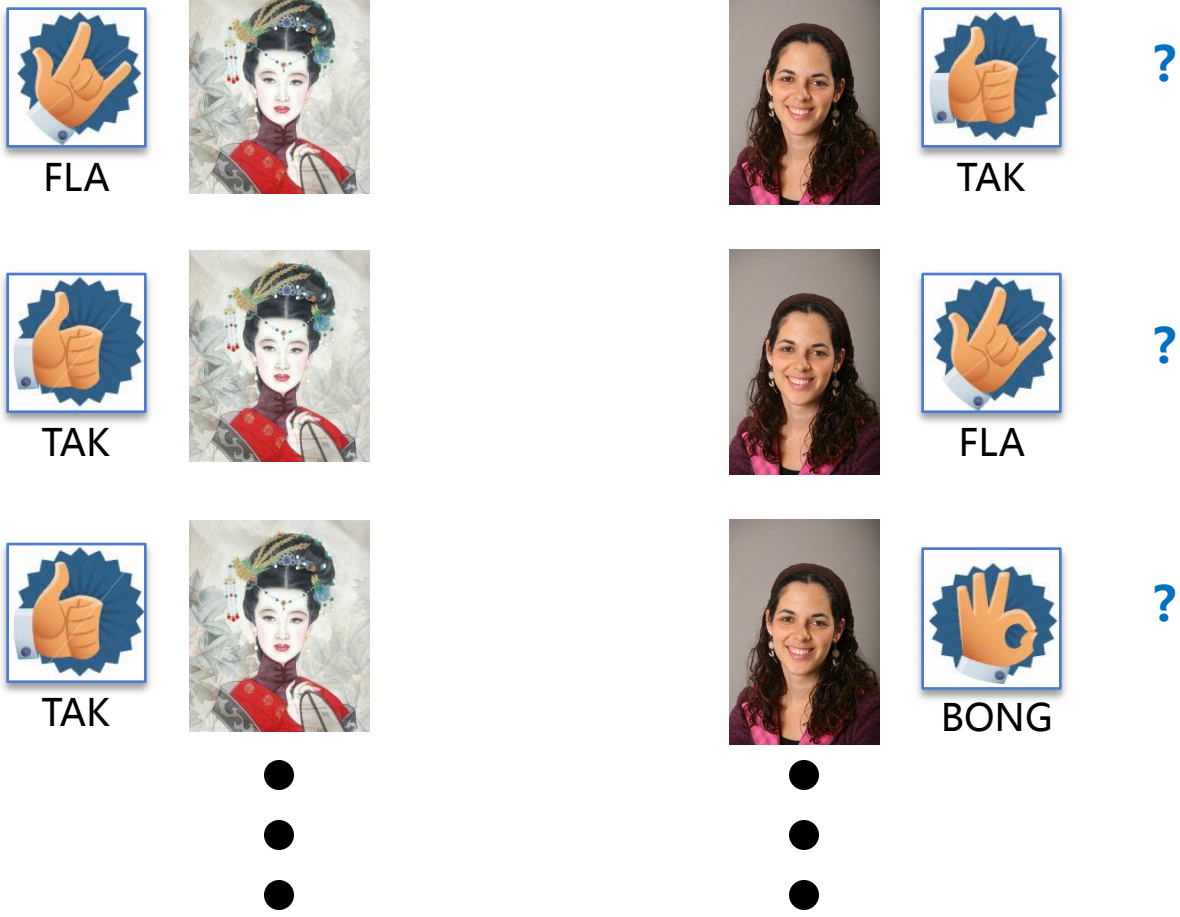
- Zero-sum
  - The benefit of one player is the loss of the other
- Symmetric
  - The two players have the same set of strategies
  - Their payoffs remain the same if their roles are reversed
- Symmetric zero-sum games
  - A(i,j) = -A(j,i)
- Each player can guarantee to herself the *value* of the game (zero) by playing the *minimax strategy*

A Visit to Beijing, 2010

0	10	-1	-10
-10	0	5	-1
1	-5	0	1
10	1	-1	0

Welcome to Beijing.
Want to play
Fla-Tak-Bong-Ding ?

mmm..
sure…



Can one perform well in a repeated symmetric game without observing a single payoff?

Intuition: **Mimic** Observed Actions

This is easy in a non-competitive environment

But is it possible to mimic an adversary, who knows he is being mimicked, and reacts to that?

Motivation: Limited Feedback

- Limited feedback from business choices
  - Example: companies make daily decisions about online advertising (e.g., choose ad location)
    - Companies often mimic the advertising campaign of a more experienced rival
    - Measuring the effect of a campaign is difficult (net profit is influenced by many factors, and it's difficult to assess how much is due to product design vs. marketing)
    - Newcomer cannot afford to invest in research or wait until they learn consumer behavior
    - Newcomer needs function effectively when competing with an existing well-informed company

Motivation: Limited Feedback (cont)

- Limited feedback from social behavior
  - Example: choose how to dress
- Sometimes feedback comes too late
  - Example: a politician gives a sequence of speeches

The Model

- Two-player, symmetric, zero-sum game, given by an n x n payoff matrix $A=\{a_{ij}\}$
  - Legal actions are $\{ 1,2, \dots , n \}$
  - Payoffs of (i,j) are $(a_{ij}, a_{ji}) \in R^2$, such that $a_{ij}+a_{ji}=0$
- The game A is finitely or infinitely repeated
- One player is informed, other is uninformed
  - Informed player knows A
  - Uninformed player does not know A and never observes a single payoff
- History on period t: sequence of actions played on periods 1, … , t-1
  - Observed by both informed and uninformed players
- Strategy: mapping from finite history to a probability distribution over $[n]$

Related Models: Imperfect Monitoring

- It is known that (almost) the value of the game can be achieved in the following settings of imperfect monitoring:
  - Adversarial multi-armed bandit problem [Auer,Cesa-Bianchi,Freund,Schapire, 2000]
    - You observe your realized payoff every period, but not the opponent's action
    - Similar results by [Megiddo, 1979] and [Banos, 1968]
  - Bayesian non-symmetric settings [Aumann&Maschler, 1968]
    - You observe the opponent's action, but not your realized payoff
- Our work complements the above literature
  - non-Bayesian settings, where uninformed player observes opponent's actions but not realized payoffs

PROPOSED STRATEGIES

- **Copycat #1**: tit-for-tat (i.e., copy opponent's play on previous round)
  - may fail in every round
    e.g., **R**ock-**P**aper-**S**cissors

R	P	S	R	P	S	R	
?	R	P	S	R	P	S	

- **Copycat #2**: copy the opponent's empirical frequency of play (fictitious play)
  - may fail badly too

R R R R R R R R R R P P P P P P P P P P

Message: one needs to be careful about how one mimics an opponent who known he is being mimicked.
A poor copycat may perform worse than making random decisions.

How to be a Strategic Copycat?

- The idea: for each pair of actions $i,j \le n$, ensure entry (i,j) is played (almost) as often as (j,i) is played
- $c_t(i,j)$ = number of periods entry (i,j) has been played in rounds $1, \ldots, t-1$
- $\Delta_t(i,j) = c_t(j,i) - c_t(i,j)$

Copycat strategy:

- On period t=1: play arbitrarily
- On period t=2,3,..
  - **Imagine** you are playing the symmetric zero-sum "pretend" game depicted by Δ_t
  - Play the mini-max strategy of Δ_t

COPYCAT STRATEGY


FLA


TAK


TAK


FLA


TAK


BONG

Δ_{θ}

	0	0	0	0
	0	0	0	0
	0	0	0	-1
	0	0	1	0

Main Result

- **Theorem**: for any symmetric *n* x *n* zero-sum game A, and any number of periods T ≥ 1, the copycat strategy ensures:

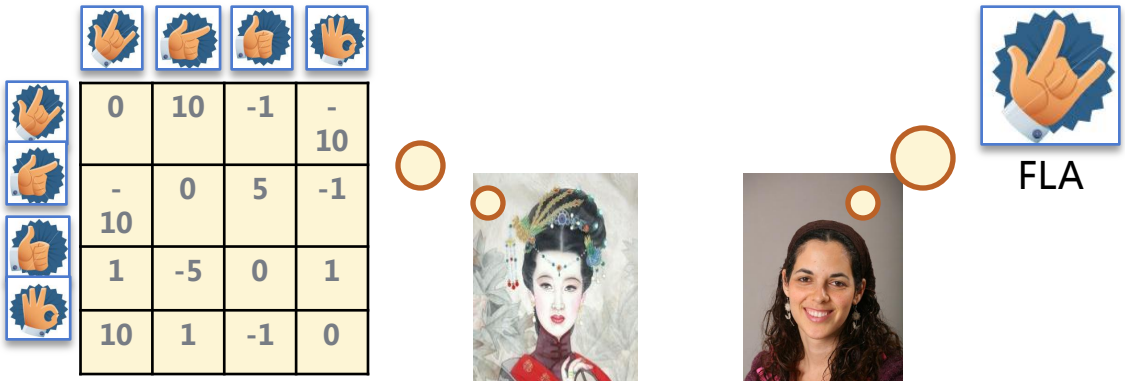
$$E[|\frac{1}{T}\sum_{t=1}^{T} A(i_t, j_t)|] \leq \frac{n}{\sqrt{2T}} \max_{i,j} |a_{i,j}|$$

The expected average payment of a copycat player

Copycat guarantees to the uninformed player (almost) the value of the game

EXTENSIONS

- General symmetric game
  - Copycat guarantees to the uninformed player (almost) the same expected payoff as that of the informed player
  - Consider the game A′ (i,j)=A(i,j)-A(j,i)
- What if even the set of actions is unknown?
  - Copycat is a strategy that uses only actions observed so far
  - Copycat delivers the same guarantees even if only a single "starting" strategy is known



Achieving Optimal Social Welfare

Theorem : In any two-player infinitely repeated symmetric game with one informed player and one uninformed player, it is possible to achieve the optimal social welfare in an (epsilon) learning equilibrium*

* Learning equilibrium: a pair of algorithms such that the algorithms themselves are in equilibrium. This is a non-Bayesian eq. notion [Brafman&Tennenholtz' 04]

Achieving Optimal Social Welfare

	j		
0,0	2,9	3,4	2,2
9,2	1,1	1,2	0,8
4,3	2,1	0,0	4,2
2,2	8,0	2,4	5,5

- (i,j) = entry maximizing sum of payoffs^{i}
- Players maximize social welfare by alternating between playing (i,j) and (j,i)
- Learning equilibrium:
  - Informed player:
    - Play i,j,i,j,… as long as protocol is followed
    - If protcol not followed: punish with safety level
  - Uninformed player:
    - Play ? in first iteration
    - Copy the last play of the informed player as long as protocol is followed
    - If protocol not followed: play **copycat**

Conclusion

- It is possible to strategically copy an adversary in symmetric games, even without observing a single payoff
- It is possible to achieve optimal welfare in epsilon- learning equilibrium in infinitely repeated symmetric games when one of the players is uninformed
- These results further our understanding of the landscape of optimization under uncertainty

Thank you.

Proof

- $c_t(i,j)$: number of plays of (i,j) on periods 1,2,…,t-1
- $\Delta_t(i,j) = c_t(i,j) - c_t(j,i)$

$$\Phi_t = \frac{1}{2}\sum_{i,j}\left(\Delta_t(i,j)\right)^2$$

$$\leq \Phi_{t-1} + 2\cdot\Delta_t(i_t, j_t) + 1$$

(the difference between Δ_t and Δ_{t-1} is only for one (i,j) pair)

$$E\left[\Phi_t\right] \leq E\left[\Phi_{t-1}\right] + 1 \leq t$$

($E[\Delta_t]=0$)

$$\left|\text{copycat payoff}\right| \leq \frac{1}{2}\sum_{i,j}\left|\Delta_T(i,j)\right|$$

(assuming $\max_{i,j}|a_{i,j}|\leq 1$)

$$\left|\text{copycat payoff}\right|^2 \leq \frac{n^2}{4}\sum_{i,j}\left(\Delta_T(i,j)\right)^2 = \frac{n^2}{2}\Phi_T \quad \text{(Cauchy - Schwartz)}$$

$E[|\text{copycat payoff}|]^2 \leq E[|\text{copycat payoff}|^2] \leq (n^2T)/2$

The expected average payoff of copycat (over T periods) $\leq \frac{n}{\sqrt{2T}}$