Learning and Efficiency of Outcomes in Games

Éva Tardos

Cornell

Large population games: traffic routing

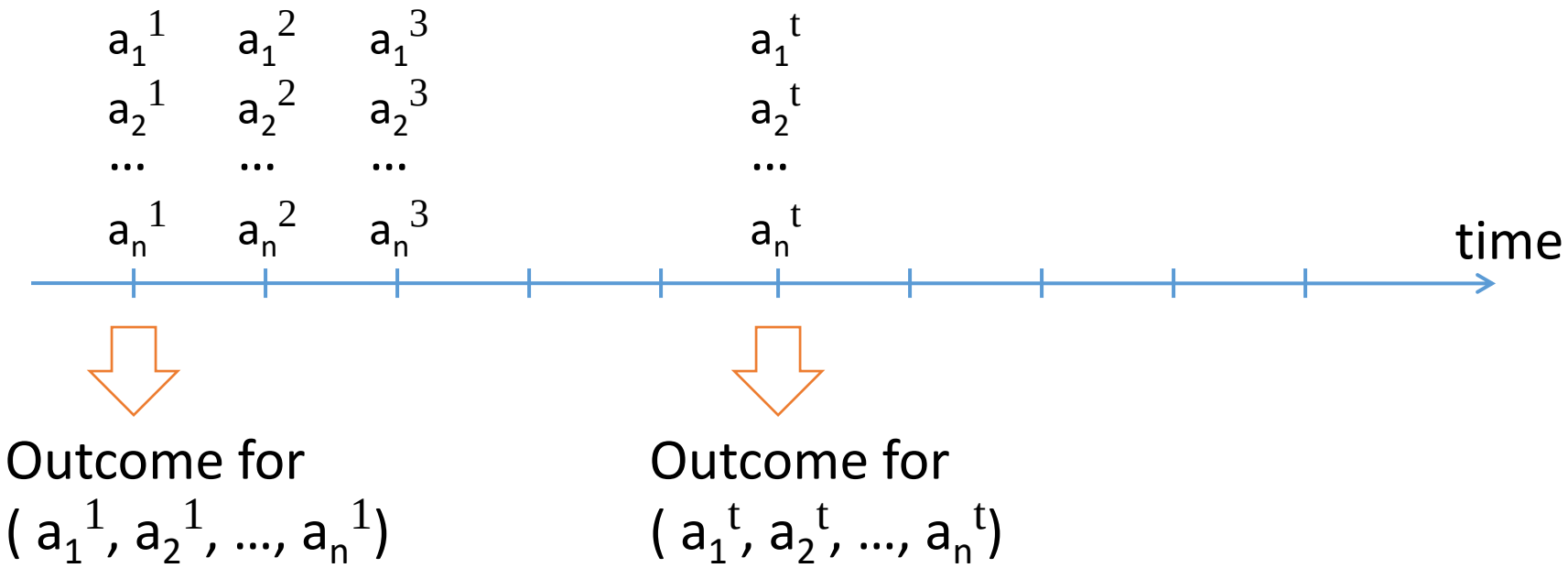
- Traffic subject to congestion delays
- cars and packets follow shortest path
- Congestion game =cost (delay) depends only on congestion on edges

Traffic streams change

e.g., popular sites may change

Changes in system setup

Repeated games



- Player's value/cost additive over periods, while playing
- Players try to learn what is best from past data

What can we say about the outcome? Part I

How fast do they learn well enough: Part II (how long do they have to stay)

Nash as Selfish Outcome ?

Can the players find Nash?

Which Nash?

They need too much information!

Correct belief about behavior of all market participants!

takes time …

Daskalakis-Goldberg-Papadimitrou'06

Nash exists, but ….

Finding Nash is

- PPAD hard in many games

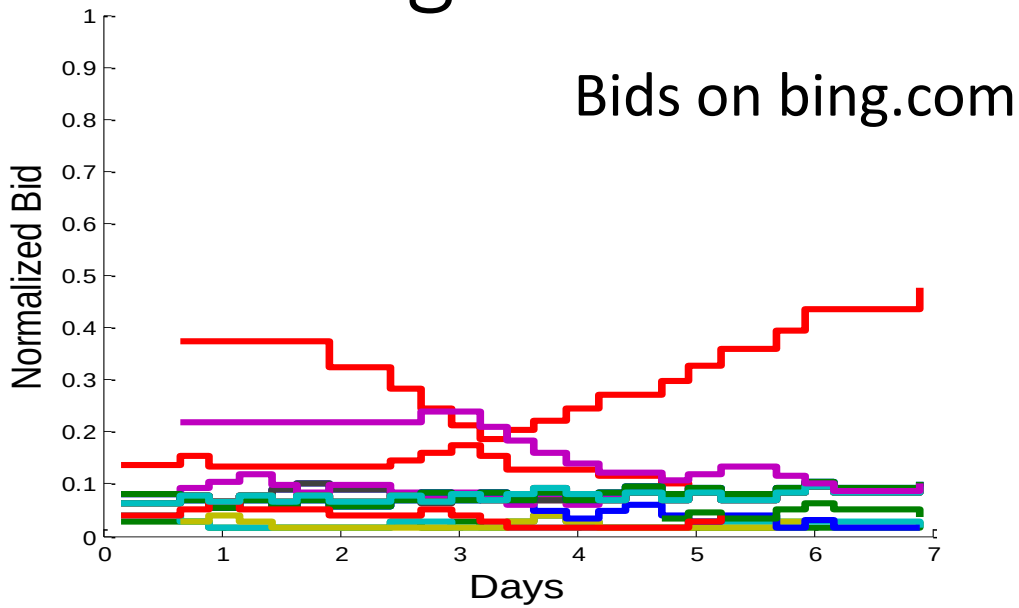
Nash equilibrium of the one-shot game?

a_1 a_1 a_1 a_1 a_1 a_1 a_1 a_1 a_1

a_2 a_2 a_2 a_2 a_2 a_2 a_2 a_2 a_2

… … … … … … … … …

a_n a_n a_n a_n a_n a_n a_n a_n a_n time



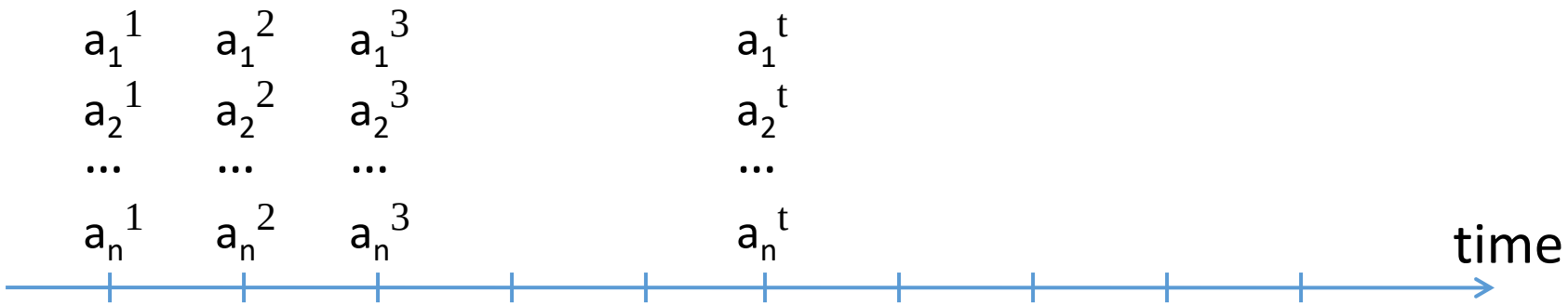
Nash equilibrium of the "one-shot" game: Stable actions a with no regret for any alternate strategy x:

$$cost_i(x, a_{-i}) \geq cost_i(a)$$

No regret

But players are not this steady

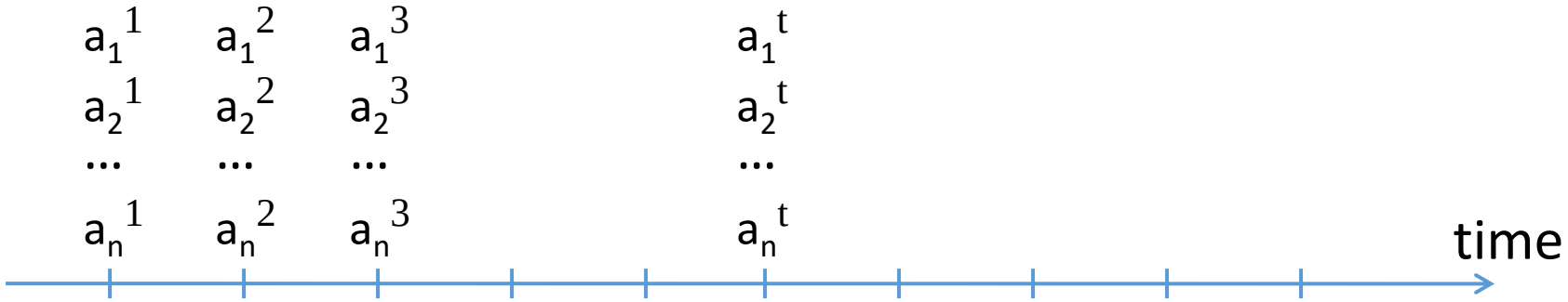
Learning outcome



Maybe here they don't know how to play, who are the other players, …

By here they have a better idea…

No-regret without stability: learning



No regret: for any fixed action x:(cost $\in [0,1]$):

$$\sum_t cost_i(a^t) \leq \sum_t cost_i(x, a_{-i}^t) + error$$

regret

$$\text{error} \leq \sqrt{T}$$ (if $o(T)$ called no-regret)

Today: approximate no-regret

For any fixed action x (with d options) :

$$\sum_t cost_i(a^t) \leq \sum_t cost_i(x, a^t_{-i}) + \sqrt{T log\ d}$$

In fact, much better bound applies!

Foster, Li, Lykouris, Sridharan, T NIPS'16

$$\sum_t cost_i(a^t) \leq (1+\epsilon) \sum_t cost_i(x, a^t_{-i}) + \frac{\log d}{\epsilon}$$

Same algorithms! MWU (Hedge), Regret Matching, etc.

T=time, d=# strategies

Quality of Learning and the Price Anarchy

Price of Anarchy [Koutsoupias-Papadimitriou'99]

$$PoA = \max_{a\ Nash} \frac{cost(a)}{Opt}$$

Assuming **no-regret learners** in stable game: [Blum, Hajiaghayi, Ligett, Roth'08, Roughgarden'09]

$$PoA = \lim_{T\to\infty} \frac{\sum_{t=1}^{T} cost(a^t)}{T\ Opt}$$

[Lykouris, Syrgkanis, T. 2016] dynamic population

$$PoA = \lim_{T\to\infty} \frac{\sum_{t=1}^{T} cost(a^t, v^t)}{\sum_{t=1}^{T} Opt(v^t)}$$

where v^t is the vector of player types at time t

Smooth games [Roughgarden'09]

Game is (λ,μ)-smooth (λ > 0; μ< 1):

if for all strategy vectors a and an optimal solution a_i^*

$$\sum_i cost_i(a) \leq \sum_i cost_i(a_i^*, a_{-i}) \leq \lambda\, OPT + \mu\, cost(a)$$

A Nash equilibrium a has cost(a) $\leq \frac{\lambda}{1-\mu}$Opt

Most price of anarchy bounds via smoothness proofs: congestion games, simple auctions

Examples of "smoothness bounds"

- Monotone increasing congestion costs (1,1) smooth

  $\Rightarrow$ Nash cost ≤ opt of double traffic rate (Roughgarden-T'02)

- affine congestion cost are (1, ¼) smooth (Roughgarden-T'02)

  $\Rightarrow$ 4/3 price of anarchy

- Atomic game (players with >0 traffic) with linear delay (5/3,1/3)-smooth (Awerbuch-Azar-Epstein & Christodoulou-Koutsoupias'05)

  $\Rightarrow$ 2.5 price of anarchy

Resulting bounds are tight

Smooth games and learning

no-regret learning results in sequence a^t :

player i would do action a_i^* in optimum

$\frac{1}{T}\sum_t cost_i(a^t) \le \frac{1}{T}\sum_t cost_i\left(a_i^*, a_{-i}^t\right) + \frac{1}{T}R$ (doesn't need to know a_i^*)

A cost minimization game is (λ,μ)-smooth (λ > 0; μ< 1):

$$\frac{1}{T}\sum_t cost(a^t) \le \frac{\lambda}{1-\mu}\mathrm{Opt} + \frac{\mathrm{n}}{(1-\mu)T}R$$

recall: $R = O(\sqrt{T}\log d)$

$\Rightarrow$ Additive error $\mathrm{O}(\frac{\mathrm{n}}{\sqrt{T}} \cdot log\ n)$

Speed of Convergence ?

Use approx no-regret learning:

$$\sum_t cost_i(a^t) \le (1+\epsilon)\sum_t cost_i\left(a_i^*, a_{-i}^t\right) + AR$$

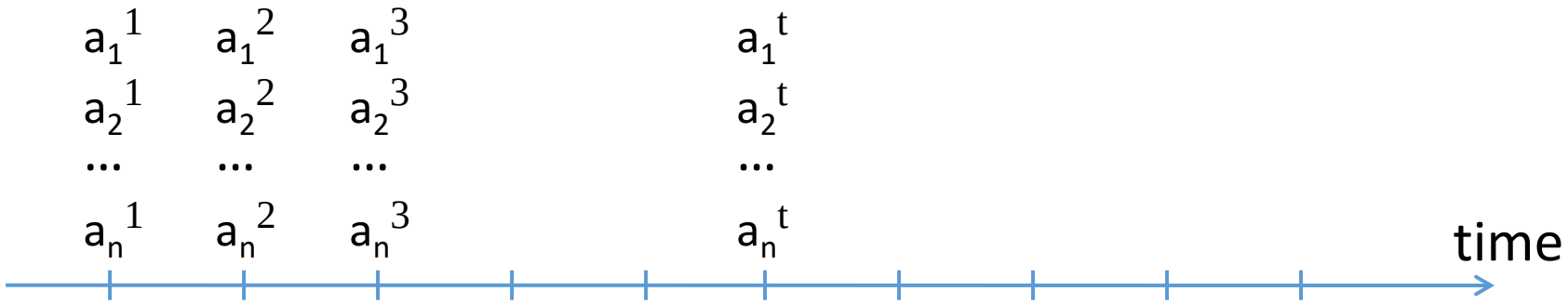
A cost minimization game is (λ,μ)-smooth (λ > 0; μ< 1):

A approx. no-regret sequence a^t has

$$\frac{1}{T}\sum_t cost(a^t) \le \frac{(1+\epsilon)\lambda}{1-(1+\epsilon)\mu}\,\mathrm{Opt} + \frac{\mathrm{n}}{\mathrm{T}(1-(1+\epsilon)\mu)}\,\mathrm{AR}$$

$AR = \frac{\log d}{\epsilon}$, so error $\frac{\mathrm{n}}{\mathrm{T}} \cdot \frac{log\ d}{\epsilon(1-(1+\epsilon)\mu)}$

No-regret: how good is this as a model of learning?



For any fixed action x (with d options) :

$$\sum_t cost_i(a^t) \leq (1+\epsilon)\sum_t cost_i(x, a^t_{-i}) + \frac{\log d}{\epsilon}$$

Pro:

- No need for common prior or rationality assumption on opponents (takes advantage if opponents play badly!)
- Behavioral assumption: if there is a consistently good strategy: please notice!
- Algorithms: Many simple rules ensure approx. regret $\log d/\epsilon$, and regret ~$\sqrt{Tlog\ d}$ for all x , Hedge, Regret Matching, Follow the perturbed leader

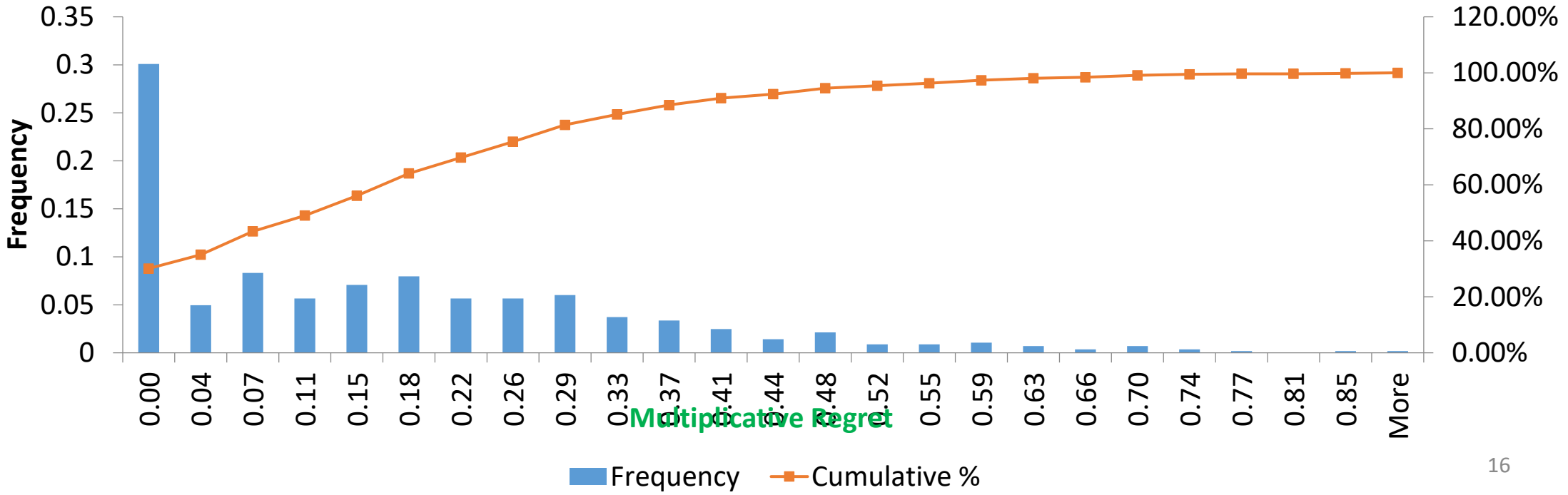
  Idea: choose at random: outcome good increase prob. Outcome bad decrease prob.

Do players really learn?
Nekipelov, Syrgkanis, T, EC'15

- Data from Microsoft Ad-Aution: 9 frequent bid changing advertisers

Value of advertiser?

Half the advertisers have <10% regret, 30% have <0 regret!



Slowly changing game:
[Lykouris, Syrgkanis, T. ‘16]

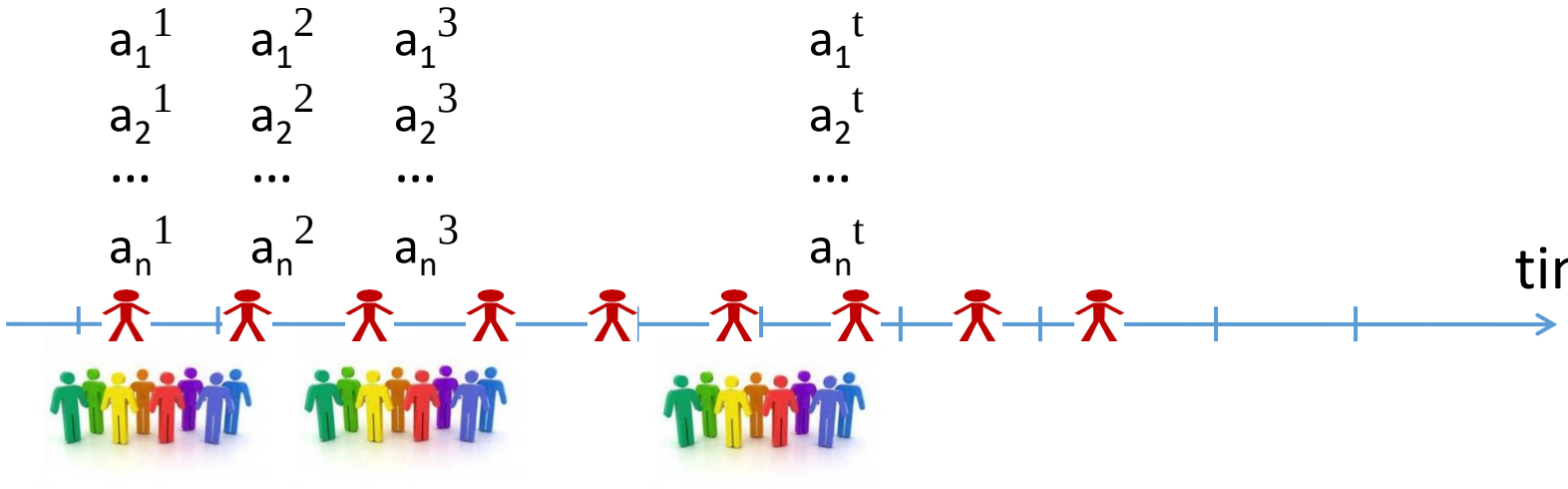
Dynamic population model:

At each step t each player i

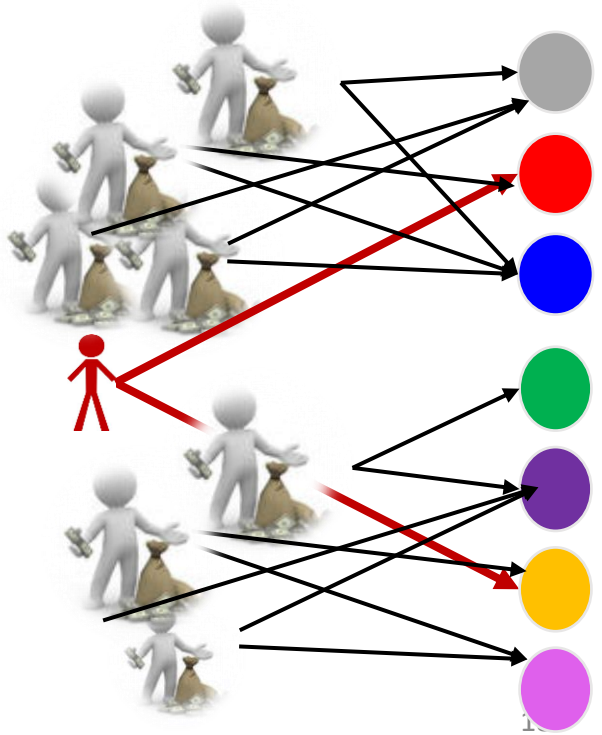
is replaced with an arbitrary new player with probability p

In a population of N players, each step, Np players replaced in expectation

Need for adaptive learning



items



Example: (matching)

- Strategy = choose an item
- Best "fixed" strategy in hindsight very weak in changing environment
- Learners can adapt to the changing environment

Result (Lykouris, Syrgkanis, T'16) :

In many games we bound average welfare close to Price of Anarchy

even when the rate of change is high, $\boldsymbol{p \approx \frac{1}{\log n}}$ with n players

assuming **adaptive** no-regret learners

- Worst case change of player type $\Rightarrow$ need for adapting to changing environment
- Sudden large change is unlikely

Summary so far:

Player populations using no-regret learning do well even in dynamic environments

- Learning guarantees high social welfare in smooth games in the limit
- Stable approx. solution + good PoA bound $\Rightarrow$ good efficiency with dynamic population

Low regret error guarantees fast convergence to high social welfare, and allows for higher population turnover

We need broad classes of learning algorithms, to make learning a good behavioral assumption!

Low approximate regret without full info?

Do we need full information feedback to achieve:

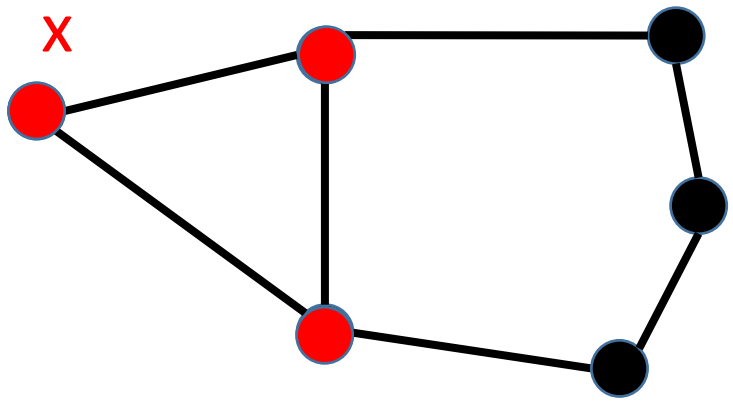
$$\sum_t cost_i(a^t) \leq (1+\epsilon)\sum_t cost_i\left(a_i^*, a_{-i}^t\right) + AR$$

Today: learning to for low approximate regret with partial feedback

joint work with Thodoris Lykouris and Karthik Sridharan

Partial information feedback: bandits

- Focus on one player: $c^t(x) = cost_i\left(x, a^t_{-i}\right)$
- Classical bandit model: $0 \leq c^t(x) \leq 1$, get $c^t(x)$ only on x selected
- Partial feedback (Mannor and Shamir NIPS'11, Alon et al NIPS'13): graph feedback:



each node: an option x

choosing x,

see feedback for $N(x)$

Partial feedback: importance sampling:
reduction to full information

- Use a full information algorithm with modified cost:

  cost $\tilde{c}^t(a) = 0$, if a not selected

  $= c^t(a)/p_a^t$ if a is selected, and had probability p_a^t

- Any fixed a expected cost $E(\sum_t \tilde{c}^t(a)) = \sum_t p_a^t \tilde{c}^t(a) = \sum_t c^t(a)$

- Imagine full information algorithm on $\widetilde{c_i}^t(a)$

  Expected cost is $E(\sum_a p_a^t \tilde{c}^t(a)) = \sum_a p_a^t E(\tilde{c}^t(a)) = \sum_a p_a^t c^t(a)$

$\Rightarrow$ Full information bound on costs $\tilde{c}^t(a)$ applies to bandit feedback…

Reduction to full information, part 2

Recall: cost $\tilde{c}^t(a) = 0$, if a not selected

$= c^t(a)/p_a^t$ if a is selected, and had probability p_a^t

Trouble: $\max \tilde{c}^t(a) = \max 1/p_a^t = \infty$

approx. regret was: $\frac{\ln d}{\epsilon} \max_{a,t} c_a^t$

- Classical solution:

  add a bit of random noise: keep $p_a^t \geq \gamma$ for some parameter γ

  regret bound of $T\gamma$, not low approximate regret

- Proposed solution:

  freeze arm if $p_a^t < \gamma$. Do not select, and do not update!

Bandits with freezing

- Use probability $\tilde{p}_a^t = 0$ if $p_a^t < \gamma$

  $\tilde{p}_a^t \sim p_a^t$ if $p_a^t \geq \gamma$

Need to make sure that $\sum_{a:p_a^t<\gamma} p_a^t \leq \epsilon$, e.g., use $\gamma \leq \frac{\epsilon}{d}$

- Any fixed a expected cost $E(\sum_t \tilde{c}^t(a)) = \sum_t \tilde{p}_a^t \tilde{c}^t(a) \leq \sum_t c^t(a)$

  estimator is negatively biased!

- Full info expected cost is $E(\sum_a p_a^t \tilde{c}^t(a)) = \sum_a p_a^t \, c_a^t \approx \sum_a \tilde{p}_a^t \, c_a$

$\Rightarrow$ full information bound applies with costs $\leq \frac{1}{\gamma}$

Benefit with freezing as a learning method

- Small and natural change for any full information algorithm.
- Black box reduction to bandit feedback
- Results in small approximate regret, and “small loss bound” on regret

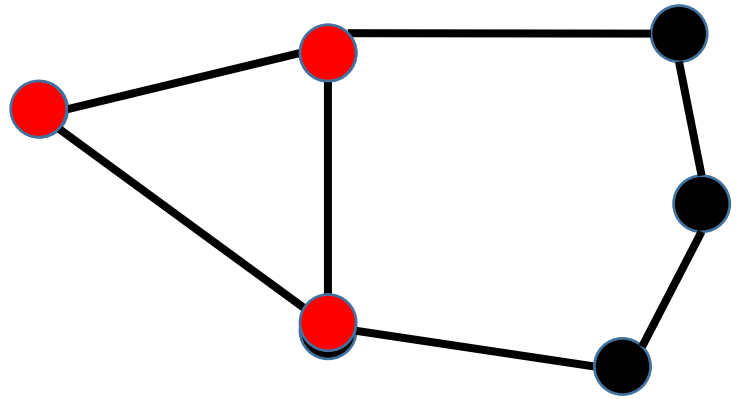
Extends to shifting comparator: probability is never too small to recover!

Regret of adaptive learning is bounded by $k \log d \, / \epsilon$

with respect to any sequence that changes k times

Freezing and partial feedback

- Partial feedback (Mannor and Shamir NIPS'11, Alon et al NIPS'13): graph feedback:



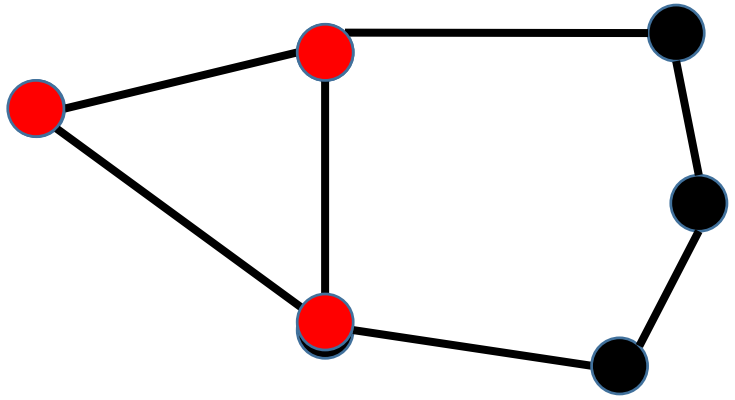
each node: an option x

choosing x, see feedback for $N(x)$

- Graph complete = full information: we have $O(\frac{\log d}{\epsilon})$
- Graph empty = bandit: we have $O(\frac{d}{\epsilon})$

Learning with Partial feedback

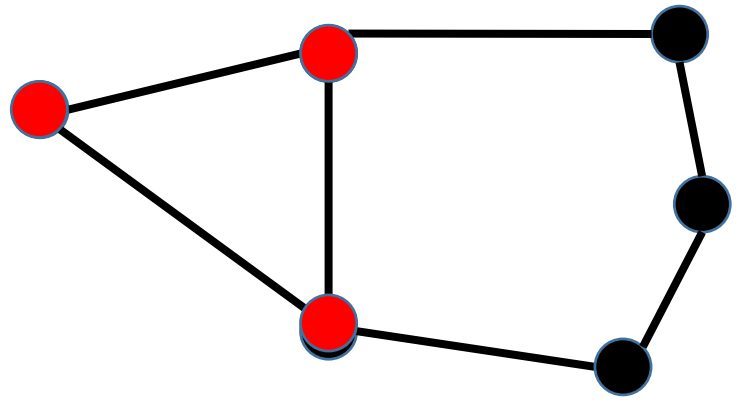
Theorem (Lykouris-Sridharan-T'17): using freezing we can approximate regret bound of $O(\alpha(G)\log d\ /\epsilon^2)$, where $\alpha(G)$ is the max independent set in G



Learning with Partial feedback

Idea: p_a^t probability arm a is played

$\pi_a^t = \sum_{b \in N(a)} p_b^t$ probability arm a is seen



Importance sampling: update cost of all arms seen

cost $\tilde{c}^t(a) = 0$, if a not seen

$= c^t(a)/\pi_a^t$ if a is seen

Freeze arms with π_a^t too small

Importance sampling with partial feedback

freeze arms with $\pi_a^t < \gamma$

- Any fixed a expected cost
  $E(\sum_t \tilde{c}^t(a)) = \sum_t (\sum_{b \in N(a)} \tilde{p}_b^t)\, \tilde{c}^t(a) \leq \sum_t c^t(a)$
  estimator is negatively biased due to freezing

- Full info Expected cost is $E(\sum_a p_a^t \tilde{c}^t(a)) = \sum_a p_a^t\, c_a \approx \sum_a \tilde{p}_a^t\, c_a$

$\Rightarrow$ full information bound applies with costs $\leq \frac{1}{\gamma}$

Question: how large can we keep γ?

Freezing arms with prob π_a^t too small

- Use $\gamma = \frac{\epsilon}{\alpha(G)}$

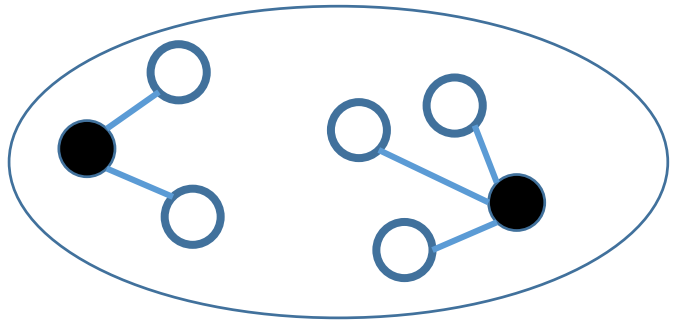
Step 0. Freeze all arms with $\pi_a^t \leq \gamma$

Lemma: total probability frozen in step 0 at most $\alpha(G)\gamma = \epsilon$

Proof: consider set of frozen arms A

Max independent set $I \subset A$

All node in A seen by a node in I



Freezing arms with prob. π_a^t too small (cont.)

- Update $\tilde{\pi}_a^t = \sum_{b \in N(a)} \tilde{p}_b^t$
- This now makes other arms seen with small probability…

Steps 1, 2, 3,… Recursively freeze all arms with $\tilde{\pi}_a^t \leq \gamma/3$

Lemma: total probability frozen in steps 1,2,… at most 3ϵ

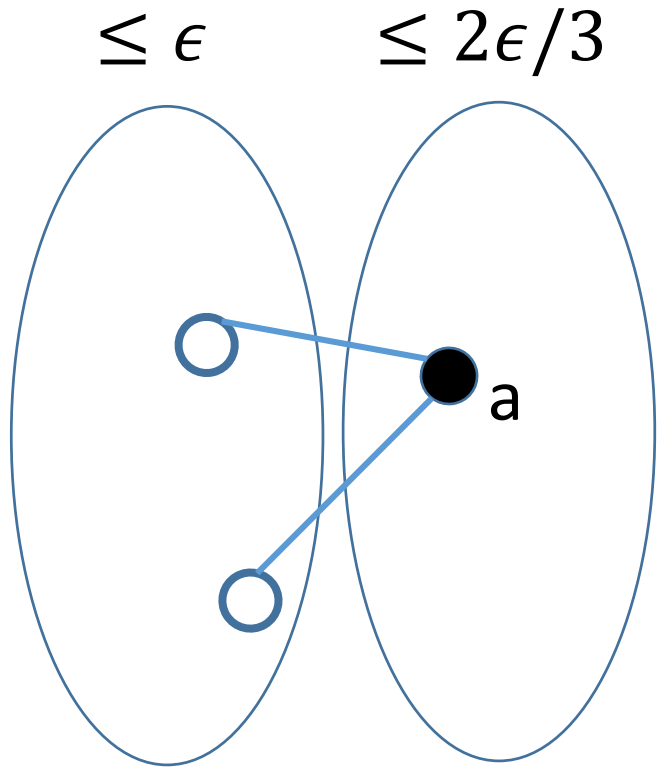
Proof:

Node a frozen in step 1: $\pi_a^t > \gamma$ and $\tilde{\pi}_a^t \leq \gamma/3$

$\Rightarrow$ Seen by $> \frac{2}{3}\gamma$ probability in step 0,

Step 0 has total probability only ϵ and each node only seen with probability $\leq \gamma$

$\Rightarrow$ total at most $O(\epsilon)$



Conclusions

Learning in games:

- Good way to adapt to opponents
- No need for common prior
- Takes advantage of opponent playing badly.
- Simple strategies guarantee low approximate regret in full information as well as with partial information

Learning players do well even in dynamic environments

- Stable approx. solution + good PoA bound $\Rightarrow$ good efficiency with dynamic population