Learning as a Solution Concept (in repeated games)

Éva Tardos

Cornell, Computer Science

Talk outline

1. Example games, and questions we want to ask:
   - What do we mean by learning?
   - What can we say about outcome of learning?
2. No-regret learning as a behavioral assumption: pros and cons
3. Quality of learning outcomes: price of anarchy
4. Limitation of no-regret as a solution concept
   - Can be hard to achieve small regret: what may be possible?
   - No-regret may lead us in the wrong direction
5. Extension on price of anarchy results via improved learning

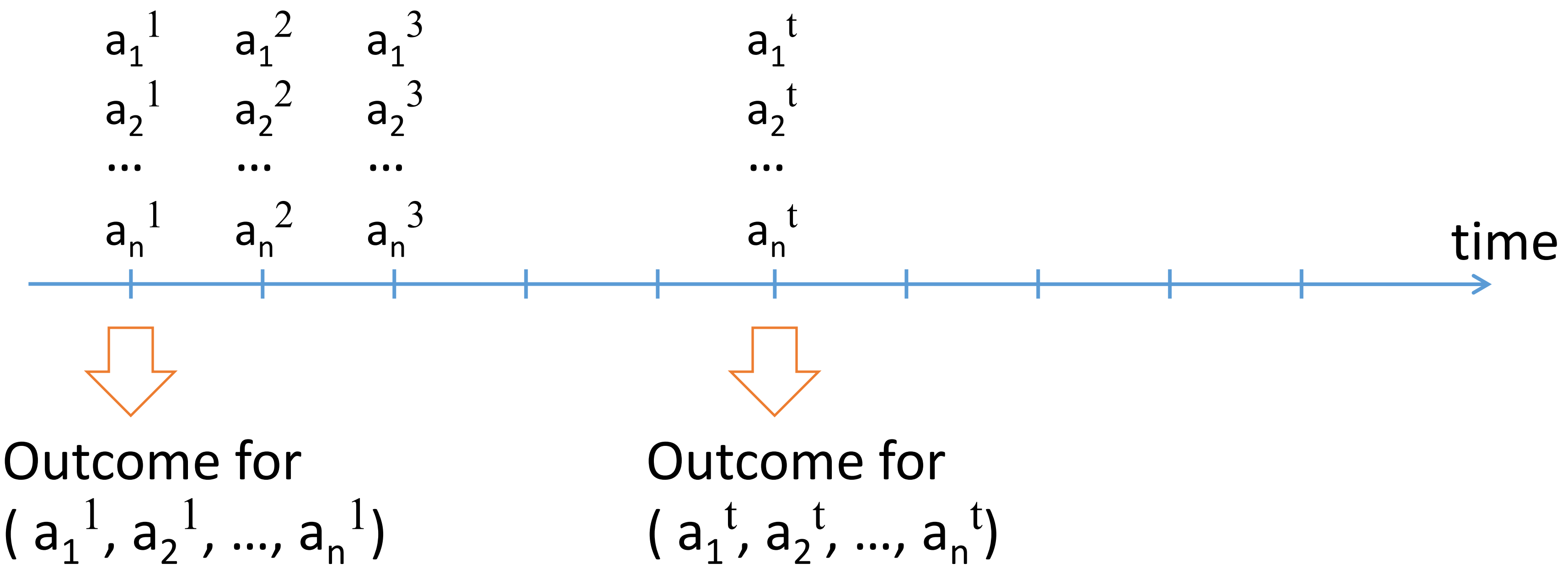
Example 1: traffic routing

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

Example 2: advertising auctions

advertising auctions

Repeated games



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

What can we say about the outcome?

What do we mean by “learning from data”?

High Social Welfare: Price of Anarchy in Routing

Theorem (Roughgarden-T'02):

In any network with continuous, non-decreasing cost functions and very small users

$\boxed{\text{cost of Nash with rates } r_i \text{ for all } i} \leq \boxed{\text{cost of opt with rates } 2r_i \text{ for all } i}$

Nash equilibrium: stable solution where no player had incentive to deviate.

Price of Anarchy= $\dfrac{\text{cost of worst Nash equilibrium}}{\text{“socially optimum” cost}}$

Games and Solution Quality

Tragedy of the Commons

- Rational selfish action can lead to outcome bad for everyone

Model:

- Value for each cow decreasing function of # of cows
- Too many cows: no value left

More examples of price of anarchy bounds

- Monotone increasing congestion costs

  Nash cost ≤ opt of double traffic rate (Roughgarden-T’02)

- affine congestion cost (Roughgarden-T’02) 4/3 price of anarchy

- Atomic game (players with >0 traffic) with linear delay (Awerbuch-Azar-Epstein & Christodoulou-Koutsoupias’05) 2.5 price of anarchy

- Bandwidth sharing (Johari-Tsitsiklis’04) 4/3 price of anarchy

Price of anarchy in auctions

- First price is auction Hassidim, Kaplan, Mansour, Nisan EC’11)

  Price of anarchy 1.58…
- All pay auction price of anarchy 2
- First position auction (GFP) is price of anarchy 2
- Variants with second price (see also Christodoulou, Kovacs, Schapira ICALP’08) price of anarchy 2

Other applications include:

- public goods
- Fair sharing (Kelly, Johari-Tsitsiklis) price of anarchy 1.33
- Walrasian Mechanism (Babaioff, Lucier, Nisan, and Paes Leme EC’13)

Learning in Repeated Game

- What is learning?
- Does learning lead to finding Nash equilibrium?

Brown’51, Robinson’51:

- fictitious play = best respond to past history of other players

Goal: “pre-play” as a way to learn to play Nash.

Outcome of Fictitious Play in Repeated Game

- Does learning lead to finding Nash equilibrium?

  mostly not

Theorem: Marginal distribution of each player actions converges to Nash in

Robinson’51: In two player 0-sum games

Miyasawa’61: In generic payoff 2 by 2 games

Finding Nash of the one-shot game?

a_1^1	a_1^2	a_1^3	a_1^t	a_1	a_1	a_1	a_1	a_1	a_1	a_1	a_1	a_1	
a_2^1	a_2^2	a_2^3	a_2^t	a_2	a_2	a_2	a_2	a_2	a_2	a_2	a_2	a_2	
…	…	…	…	…	…	…	…	…	…	…	…	…	
a_n^1	a_n^2	a_n^3	a_n^t	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:

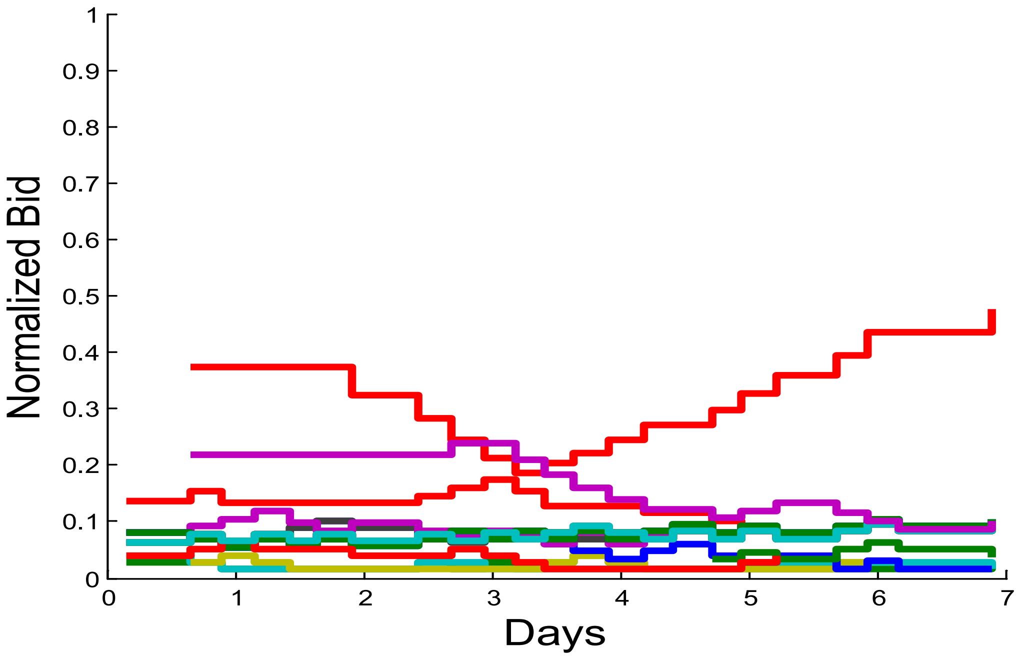
- Stable actions a
- with no regret for any alternate strategy x:

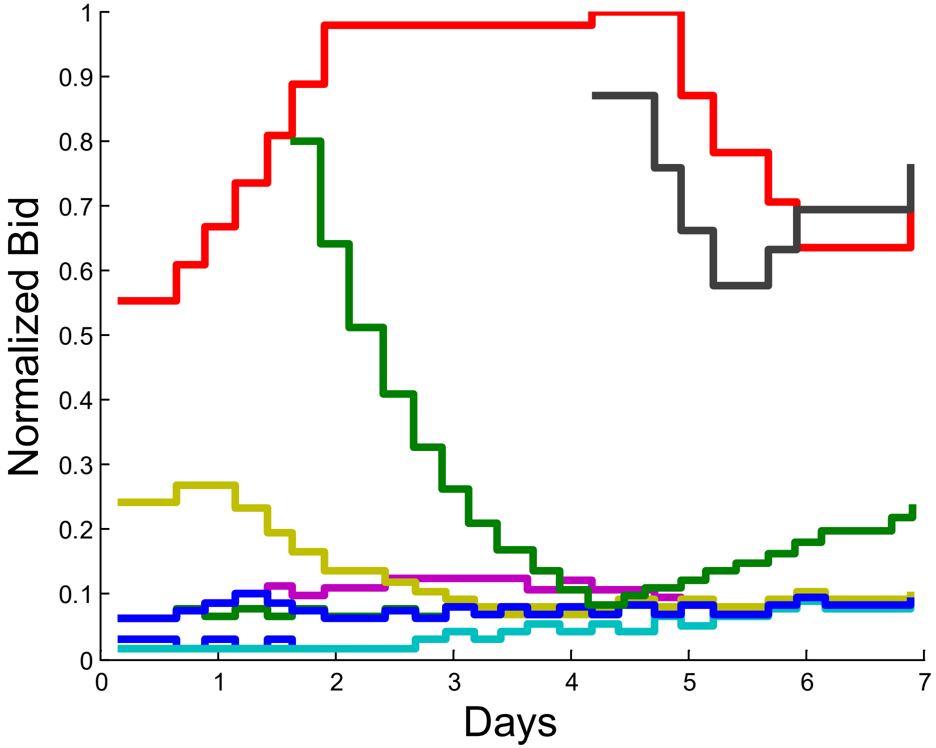
$$cost_i(x, a_{-i}) \geq cost_i(a)$$

No regret

Behavior is far from stable

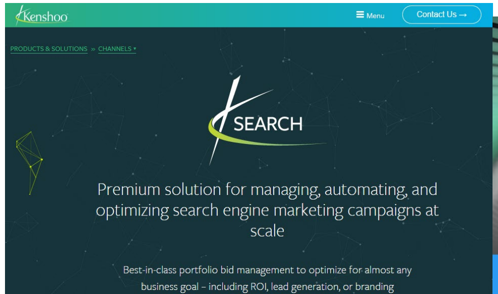
data from Nekipelov, Syrgkanis, T.'15

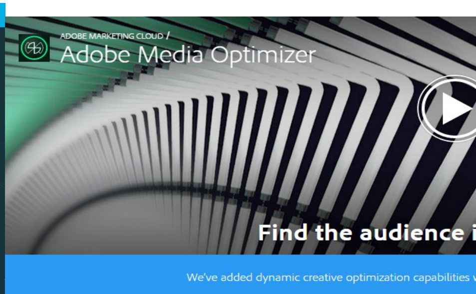




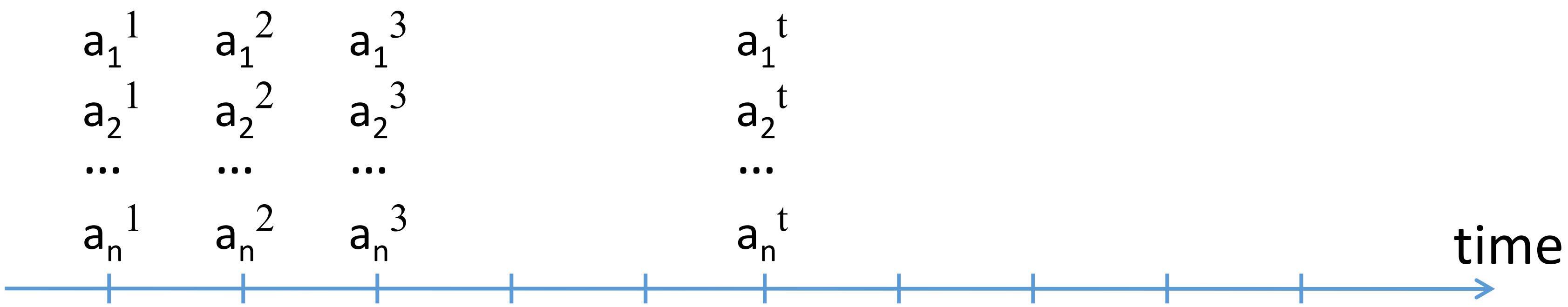
Bing search advertisement bid

Bidders use sophisticated bidding tools





Change of focus: Outcome of learning while playing



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

By here they have a better idea…

Recall: No regret at Nash:

$a_1 \quad a_1 \quad a_1 \quad a_1 \quad a_1 \quad a_1 \quad a_1 \quad a_1 \quad a_1$

$a_2 \quad a_2 \quad a_2 \quad a_2 \quad a_2 \quad a_2 \quad a_2 \quad a_2 \quad a_2$

$\dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots$

$a_n \quad a_n \quad a_n \quad a_n \quad a_n \quad a_n \quad a_n \quad a_n \quad a_n$ time

- Stable actions a
- with no regret for any alternate strategy x:

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

No regret

No-regret without stability: learning

a_1^1 a_1^2 a_1^3 a_1^t

a_2^1 a_2^2 a_2^3 a_2^t

… … … …

a_n^1 a_n^2 a_n^3 a_n^t

time

No regret: for any fixed action x:

regret

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

$\text{error} \le \sqrt{T}$ assuming cost ∈ [0,1]

(if $o(T)$ called no-regret)

Many classical online learning algorithms

Hannan consistency [Hannan'57]

Multiplicative weights (Hedge) [Freund-Schapire'97]

Follow the perturbed leader [Kalai-Vempala'03]

Alternate: approximate no-regret

For any fixed action x (with d options) :

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

T=time, d=# strategies

In fact, much better bound applies:

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

Same algorithms!

Multiplicative weights (Hedge) [Freund-Schapire'97]

Follow the perturbed leader [Kalai-Vempala'03]

Outcome of no-regret learning = (Coarse) correlated equilibrium

Coarse correlated equilibrium: probability distribution of outcomes such that for all players

expected payoff ≥ exp. payoff of any fixed strategy

Coarse correlated eq. & players independent = Nash

Theorem [Freund and Schapire'99, Robinson'51] In two-person 0-sum games play converges to Nash value, and Nash strategy for all players

but play is correlated

	R	P	S
R	0 / 0	1 / -1	-1 / 1
P	-1 / 1	0 / 0	1 / -1
S	1 / -1	-1 / 1	0 / 0

Outcome of no-regret learning in a fixed game

Limit distribution σ of play (action vectors a=$(a_1, a_2, \dots, a_n)$)

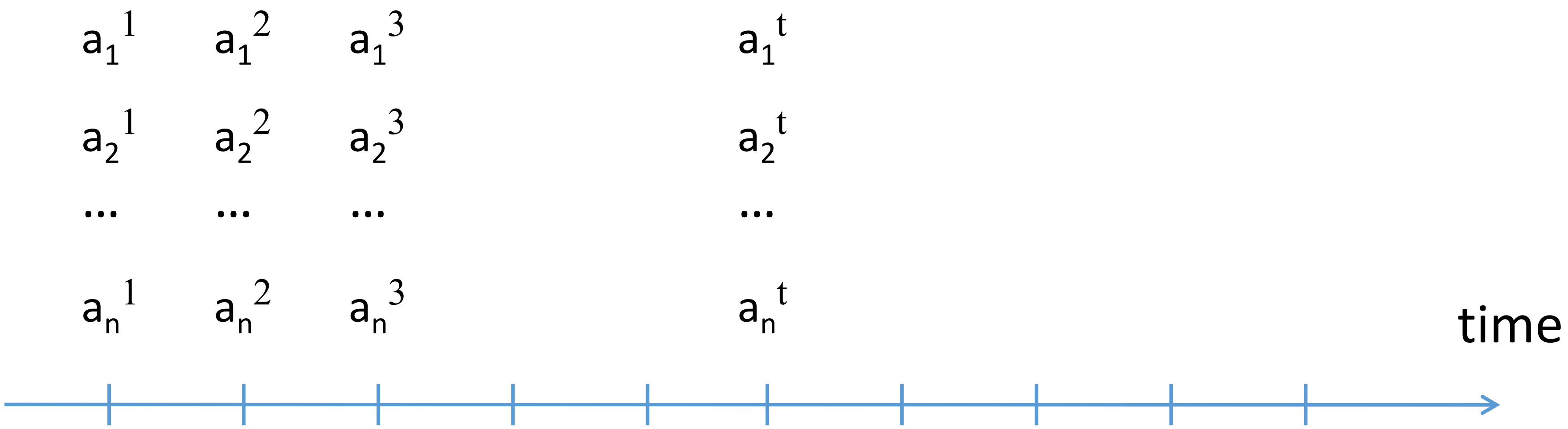
- all players i have no regret for all strategies x

$$E_{a\sim\sigma}\big(cost_i(a)\big) \leq E_{a\sim\sigma}(cost_i(x, a_{-i}))$$

Hart & Mas-Colell: Long term average play is (coarse) correlated equilibrium

Players update independently, but correlate on shared history

No-regret as a model of learning?



For any fixed action x (with d options) :

$$\sum_t cost_i(a^t) \leq (1+\epsilon)\sum_t cost_i(x, a_{-i}^t) + \epsilon T$$ T=time horizon

Behavioral model, first suggested Blum, Hajiaghayi, Ligett, Roth'08 in the context of traffic routing and Christodoulou, Kovacs, Schapira '08 in context of auctions (as opposed to analyzing outcomes of algorithms).

Behavioral assumption: if there is a consistently good strategy: please notice!

No-regret as a model of learning?

Behavioral assumption: if there is a consistently good strategy: please notice!

For any fixed action x (with d options) :

$$\sum_t cost_i(a^t) \leq (1+\epsilon)\sum_t cost_i(x, a_{-i}^t) + \epsilon T$$

T=time horizon

Pros: Behavioral model that can be used in theory!

- Algorithms: Many simple rules ensure small regret
- No need for common prior or rationality assumption on opponents

Cons:

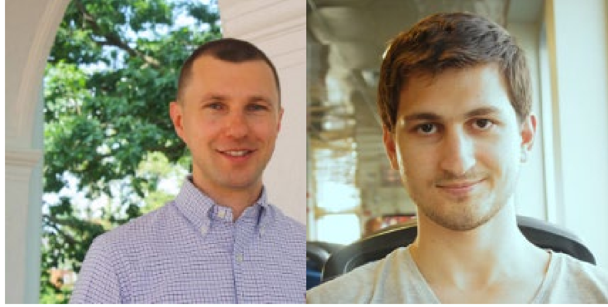
- Can we too hard to do in multi-parameter problems: Yang-Papadimitriou'14, Daskalakis-Syrgkanis'16
- It may not be best response if others use no-regret learning:
- We can except players do to better than no regret: changing environment, policy regret

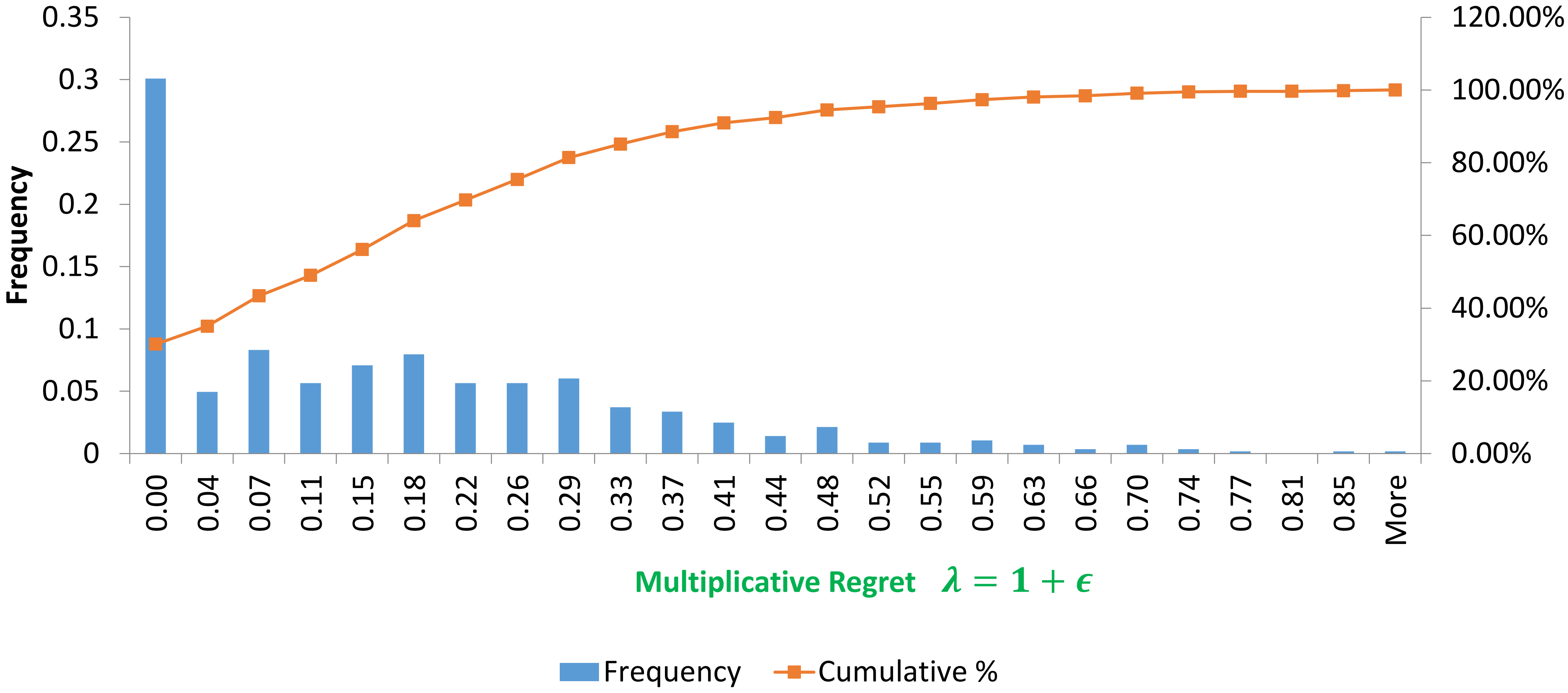
No-regret learning as a behavioral model?

- Er'ev and Roth'96
  lab experiments with 2 person coordination game
- Fudenberg-Peysakhovich EC'14
  lab experiments with seller-buyer game
  recency biased learning
- Nekipelov-Syrgkanis-T. EC'15
  Bidding data on Bing-Ad-Auctions
- Nisan-Noti WWW'17
  Lab experiment with ad-auction games
- Nekipelov-Jalaly-Tardos '18
  Zillow ad-data

Distribution of smallest rationalizable multiplicative regret

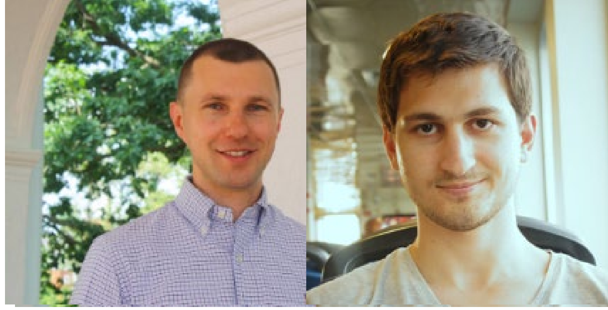
data from Nekipelov, Syrgkanis, T.'15

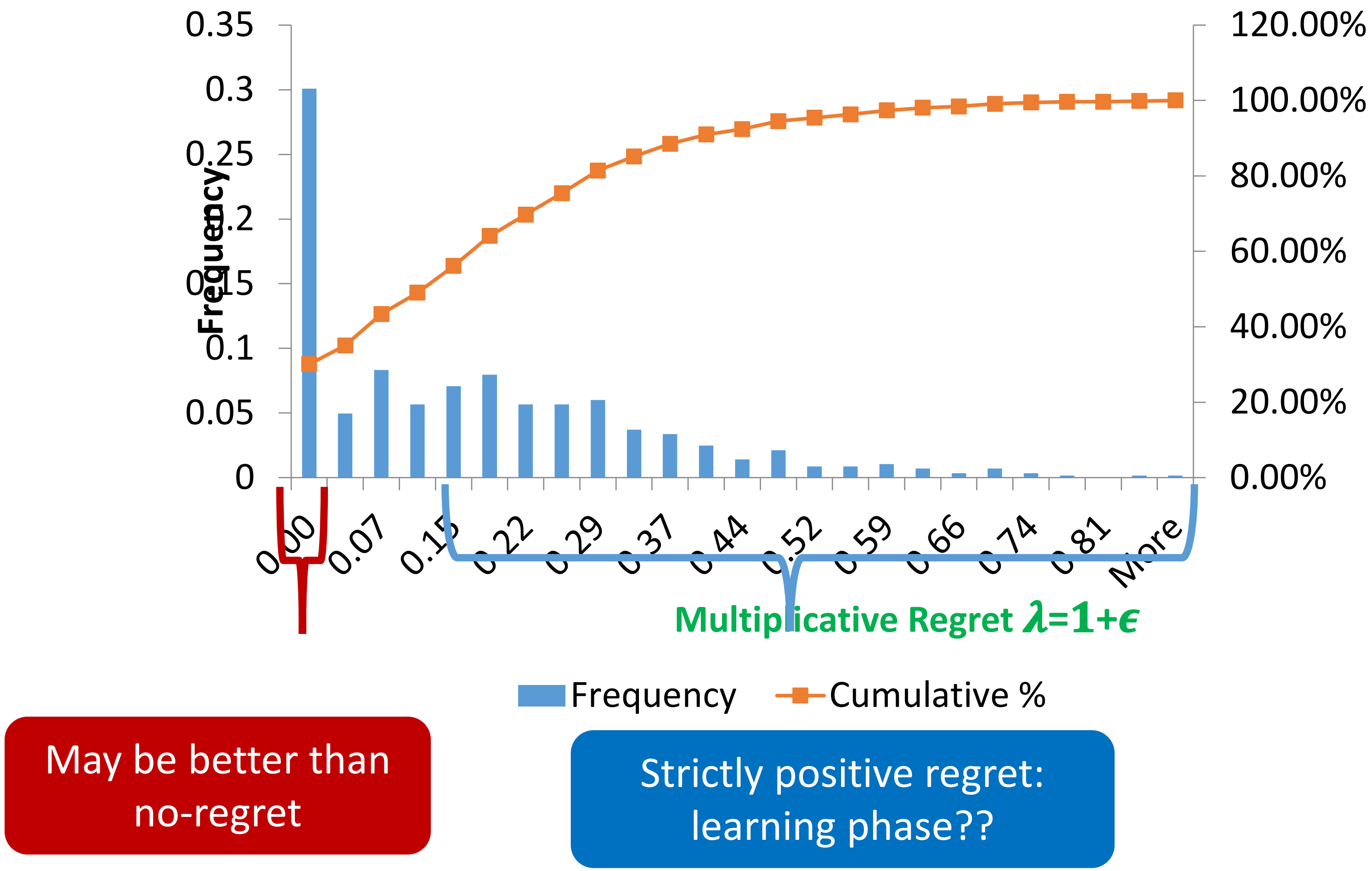




Distribution of smallest rationalizable multiplicative regret

data from Nekipelov, Syrgkanis, T.'15





Nekipelov, Syrgkanis, T'15:

Economerics for learners: using learning (instead of Nash) as an assumption to infer values

Change of focus: Quality of Learning Outcome

Price of Anarchy [Koutsoupias-Papadimitriou'99]

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

Assuming **no-regret learners** in fixed 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

Proof Technique: Smoothness (Roughgarden'09)

Consider optimal solution: player i does action a_i^* in optimum

Nash: $\mathrm{cost_i(a)} \le cost_i\,(\boldsymbol{a_i^*}, a_{-i})$ (doesn't need to know a_i^*)

A game is (λ,μ)-smooth ($\lambda > 0$; $\mu < 1$): if for all strategy vectors a

$$\sum_i cost_i(a) \overset{\text{Nash}}{\le} \sum_i cost_i(a_i^*, a_{-i}) \le \boxed{\lambda\, OPT} + \mu \boxed{cost(a)}$$

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

If Opt <<cost(a), some player will want to deviate to a_i^*

as $\lambda\, OPT + \mu\, cost(a) < cost(a)$

Learning and price of anarchy

Use approx no-regret learning:

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

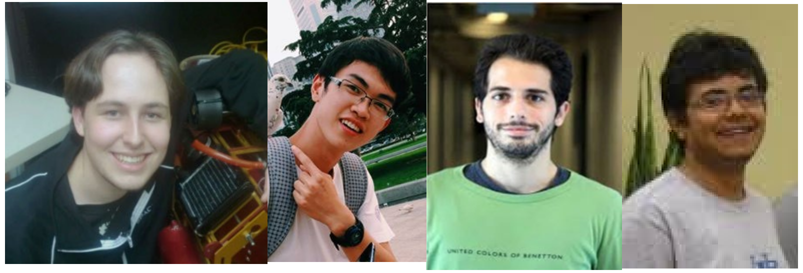
A cost minimization game is (λ,μ)-smooth (λ > 0; μ< 1):

$$\sum_t \sum_i cost_i\left(a_i^*, a_{-i}^t\right) \leq \lambda \sum_t Opt + \mu \sum_t cost(a^t)$$

A approx. no-regret sequence a^t has

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

Speed of Convergence



Special method (e.g., optimistic graduate decent):

2-person 0 sum games: Popov'80

Daskalakis, Deckelbaum, Kim'11, Rakhlin, Sridharan'13

General game:

Syrgkanis, Agarwal, Luo, Schapire'15

General game and **no-regret as a behavioral model**:

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

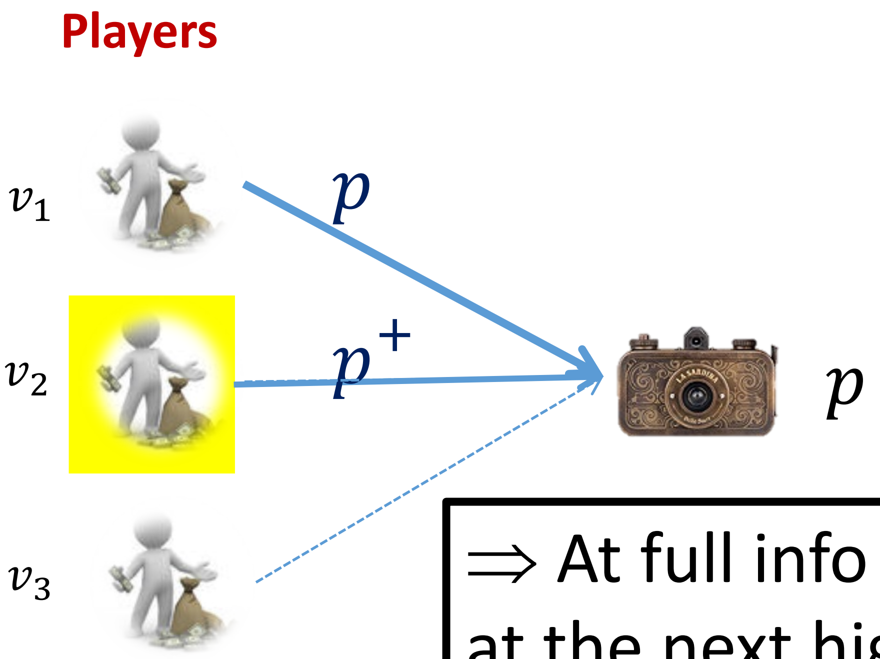
$$\frac{1}{T}\sum_t cost(a^t) \leq \frac{(1+\epsilon)\lambda}{1-(1+\epsilon)\mu}\mathrm{Opt} + \frac{\mathrm{n}}{\mathrm{T}(1-(1+\epsilon)\mu)}\mathrm{AR}$$

Note the convergence speed! $AR = \frac{\log d}{\epsilon}$, so error $\boxed{\frac{\mathrm{n}}{\mathrm{T}} \cdot \frac{log\ d}{\epsilon(1-(1+\epsilon)\mu)}}$

Illustrative example: A utility game: Auction

First Example: Single item first price

- Auction sets a price p (full info, pure Nash).

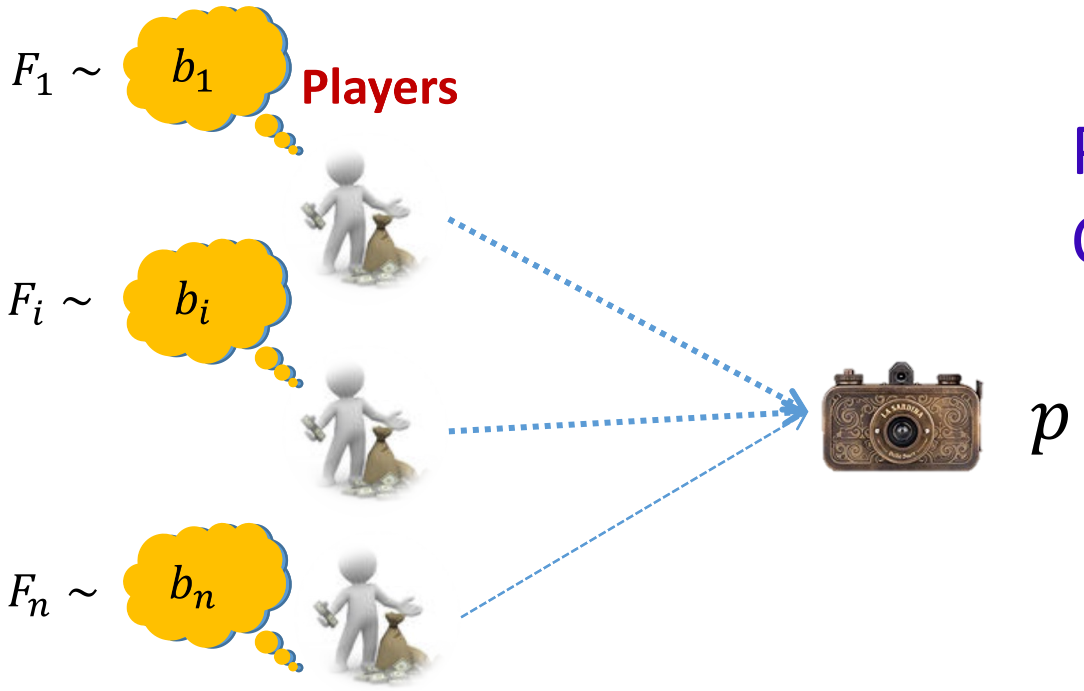


Can win by bidding p^+

$\Rightarrow$ At full info Nash player with max v_i wins at the next highest price

First price auction with uncertainty?

- Bayesian game
- Randomized bid



Price p random

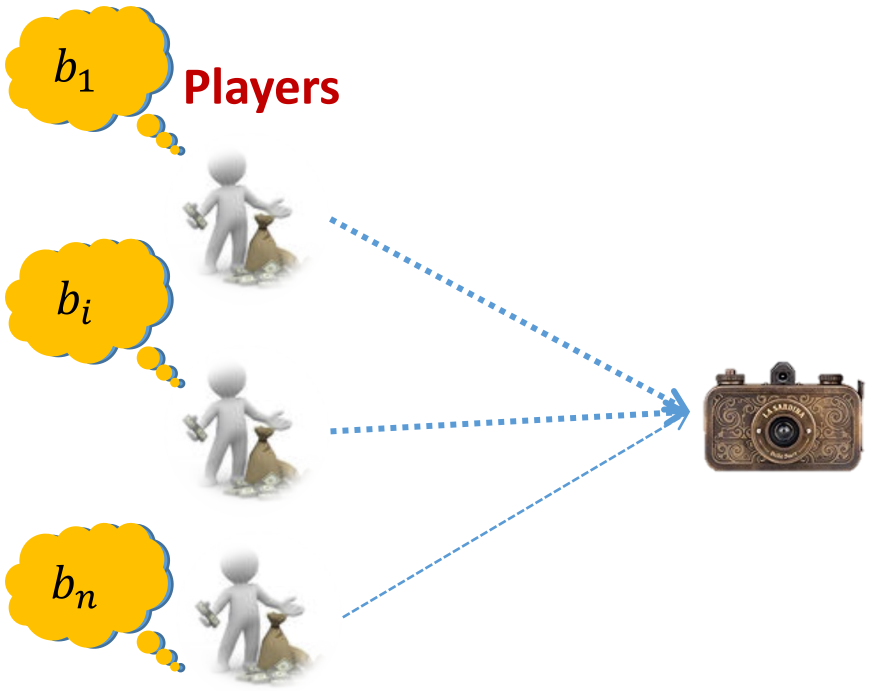
Cannot bid p^+

Bayes Nash analysis

Strategy: bid as a function of value $b_i(v)$

Nash: $E_{v_{-i}b}\left[u_i\big(b(v)\big)|v_i\right] \geq E_{v_{-i}b_{-i}}\left[u_i\big(b_i', b_{-i}(v_{-i})\big)|v_i\right]$

for all b_i'



Auction games:

- Finite set of players 1,…,n
- strategy sets S_i for player i: bid on some items (not a finite set)
- Resulting in strategy vector: s=$(s_1, \dots, s_n)$ for each $s_i \in S_i$
- Utility player i: $u_i(s)$ or $u_i(s_i, s_{-i})$
  - We assume quasi-linear utility, and no externalities:
  - If player wins set if items A_i and pays p_i her value is $u_i(A_i, p_i) = v_i(A_i) - p_i$

- Social welfare? (include auctioneer): $\sum_i v_i(A_i) = \sum_i u_i(A_i) + \underbrace{\sum_i p_i}_{\text{Revenue}}$

Smoothness variant for auctions

Smoothness in games: there exists strategies s_i^* :

$$\sum_i cost_i(s_i^*, s_{-i}) \leq \lambda\, OPT + \mu\, cost(s)$$

For utility games: $\sum_i u_i(s_i^*, s_{-i}) \geq \lambda\, OPT - \mu\, SW(s)$

Variant [Syrgkanis-T'13]: Auction game is λ-smooth if for some λ>0 and some strategy s* and all s we have

$$\sum_i u_i(s) \underset{\text{Nash}}{\leq} \sum_i u_i(s_i^*, s_{-i}) \geq \lambda opt - Rev(s)$$

Theorem: λ-smooth auction game ⇒ Price of anarchy for any $\leq \frac{1}{\lambda}$

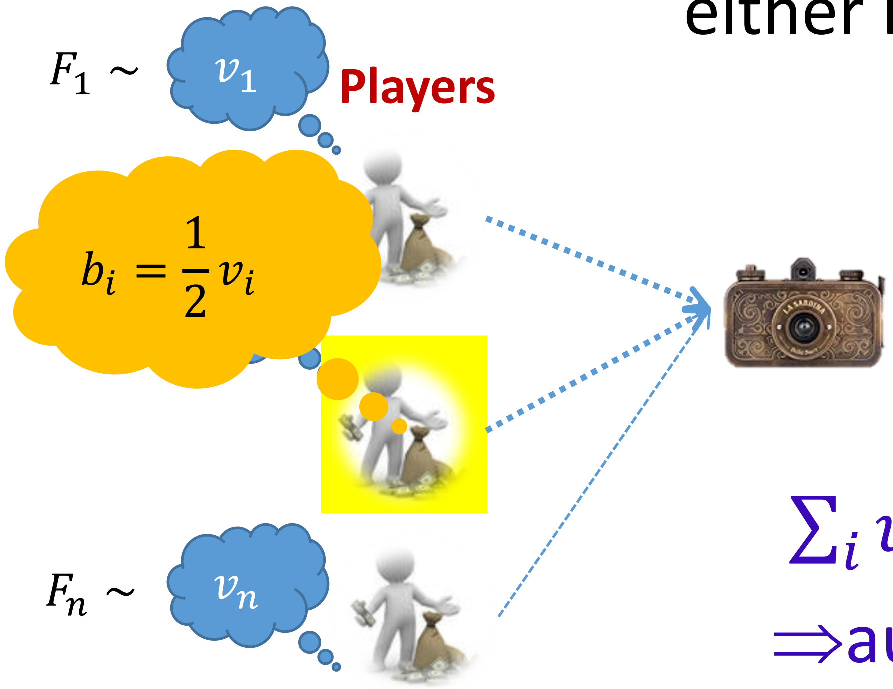
Social welfare: $\mathrm{SW}(s) = \sum_i u_i(s) + Rev(s)$

revenue

Robust Analysis: first price auction

No regret: $u_i(b) \geq \quad u_i\left(\frac{1}{2}v_i, b_{-i}\right) \geq \frac{1}{2}v_i - p\ ,0$

either i wins or price above p $\geq \frac{1}{2}v_i$



- Apply this to the top value

\+ winner doesn't regret paying

$\sum_i u_i\left(\frac{v_i}{2}, b_{-i}\right) \geq \left(\max\left(\frac{v_i}{2}\right) - p\right) + \sum_i 0$

$\Rightarrow$auction is 1/2- smooth

$\Rightarrow$a price of anarchy of 2

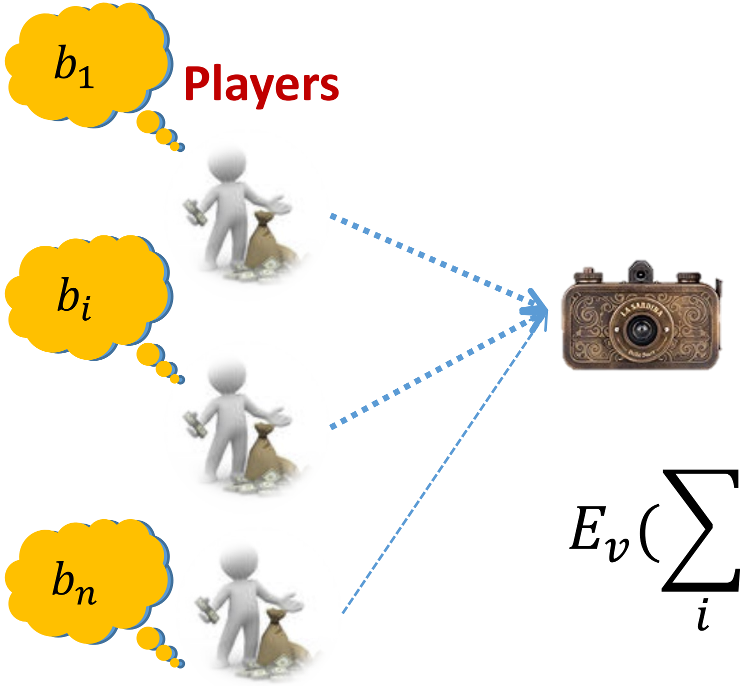
(actually… $(e-1)/e \approx 0.63$)

Bayes Nash analysis: Bayesian extension (I)

Strategy: bid as a function of value $b_i(v)$

Nash: $E_{v_{-i}b}\left[u_i\big(b(v)\big)|v_i\right] \geq E_{v_{-i}b_{-i}}\left[u_i\big(b_i', b_{-i}(v_{-i})\big)|v_i\right]$

for all b_i'



Same bound on price of anarchy, same prof (take expectation) or no-regret learning outcome

$$E_v\Big(\sum_i u_i(b)\Big) \geq \sum_i E_v\left(u_i\left(\frac{v_i}{2}, b_i\right)\right) \geq \lambda E_v(Opt(v)) - \mu E_v(Rev(b)$$

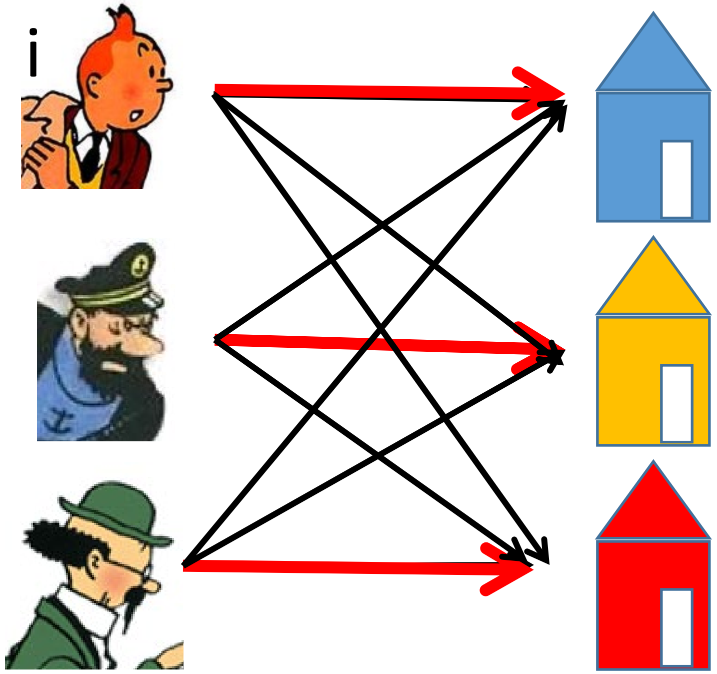
No need to bid $\frac{v_i}{2}$ just don’t regret it!

Smoothness and Bayesian games

We had $b_i^*(v) = \frac{v_i}{2}$, 0.5-smooth: Bid depends only on the players own value!

Theorem: Auction is λ-smooth and b_i^* is a function of v_i only, then price of anarchy bounded by $1/\lambda$ for arbitrary (private value) type distributions. True for Bayesian Nash equilibria as well as all no-regret learning outcomes.

Multiple items (e.g. unit demand bidders)



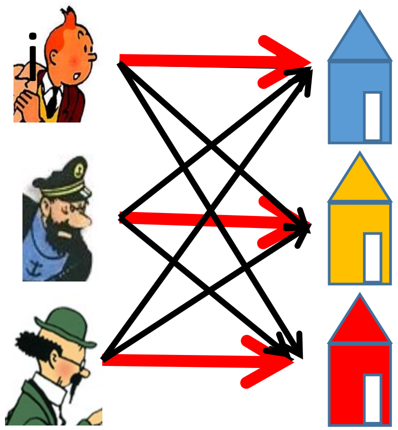
Value if i gets subset S is $v_i(S)$

for example: $v_i(S) = \max_{j \in S} v_{ij}$

Optimum is max value matching!

$$\max_{M^*} \sum_{ij \in M^*} v_{ij}$$

Multi-item first prize auction with unit demand bidders



- Optimal solution $\max\limits_{M^*} \sum_{ij \in M^*} v_{ij}$
- A bid vector b^* inducing optimal solution i bids $v_{ij}/2$ on item j_i^* assigned in i in opt ($(i, j_i^*) \in M^*$)

- Smoothness?
- $\sum_i u_i(b_i^*, b_{-i}) \geq 1/2 \sum_i v_{i j_i^*} - \sum_j \max\limits_i b_{ij} = \frac{1}{2} OPT - Rev$
- True item by item!

Bayesian extension theorem

Theorem [Roughgarden'12, Syrgkanis'12, Syrgkanis-T'13] Auction game is λ-auction smooth, and values are drawn from independent distributions, then the Price of anarchy in the Bayesian game is at most $1/\lambda$.

In addition [Hartline, Syrgkanis-T'15] also extends to learning out come in Bayesian games.

Extension theorem: OK to only think about the full information game!

Proof idea: bid b*(v)….

Trouble: depends on other players and hence we don't know……

Instead: sample opponents $\bar{v}_j$ and bid $b^\star(v_i, \bar{v}_{-i})$.

Trouble: bidding is very hard!

So many bids to consider $(b_1, b_2, \ldots, b_n)$ all possible bids on all items

Simplifications:

- Do not bid $b_j > v_j$, still bid space is $\prod_j [0, v_j]$
- Discretize, only bid multiples of ϵ, being off my an ϵ can only cause ϵ regret! Only $\prod_j v_j/\epsilon$ options
  - Assume (k-1)$\epsilon < b < k\epsilon$
  - If b wins: so does kϵ and pays too much by ϵ
  - If kϵ wins and b looses kϵ is better off.

Daskalakis-Syrgkanis'16: optimal bid is NP-hard to find or even approximate. Reduction from set-cover

Extensions beyond coarse correlated equilibria

1. What is possible when no-regret is too hard to reach
2. What can we say when there is churn: games/participants change/evolve
3. What is possible to say when there is carryover effects between iterations
4. What may be a good way to learn when cooperation may be constructive? Mostly open

Bidding options that are possible to not regret[Daskalakis-Syrgkanis'16]

- Idea: strategy space names set S of items to buy, regardless of price

- Alternate notion of no regret:

$$\frac{1}{\mathrm{T}}\sum_\tau u_i(b^\tau) \geq (1-\epsilon)\max_{S_i}\left(v_i(S_i) - \frac{1}{T}\sum_\tau p^\tau(S_i)\right) - Regret$$

Items in j ∈ S_i are evaluated against their average price! $v_j - \frac{1}{T}\sum_\tau p^\tau(j)$

No-regret for sets versus bids

- This is achievable using a variant of follow the perturbed leader.

  Need subroutine: select the set you would prefer on the average prices so far

- Is this form of no regret good enough for social welfare?

  Let S_i^* be set awarded to i in optimum. We get

$$\sum_\tau u_i(S^\tau) \geq T v_i(S_i^*) - \sum_\tau Rev^\tau(S_i^*) \text{- regret}$$

  Sum over all players

$$\sum_\tau \sum_i u_i(s^\tau) \geq T \sum_i v_i(S_i^*) - \sum_\tau \sum_i Rev^\tau(S_i^*) = T\ OPT - \sum_\tau Rev^\tau$$

Learning in Dynamic Game: [Lykouris, Syrgkanis, T. '16]

Dynamic population model:

At each step t each player i

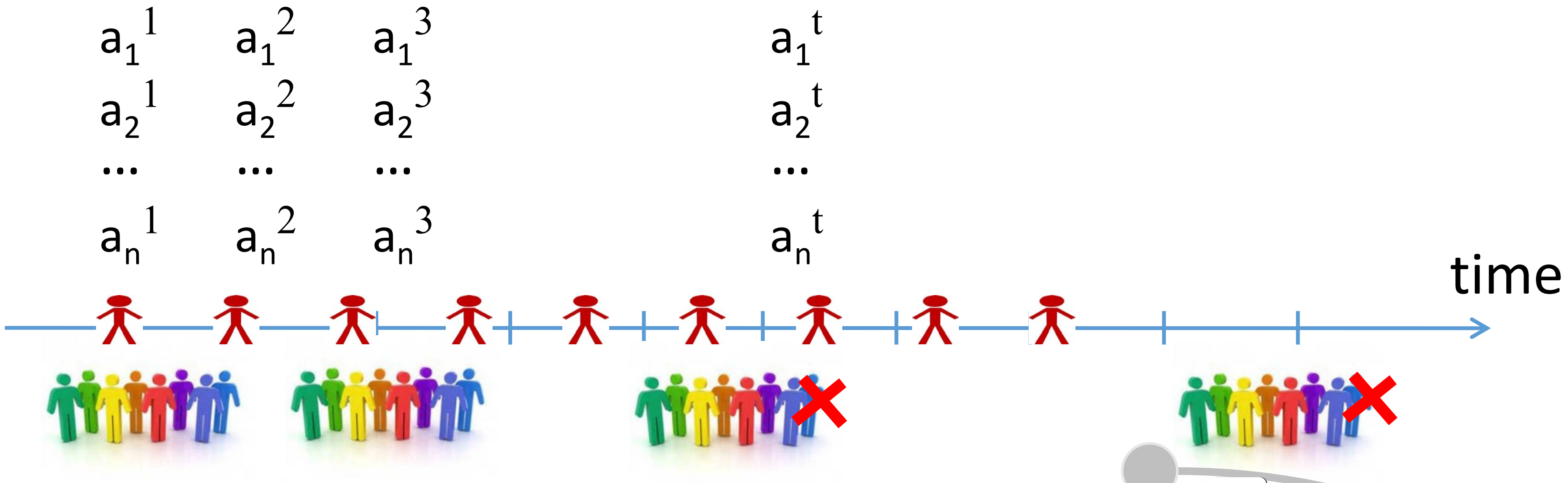
is replaced with an arbitrary new player with probability p

What should they learn from data?

No regret good enough?

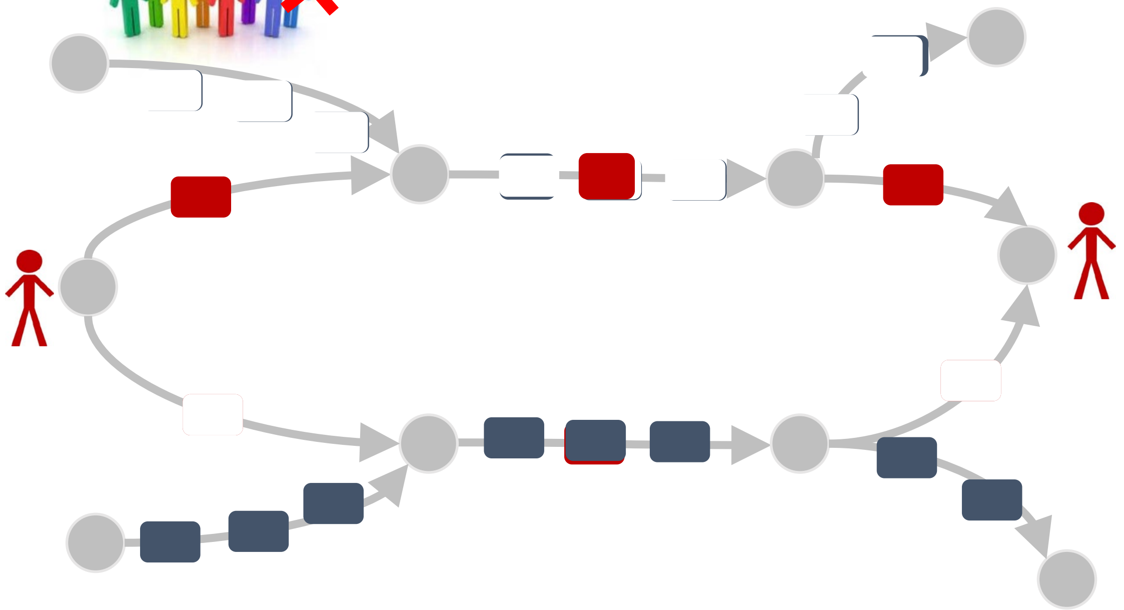
$$\sum_t cost_i(a^t) \le (1+\epsilon) \sum_t cost_i(a_i^*, a_{-i}^t) + AR$$

Need for adaptive learning



Example routing

- Strategy = path
- Best “fixed” strategy in hindsight very weak in changing environment
- Learners should/can adapt to the changing environment



Adapting result to dynamic populations

Inequality we "wish to have"

$$\sum_t cost_i(a^t; v^t) \leq \sum_t cost_i(a_i^{*t}, a_{-i}^t; v^t)$$

where a_i^{*t} is the optimum strategy for the players at time t.

with stable population = no regret for a_i^*

Too much to hope for in dynamic case:

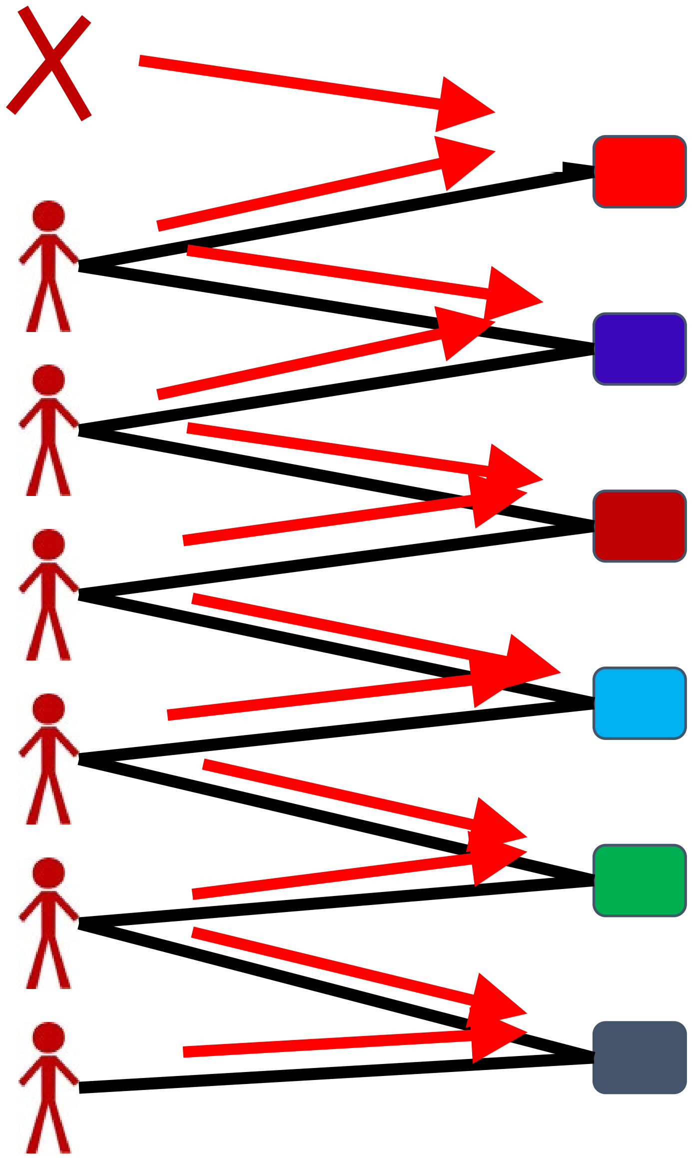
- sequence a^{*t} of optimal solutions changes too much.
- No hope of learners not to learn this well!

Change in Optimum Solution

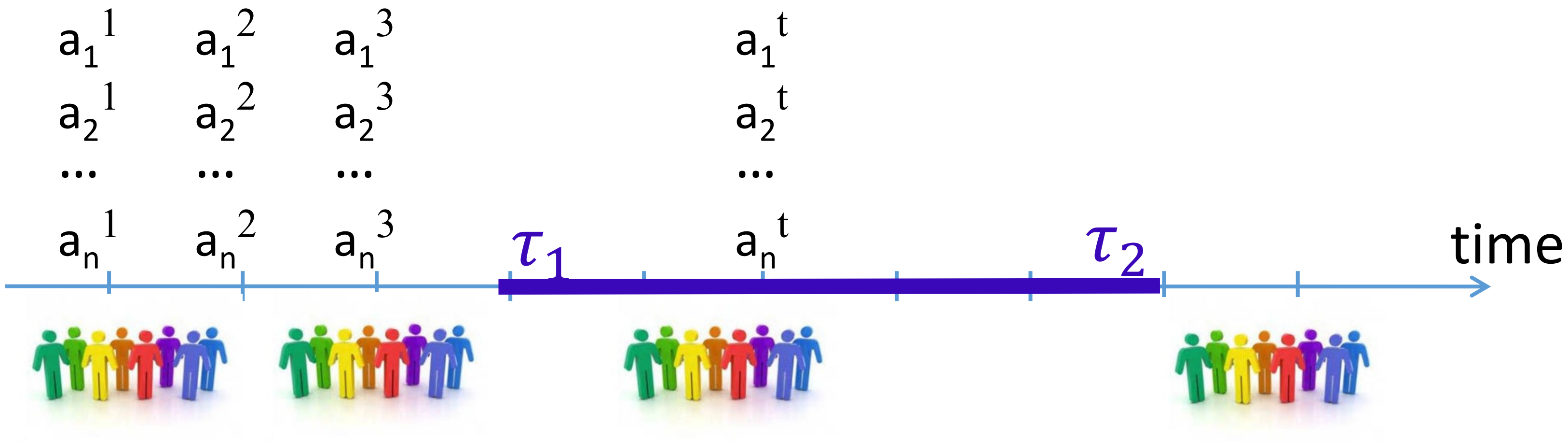
True optimum is too sensitive

- Example using matching
- The optimum solution
- One person leaving
- Can change the solution for everyone

- Np changes each step $\rightarrow$ No time to learn!! (we have p>>1/N)



Adaptive Learning



Theorem Approximate Regret [e.g., Foster,Li,Lykouris,Sridharan,T. NIPS'16]

for all player i, strategy x^τ sequence that changes k times

$$\sum_\tau u_i(s^\tau, v^\tau) \geq \sum_\tau (1+\epsilon)\, u_i(x^\tau, s_{-i}^\tau; v^\tau) + O\left(\frac{k}{\epsilon} \log m\right)$$

Using any classical learning mixed with a bit of **recency bias**

Theorem (high level)

If a game satisfies a "smoothness property"

The welfare optimization problem admits an approximation algorithm whose outcome $\tilde{a}^\star$ is stable to changes in one player's type

Then any adaptive learning outcome is approximately efficient

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

Proof idea: use this approximate solution as $\tilde{a}^\star$ in Price of Anarchy proof

With $\tilde{a}^\star$ not changing much, learners have time to learn not to regret following $\widetilde{\boldsymbol{a}^\star}$

Result (Lykouris, Syrgkanis, T'16) :

In many smooth games welfare close to Price of Anarchy even when the rate of change is high, $\boldsymbol{p} \approx \frac{\mathbf{1}}{\log \boldsymbol{n}}$ with n players, assuming **adaptive** no-regret learners

- Worst case change of player type $\Rightarrow$ need for learning players
- Bound $\boldsymbol{\alpha} \cdot \boldsymbol{\beta} \cdot \boldsymbol{\gamma}$ depends on
  - $\boldsymbol{\alpha}$ price of anarchy bound — as game gets large, goes to 1 in auctions, goes to 4/3 in linear congestion games
  - $\boldsymbol{\gamma}$ loss due to regret error — goes to 1 as $p \to 0$
  - $\boldsymbol{\beta}$ loss in opt for stable solutions — goes to 1 as $p \to 0$ & game is large

Social Welfare of Learning Outcomes

Critical Assumption: new copy of the same game is repeated (no carryover effect between rounds other than through learning)

Is this reasonable?

Cooperative Games: when no-regret is the wrong thing to do

- Simple example: repeated prisoner's dilemma: the only no-regret strategy is to defect, as defect is dominant strategy!

But defecting induces the opponent to defect:

Has effect on next round beyond the learning!

Suggested learning:

de Farias, Megiddo'06

Arora, Dekel, Tewari'12 policy regret

	C	D
C	0 0	1 -11
D	-11 1	-10 -10

Large population games: traffic routing

Morning rush-hour traffic

No carryover effect (except through the learning of the agents)

Second-by-second packet traffic

Packets take time to clear, dropped packets need to be resent in the next round

Price of Anarchy in Stateful Systems

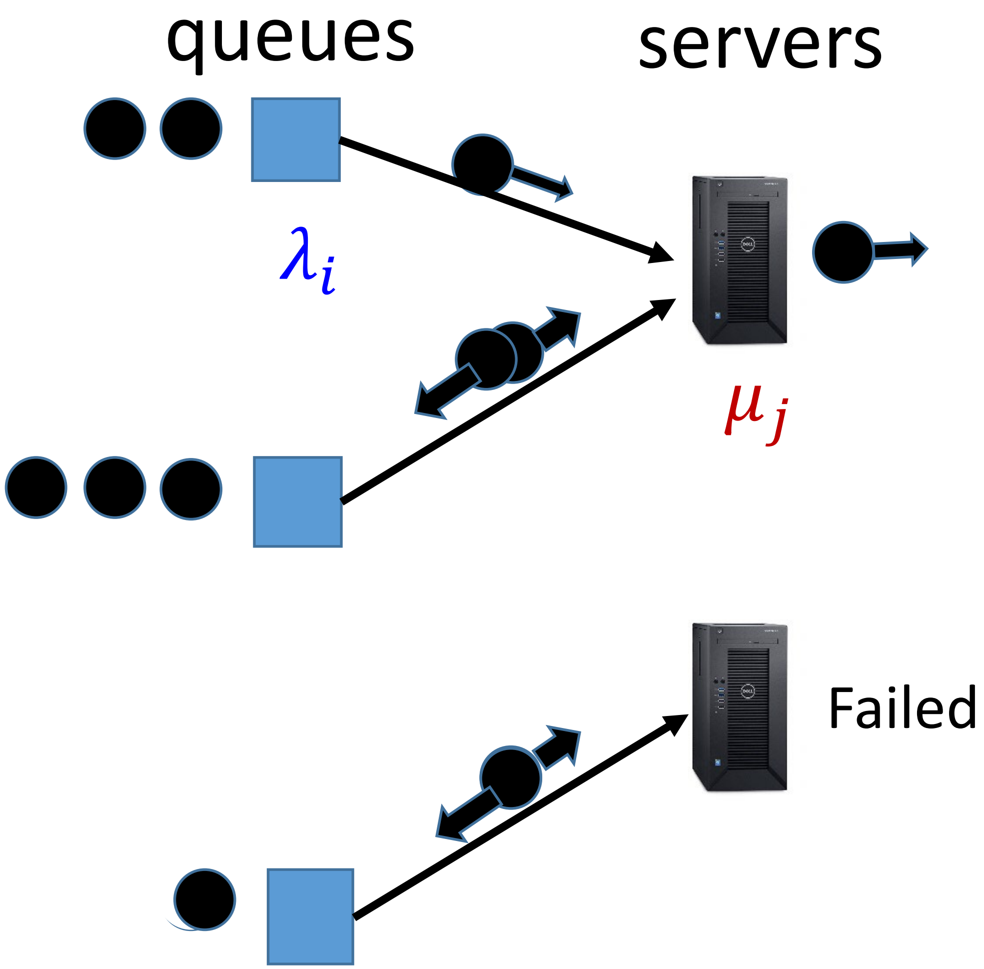
- Not as well understood: do PoA-style bounds still hold with dependence between games in each round?

Questions:

- How much extra capacity ensures good system performance despite selfish users
- Is no-regret learning the right way to learn in presence of dependence between rounds

Simple Model of Queuing

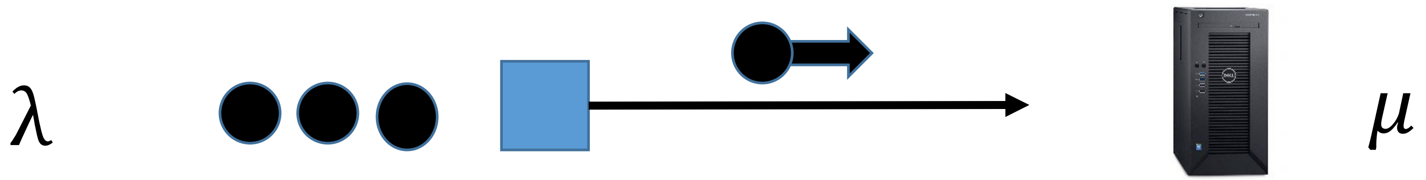
- Queue i gets new packets with a Bernoulli process with rate λ_i
- Server j succeeds at serving a packet with probability μ_j
- Each time step: each queue can send **one** packet to one of the servers to try to get serviced
- **Server can process at most one packet and unserved packets get returned to queue**
- Servers attempt to serve oldest packet



Our Main Question

How large should the server capacity be to ensure competitive, no-regret queues remain bounded in expectation over time?

- Example: one queue, one server (no learning, no competition)



- $\lambda < \mu$: expected queue size bounded (biased r.w. on the half-line)
- $\lambda = \mu$: expected queue size grows like $\Theta(\sqrt{t})$ (unbiased r.w.)
- $\lambda > \mu$: expected queue size grows linearly in t ➔ sharp threshold

One queue many servers

- The one queue faces a Bayesian multi-arm bandit learning problem to find the best server

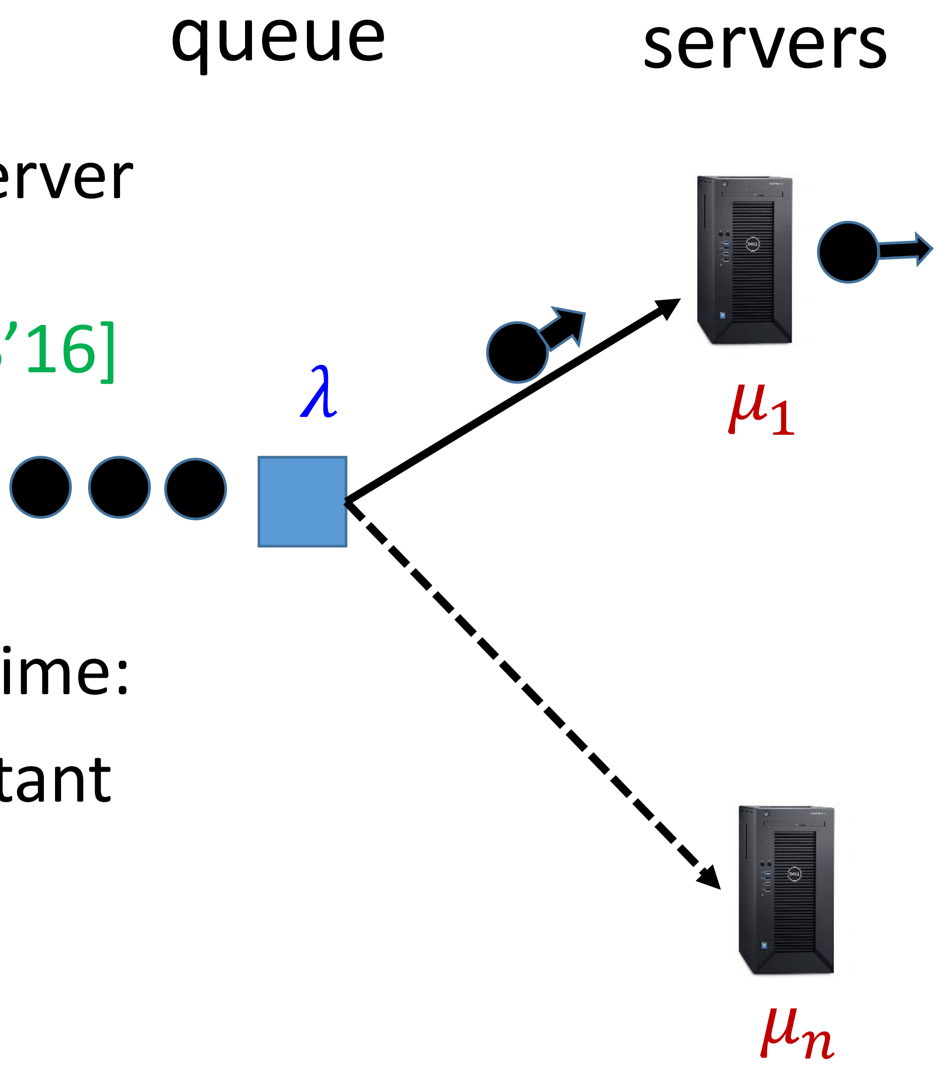
[Krishnasamy, Sen, Johari, & Shakkottai NIPS'16]

- Queue is searching for the best server:

needs $\lambda < \mu_i$

- Study the evolution of queue length over time:

goes up to $O(\log t)$ and then back to a constant once the best server is identified

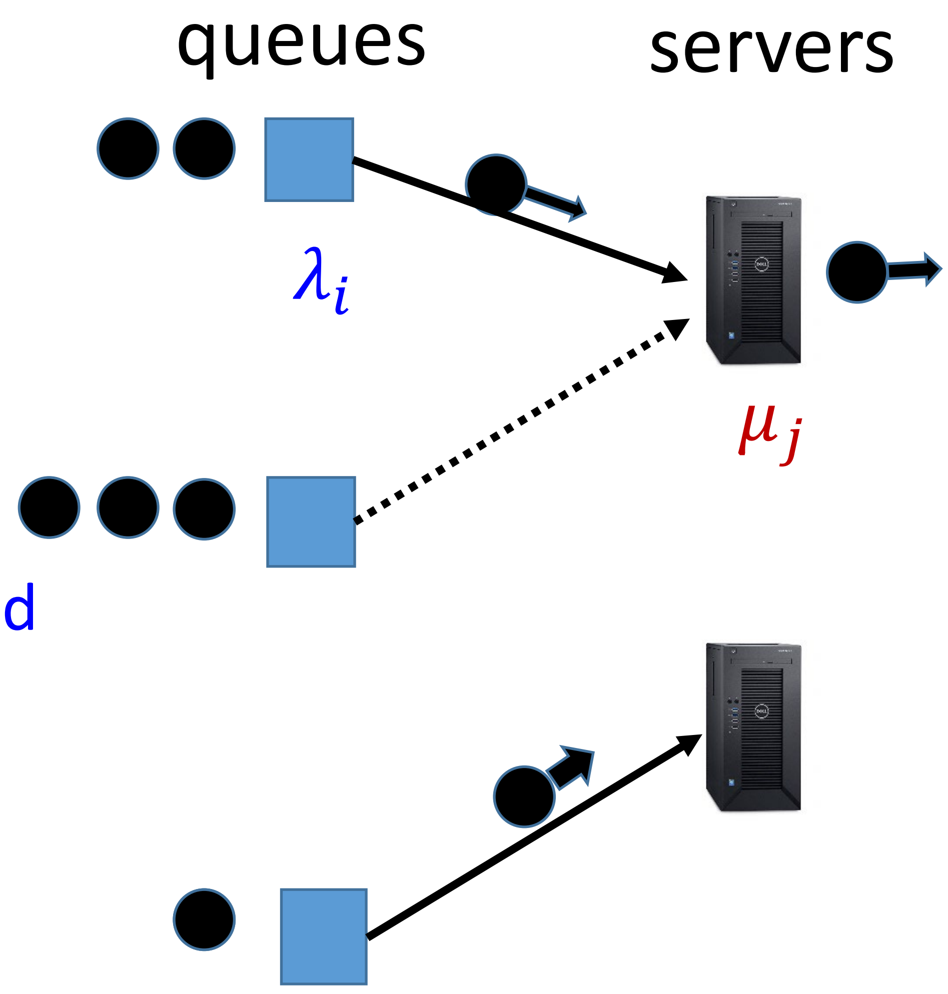


Baseline Measure: Coordinated Queues

Assume queues and servers are sorted:

$$1 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

$$1 \geq \mu_1 \geq \mu_2 \geq \cdots \geq \mu_m > 0$$

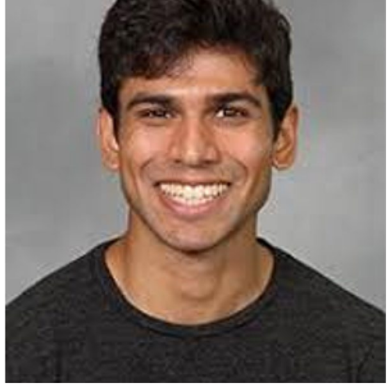


Claim: necessary/sufficient condition for centralized stability: for all *k*,

$$\sum_{i=1}^{k} \lambda_i < \sum_{i=1}^{k} \mu_i$$

(Recall can only send one packet each round)

Selfish Queuing with Priorities



- Main Theorem [informal, Gaitonde-T '20]: suppose that:
  - Servers attempt to serve oldest packet received in each round,
  - Queues use no-regret learning algorithms,
  - and for all k,

$$\sum_{i=1}^{k} \lambda_i < \frac{1}{2} \sum_{i=1}^{k} \mu_i$$

Then, **all queue sizes remain bounded** in expectation uniformly over time. Moreover, factor 1/2 is tight.

Proof Ideas

- Use potential function

$$\Phi \approx \sum_\tau \Phi_\tau$$

with $\Phi_\tau = \#$ packets aged τ or older in the system

- [Pemantle, Rosenthal '04]: random process satisfying
  i. Sufficiently regular
  ii. Negative drift when large

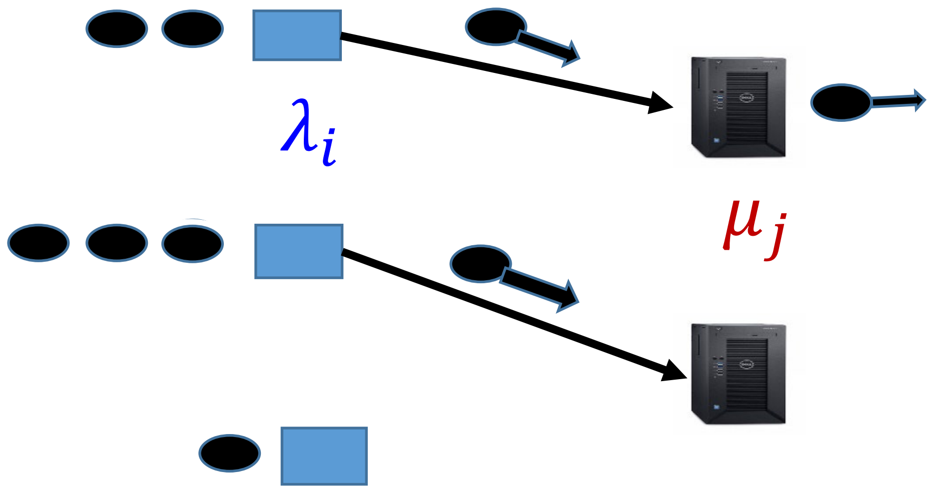
  remains bounded in expectation for all times

- No-regret + factor 2 slack implies negative drift when queues have large backup

Why Φ and How No-Regret Helps

- Look at queues with packets at least τ-old; they have priority
- Fix long window and look at best/fastest servers
- Either: i) many τ-old queues send there throughout window ➔ Φ_τ decreases by a lot

  ii) they do not ➔ had priority there so no-regret kicks in:



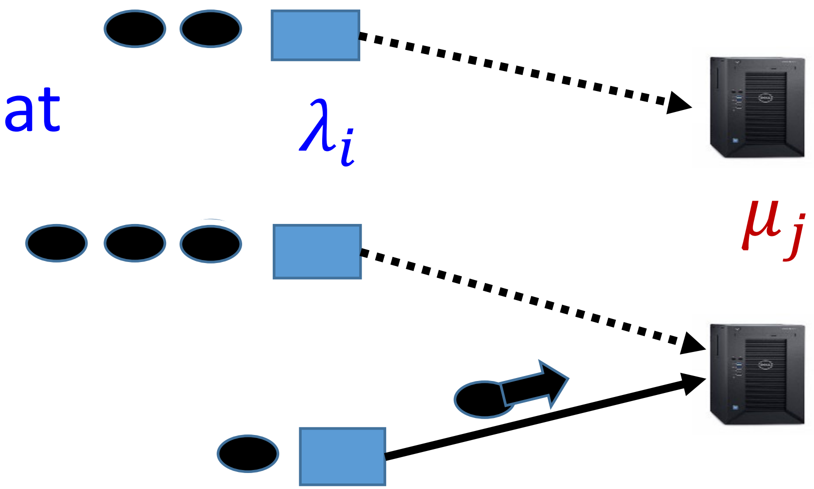
Why Φ and How No-Regret Helps

- Look at queues with packets at least τ-old; they have priority
- Fix long window and look at best/fastest servers
- Either: i) many τ-old queues send there throughout window ➔ decrease in queue size, OR

  ii) they do not ➔ had priority there so no-regret kicks in:

  Any queue with τ-old packets would have regret, unless it managed to get service for at least this much!

Apply at all thresholds τ simultaneously to get no-regret at all scales ➔ implies negative drift



Extra Technical Details

- Need no-regret to hold on specific windows of long enough size with high-probability

  unlikely bad situations **will** happen, need to be able to recover

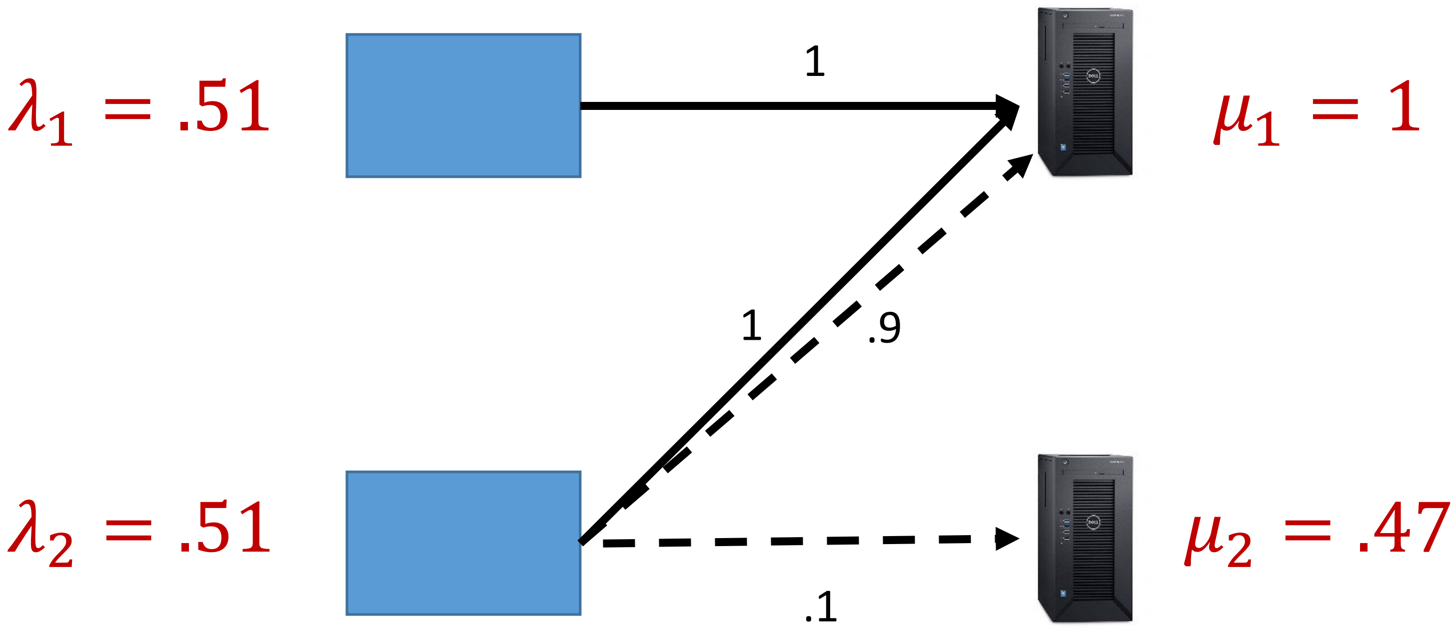
- Other technical issues for applying Pemantle/Rosenthal result: use model with deferred decisions, study *ages* instead of *sizes*

  age of oldest packet T_i^t in queue i

  $\Phi_\tau = \sum_{i:T_i^t > \tau} \lambda_i (T_i^t - \tau) \approx$ # packets age τ or older in the system

- apply concentration bounds,
- "sufficiently regular" = bounded moments

Myopia in No-Regret: Example

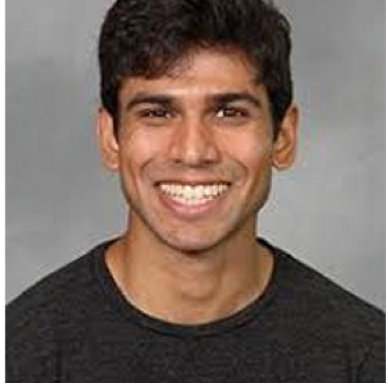


- Both sending to top server has no-regret
- Deviating gives regret
- Age/split top server equally → linear growth

- Moving to inferior server selfishly helps
- Helps top queue clear, indirectly helping both queues clear!

But the 0.47 rate causes regret!

Selfish Queuing: Price of Anarchy



Theorem 1 [Gaitonde-T ’20]: if queues use no-regret algorithms to select servers and for all k,

$$\sum_{i=1}^{k} \lambda_i < 0.5 \sum_{i=1}^{k} \mu_i$$

Then queue lengths/ages grow sublinearly.

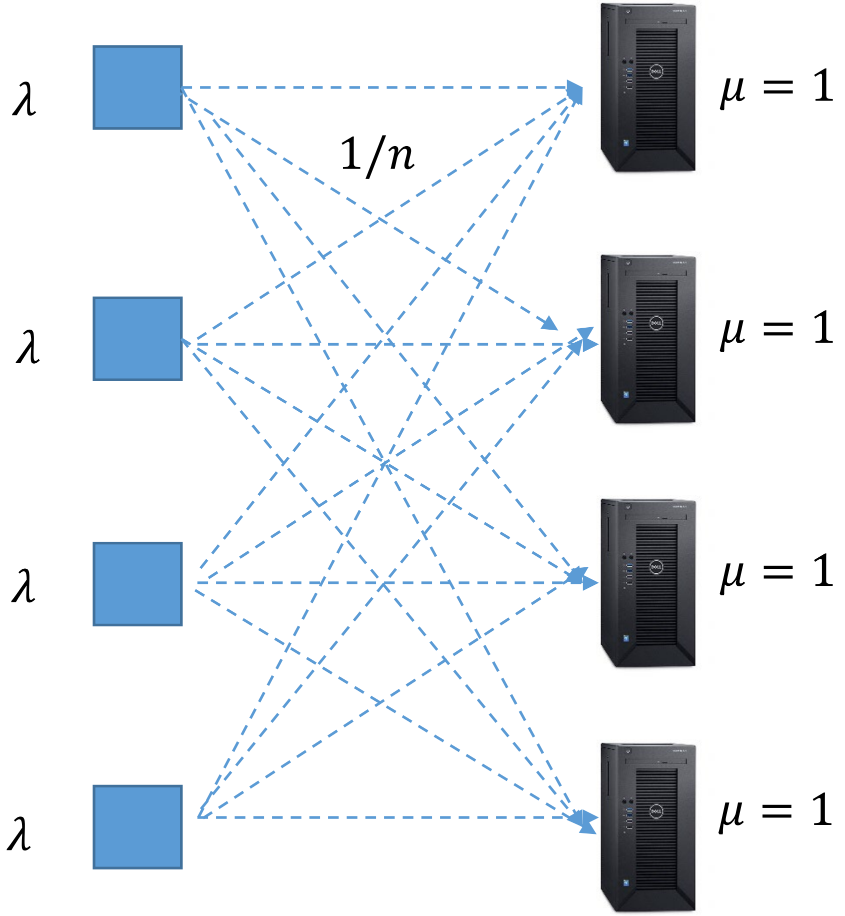
Theorem 2 [Gaitonde-T‘21]: If queues choose servers patiently, and for all k

$$\sum_{i=1}^{k} \lambda_i < 0.63 \sum_{i=1}^{k} \mu_i$$

then in every equilibrium, queue lengths/ages grow sublinearly. $0.63 = (e-1)/e$

Price of Anarchy

- Worst-case (intuitively): n equal queues, n servers with rate 1, uniform mixing → worst case needs at least $\frac{e}{e-1}$ slack
- In general: fastest-aging queue cannot benefit from deviation at equilibrium, but not clear *why*



What's Going On?

- Too myopic: not patient enough to see asymptotic benefit of "bad" servers:
- What we do: evaluate alternate outcome without considering long-term effect of the change

$$\sum_t cost_i(a^{1:t}) \leq \sum_t cost_i\left(\left(a_i^{1:t-1}, x\right), a_{-i}^{1:t}\right) + o(T)$$

- What we may want (?):

$$\sum_t cost_i(a^{1:t}) \leq \sum_t cost_i\left(x^{1:t}, a_{-i}^{1:t}\right) + o(T)$$

- We study the patient queuing game with stationary strategies

Conclusions

Thanks!

Learning in games:

- Good way to adapt to opponents
  - Takes advantage of opponent playing badly.
- No need for common prior

Learning players do well even in dynamic environments

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

Do OK in some games with carryover effect.

Question: can other forms of learning do better?

e.g., policy regret? [Arora, Dekel, Tewari'12]

Unfortunately, doesn't help in queueing Sentenac, Boursier, Perchet'21