Nonstochastic Bandit Bros: Vanilla, Partial, Delayed, Composite, Contextual

Claudio Gentile

INRIA and Google NY

cla.gentile@gmail.com

Toulouse

September 13th, 2018

Based on joint work with:

N. Alon, N. Cesa-Bianchi, P. Gaillard, S. Gerchinovitz, Y. Mansour, S. Mannor, O. Shamir

Goal of this presentation

Recent activity in the analysis of bandit problems in nonstochastic settings under various modeling assumptions, and kind of available feedback

Outline :

- Nonstochastic bandit game:
  - vanilla
  - delayed
  - composite anonymous
  - graph
- Contextual bandits for nonparametric policies

Examples thereof

Goal of this presentation

Recent activity in the analysis of bandit problems in nonstochastic settings under various modeling assumptions, and kind of available feedback

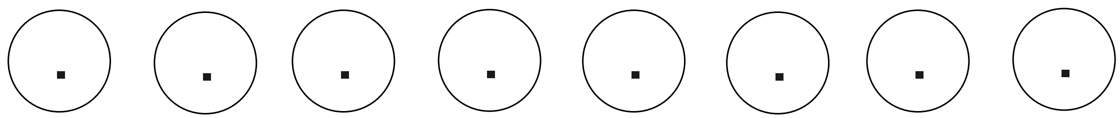
Outline :

- Nonstochastic bandit game:
  - vanilla
  - delayed
  - composite anonymous
  - graph
- Contextual bandits for nonparametric policies

Examples thereof

Nonstochastic bandit game/1

N actions for Player



For $t = 1, 2, \ldots$:

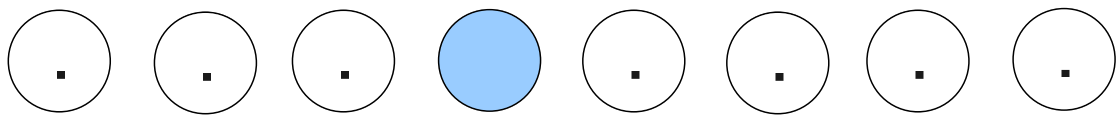
1. Losses $\ell_t(i) \in [0,1]$ are assigned by opponent to every action $i = 1 \ldots N$ (hidded to player)

2. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Player gets feedback information: $\ell_t(I_t)$

Nonstochastic bandit game/1

N actions for Player



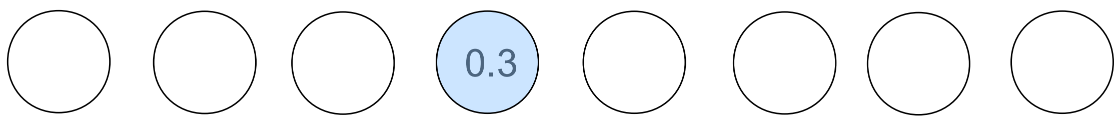
For $t = 1, 2, \ldots$:

1. Losses $\ell_t(i) \in [0,1]$ are assigned by opponent
   to every action $i = 1 \ldots N$ (hidded to player)

2. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Player gets feedback information: $\ell_t(I_t)$

Nonstochastic bandit game/1



For $t = 1, 2, \ldots$:

1. Losses $\ell_t(i) \in [0, 1]$ are assigned by opponent
   to every action $i = 1 \ldots N$ (hidded to player)

2. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Player gets feedback information: $\ell_t(I_t)$

Nonstochastic bandit game/2

Goal [external regret]: Given T rounds, Player's total loss

$$\sum_{t=1}^{T} \ell_t(I_t)$$

must be close to that of single best action in hindsight for Player

(Pseudo) Regret of Player for T rounds:

$$R_T = \max_{i=1\ldots N} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t) - \sum_{t=1}^{T} \ell_t(i)\right]$$

Want : $R_T = o(T)$ as T grows large ("no regret")

Lower bound: $\Omega(\sqrt{TN})$

Regret:

$$R_T^* = \max_{i=1\ldots N} \left(\sum_{t=1}^{T} \ell_t(I_t) - \sum_{t=1}^{T} \ell_t(i)\right)$$

Want : $R_T^* = o(T)$ as T grows large w.h.p

Nonstochastic bandit game/3: Exp3 Alg. [Auer et al. 02]

At round t pick action $I_t = i$ with probability proportional to

$$\exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right), \qquad i = 1 \dots N$$

$$\widehat{\ell}_s(i) = \begin{cases} \frac{\ell_s(i)}{\text{Pr}_s(\ell_s(i) \text{ is observed in round } s)} & \text{if } \ell_s(i) \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

- Only one nonzero component in $\widehat{\ell}_t$
- Exponentially-weighted alg with (importance sampling) loss estimates
$$\widehat{\ell}_t(i) \approx \ell_t(i)$$
- Upper bound on regret:
$$R_T \le \sqrt{TN \ln N}$$
- Improved upper bound: $O(\sqrt{TN})$ (the INF alg.) [AB09]

Nonstochastic bandit game with delay/1

For $t = 1, 2 \ldots$:

1. Losses $\ell_t(i)$ are assigned by opponent
   to every action $i = 1 \ldots N$ (hidded to player)

2. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Player gets delayed feedback information: $\ell_{t-d}(I_{t-d})$ $\qquad [t > d]$

Lower bound : $R_T \geq \sqrt{T(d+N)}$

Nonstochastic bandit game with delay/1

For $t = 1, 2 \ldots$:

1. Losses $\ell_t(i)$ are assigned by opponent
   to every action $i = 1 \ldots N$ (hidded to player)

2. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Player gets delayed feedback information: $\ell_{t-d}(I_{t-d})$ $\quad [t > d]$

Lower bound : $R_T \geq \sqrt{T(d + N)}$

Nonstochastic bandit game with delay/1

For $t = 1, 2 \ldots$:

1. Losses $\ell_t(i)$ are assigned by opponent
   to every action $i = 1 \ldots N$ (hidded to player)

2. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Player gets delayed feedback information: $\ell_{t-d}(I_{t-d})$ $\qquad [t > d]$

Lower bound : $R_T \geq \sqrt{T(d+N)}$

Nonstochastic bandit game with delay/1

For $t = 1, 2 \ldots$:

1. Losses $\ell_t(i)$ are assigned by opponent
   to every action $i = 1 \ldots N$ (hidded to player)

2. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Player gets delayed feedback information: $\ell_{t-d}(I_{t-d})$ $\qquad [t > d]$

Lower bound : $R_T \geq \sqrt{T(d + N)}$

Nonstochastic bandit game with delay/2 [CB+16]

Upper bound :

- Use importance-sampling estimate within Exp3, and update as soon as loss becomes available:

$$\widehat{\ell}_t(i) = \begin{cases} \frac{\ell_{t-d}(i)}{\mathrm{Pr}_{t-d}(I_{t-d}=i)} & \text{if } I_{t-d} = i \\ 0 & \text{otherwise} \end{cases}$$

- Cumulative regret (matching lower bound up to logs):

$$R_T = \tilde{\mathcal{O}}\left(\sqrt{T(d+N)}\right)$$

Unknown delays: [Li+18]

Collect (delayed) loss observations at time t, but use Pr_t instead of Pr_{t-d}

Composite anonymous feedback/1 [D+14,A+15,PB+17,CB+18]

- Loss of action is not charged immediatedly but spread arbitrarily over d consecutive steps

- Generalizes d-delayed feedback

- Several motivating examples in online businesses:
  - impression resulting in immediate clickthrough, later followed by conversion
  - user interacting with a recommended item (e.g. media content) multiple times over several days

- Loss observed by player at time t is composite loss i.e. sum of d loss components (accumulated effect of d-many past actions) :

$$\ell_t^{(0)}(I_t) + \ell_{t-1}^{(1)}(I_{t-1}) + \ldots + \ell_{t-d+1}^{(d-1)}(I_{t-d+1})$$

$\ell_{t-s}^{(s)}(I_{t-s}) = s$-th loss component from action I_{t-s}

Composite anonymous feedback/2 [D+14,A+15,PB+17,CB+18]

$N = 3$ actions = { 1 , 2 , 3 }

$d = 4$ loss components

					time t				
$I_{t-3} =$	2	$\ell^0_{t-3}(2)$	$\ell^1_{t-3}(2)$	$\ell^2_{t-3}(2)$	$\ell^3_{t-3}(2)$				$\in [0,1]$
$I_{t-2} =$	3		$\ell^0_{t-2}(3)$	$\ell^1_{t-2}(3)$	$\ell^2_{t-2}(3)$	$\ell^3_{t-2}(3)$			$\in [0,1]$
$I_{t-1} =$	1			$\ell^0_{t-1}(1)$	$\ell^1_{t-1}(1)$	$\ell^2_{t-1}(1)$	$\ell^3_{t-1}(1)$		$\in [0,1]$
$I_t =$	2				$\ell^0_t(2)$	$\ell^1_t(2)$	$\ell^2_t(2)$	$\ell^3_t(2)$	$\in [0,1]$
					$\notin [0,1]$				

Composite anonymous feedback/3 [D+14,A+15,PB+17,CB+18]

For $t = 1, 2 \ldots$:

1. Losses $\ell_t(i) \in [0, 1]$ are assigned (obliviously) by opponent to every action $i = 1 \ldots N$ (hidded to player)

2. Losses $\ell_t(i)$ broken up into d components (arbitrarily but obliviously):

$$\ell_t(i) = \ell_t^{(0)}(i) + \ell_t^{(1)}(i) + \ldots + \ell_t^{(d-1)}(i)$$

3. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

4. Player gets composite loss feedback information:

$$\ell_t^{(0)}(I_t) + \ell_{t-1}^{(1)}(I_{t-1}) + \ldots + \ell_{t-d+1}^{(d-1)}(I_{t-d+1})$$

Composite anonymous feedback/3 [D+14,A+15,PB+17,CB+18]

For $t = 1, 2 \ldots$:

1. Losses $\ell_t(i) \in [0,1]$ are assigned (obliviously) by opponent to every action $i = 1 \ldots N$ (hidded to player)

2. Losses $\ell_t(i)$ broken up into d components (arbitrarily but obliviously):

$$\ell_t(i) = \ell_t^{(0)}(i) + \ell_t^{(1)}(i) + \ldots + \ell_t^{(d-1)}(i)$$

3. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

4. Player gets composite loss feedback information:

$$\ell_t^{(0)}(I_t) + \ell_{t-1}^{(1)}(I_{t-1}) + \ldots + \ell_{t-d+1}^{(d-1)}(I_{t-d+1})$$

Composite anonymous feedback/3 [D+14,A+15,PB+17,CB+18]

For $t = 1, 2 \ldots$:

1. Losses $\ell_t(i) \in [0, 1]$ are assigned (obliviously) by opponent to every action $i = 1 \ldots N$ (hidded to player)

2. Losses $\ell_t(i)$ broken up into d components (arbitrarily but obliviously):

$$\ell_t(i) = \ell_t^{(0)}(i) + \ell_t^{(1)}(i) + \ldots + \ell_t^{(d-1)}(i)$$

3. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

4. Player gets composite loss feedback information:

$$\ell_t^{(0)}(I_t) + \ell_{t-1}^{(1)}(I_{t-1}) + \ldots + \ell_{t-d+1}^{(d-1)}(I_{t-d+1})$$

Composite anonymous feedback/3 [D+14,A+15,PB+17,CB+18]

For $t = 1, 2 \ldots$:

1. Losses $\ell_t(i) \in [0, 1]$ are assigned (obliviously) by opponent to every action $i = 1 \ldots N$ (hidded to player)

2. Losses $\ell_t(i)$ broken up into d components (arbitrarily but obliviously):

$$\ell_t(i) = \ell_t^{(0)}(i) + \ell_t^{(1)}(i) + \ldots + \ell_t^{(d-1)}(i)$$

3. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

4. Player gets composite loss feedback information:

$$\ell_t^o(I_{t-d+1} \ldots I_t) = \ell_t^{(0)}(I_t) + \ell_{t-1}^{(1)}(I_{t-1}) + \ldots + \ell_{t-d+1}^{(d-1)}(I_{t-d+1})$$

Composite Loss Wrapper [CB+18]

- Take Base MAB(η) as input
- $I_0 \sim \boldsymbol{p}_1 =$ uniform on actions $1 \ldots N$

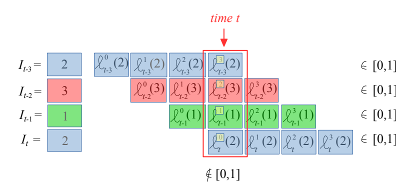
Interleave update (*up*), draw (*dr*), stay (*st*) rounds :

$$\ldots\ up\ dr\ \underbrace{st\ st\ st\ \ldots\ st}_{\geq 2d-2}\ up\ dr\ \underbrace{st\ st\ st\ \ldots\ st}_{\geq 2d-2}\ up\ dr\ \underbrace{st\ st\ st\ \ldots\ st}_{\geq 2d-2}\ \ldots$$

Stretch of stay rounds: $2d - 2 + \text{Geom}(1/(2d))$ long

- draw round: $\quad I_t \sim \boldsymbol{p}_t$ without updating $\boldsymbol{p}_t$
- stay round: $\quad I_t = I_{t-1}$ without updating $\boldsymbol{p}_t$
- update round: $I_t = I_{t-1}$, but $\boldsymbol{p}_t \to \boldsymbol{p}_{t+1}$ by feeding Base MAB with average composite loss

$$\bar{\ell}_t = \frac{1}{2d} \sum_{\tau=t-d+1}^{t} \ell^o_\tau(I_{\tau-d+1} \ldots I_\tau)$$



Stability and regret bounds [CB+18]

Stability: Base MAB $A(\eta)$ generating $\boldsymbol{p}_1, \boldsymbol{p}_2 \dots \boldsymbol{p}_t \dots$ ξ-stable if

$$\mathbb{E}\left[\sum_{i\,:\,p_{t+1}(i) > p_t(i)} p_{t+1}(i) - p_t(i)\right] \leq \xi$$

Regret of Base MAB: $R_A(T, N, \eta) \implies$ regret of Composite Loss Wrapper

$$R_T \leq T\xi + \mathcal{O}(d \cdot R_A(T/d, N, \eta))$$

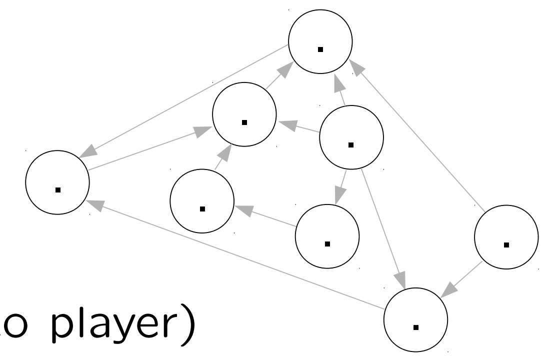
Examples:

- Exp3 ξ-stable with $\xi = \eta \quad \implies \quad R_T = \mathcal{O}\left(\sqrt{dNT \log N}\right)$
- Reduction is far more general (still pay factor $\sqrt{d}$):
  - Combinatorial Bandits
  - Bandit/Linear Convex Optimization

Lower bound (for vanilla MAB): $R_T = \Omega\left(\sqrt{dNT}\right)$

Feedback graphs/1

[MS11,A+13,K+15]



N actions for Player

Before game starts, sequence of feedback graphs $G_t = (V, E_t)$
$V = \{1, \ldots, N\}$
generated by exogenous source (hidden to player)
All self-loops included

For $t = 1, 2, \ldots$:

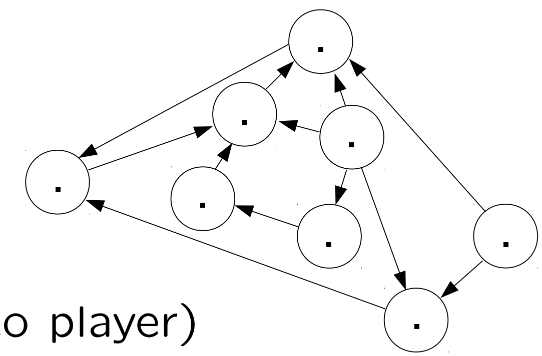
1. Losses $\ell_t(i) \in [0, 1]$ are assigned by opponent to every action $i = 1 \ldots N$ (hidded to player)

2. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Player gets feedback information: $\{\ell_t(j) : (I_t, j) \in E_t\}$

Feedback graphs/1

[MS11,A+13,K+15]



N actions for Player

Before game starts, sequence of
feedback graphs $G_t = (V, E_t)$
$V = \{1, \ldots, N\}$
generated by exogenous source (hidden to player)
All self-loops included

For $t = 1, 2, \ldots$:

1. Losses $\ell_t(i) \in [0, 1]$ are assigned by opponent to every action $i = 1 \ldots N$ (hidded to player)

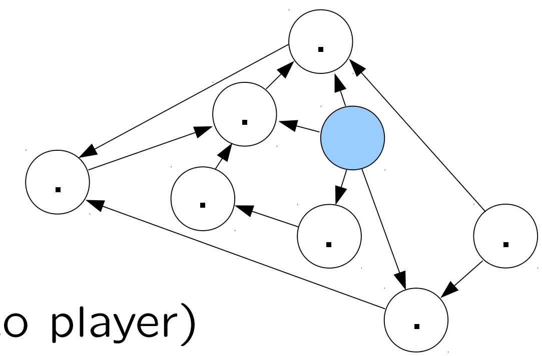
2. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Player gets feedback information: $\{\ell_t(j) : (I_t, j) \in E_t\}$

Feedback graphs/1

[MS11,A+13,K+15]

N actions for Player



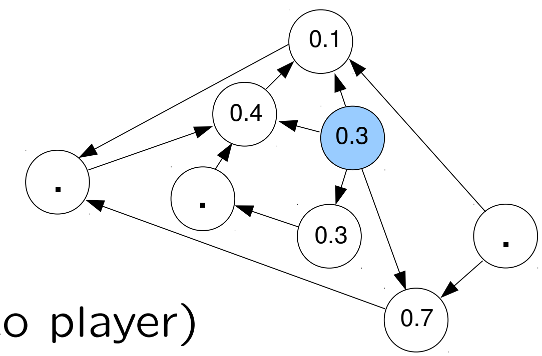
Before game starts, sequence of
feedback graphs $G_t = (V, E_t)$
$V = \{1, \ldots, N\}$
generated by exogenous source (hidden to player)
All self-loops included

For $t = 1, 2, \ldots$:

1. Losses $\ell_t(i) \in [0, 1]$ are assigned by opponent to every action $i = 1 \ldots N$ (hidded to player)
2. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$
3. Player gets feedback information: $\{\ell_t(j) : (I_t, j) \in E_t\}$

Feedback graphs/1

[MS11,A+13,K+15]



N actions for Player

Before game starts, sequence of feedback graphs $G_t = (V, E_t)$
$V = \{1, \dots, N\}$
generated by exogenous source (hidden to player)
All self-loops included

For $t = 1, 2, \dots$:

1. Losses $\ell_t(i) \in [0, 1]$ are assigned by opponent to every action $i = 1 \dots N$ (hidded to player)

2. Player picks action I_t (possibly using randomization) and incurs loss $\ell_t(I_t)$

3. Player gets feedback information: $\{\ell_t(j) : (I_t, j) \in E_t\}$

Feedback graphs/2: Exp3-IX Alg. [K+15]

At round t pick action $I_t = i$ with probability proportional to

$$\exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(i)\right), \qquad i = 1 \ldots N$$

$$\widehat{\ell}_s(i) = \begin{cases} \frac{\ell_s(i)}{\gamma_t + \Pr_s(\ell_s(i) \text{ is observed in round } s)} & \text{if } \ell_s(i) \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

- Note: prob. of observing loss of action $\neq$ prob. of playing action
- Exponentially-weighted alg with γ_t-biased (importance sampling) loss estimates

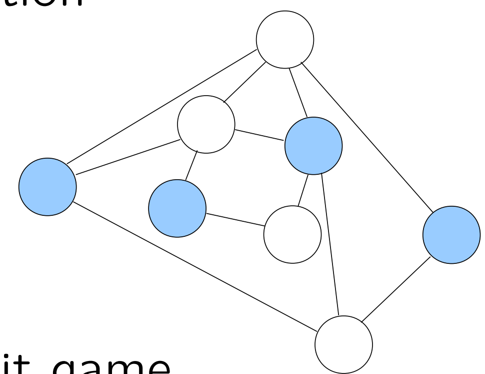
$$\widehat{\ell}_t(i) \approx \ell_t(i)$$

- Bias is controlled by $\gamma_t = 1/\sqrt{t}$

Feedback graphs/3

[A+13,K+15]

Independence number $\alpha(G_t)$: disregard edge orientation



$$\underbrace{1}_{\text{clique: full info game}} \quad \le \quad \alpha(G_t) \quad \le \quad \underbrace{N}_{\text{edgeless: bandit game}}$$

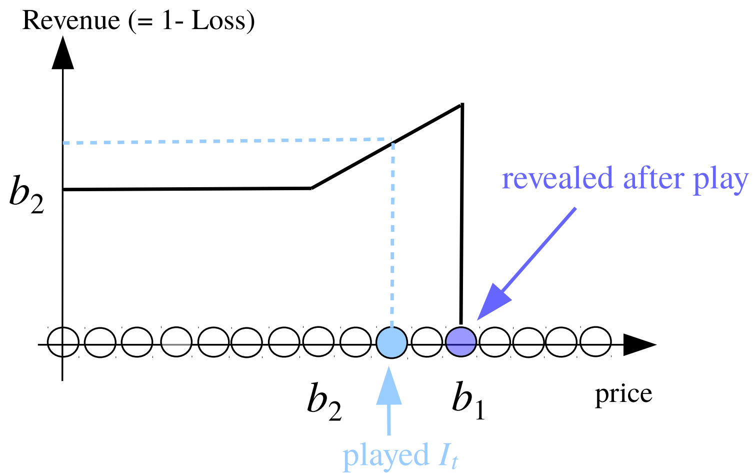
Regret analysis:

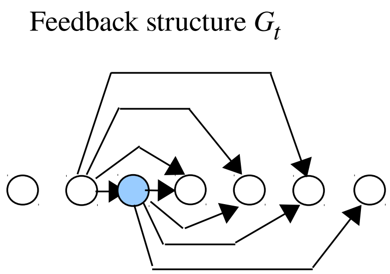
$$R_T = O\left(\ln(TN)\sqrt{\sum_{t=1}^{T}\alpha(G_t)}\right)$$

If $G_t = G \ \forall t$:

$$R_T = \tilde{O}\left(\sqrt{T\alpha(G)}\right)$$

Feedback graphs/4: Simple example





- Second-price auction with reserve (seller side)
  highest bid revealed to seller (e.g. AppNexus)
- Auctioneer is third party
- After seller plays reserve price I_t, both seller's revenue and highest bid revealed to him/her
- Seller/Player in a position to observe all revenues for prices $j \geq I_t$
- $\alpha(G) = 1$: $R_T = O\left(\ln(TN)\sqrt{T}\right)$ (full info game up to logs)

Goal of this presentation

Recent activity in the analysis of bandit problems in nonstochastic settings under various modeling assumptions, and kind of available feedback

Outline :

- Nonstochastic bandit game:
  - vanilla
  - delayed
  - composite anonymous
  - graph
- Contextual bandits for nonparametric policies

Examples thereof

Learning against Lipschitz policies/1

Ingredients:

- Context (metric) space $\mathcal{X}$ (e.g., $\mathcal{X} = \mathbf{R}^n$)
- Action (metric) space $\mathcal{Y}$ (e.g., $\mathcal{Y} = [0,1]$)
- Class of Lipschitz (and bounded) policies $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$
- (One-sided) Lipschitz loss functions $\ell_t : \mathcal{Y} \to [0,1]$

Learning protocol(s):

- Opponent picks context $x_t \in \mathcal{X}$ and loss $\ell_t(\cdot)$
- Player observes x_t, picks action $\widehat{y}_t \in \mathcal{Y}$, and incurs loss $\ell_t(\widehat{y}_t)$
- Player observes:
  - $\ell_t(\widehat{y}_t)$ only [bandit info: contextual bandit]
  - $\ell_t(y)$ $\forall y \geq \widehat{y}_t$ [one-sided full info: contextual one-sided expert]
  - $\ell_t(y)$ $\forall y \in \mathcal{Y}$ [full info: contextual expert]

Learning against Lipschitz policies/2

(Pseudo) Regret of Player for T rounds w.r.t. $\mathcal{F}$:

$$R_T(\mathcal{F}) = \max_{f \in \mathcal{F}} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\widehat{y}_t) - \sum_{t=1}^{T} \ell_t(f(x_t))\right]$$

Want : $R_T = o(T)$ as T grows large ("no regret")
for any sequence of contexts $x_1, x_2, \ldots x_t, \ldots$

Yardstick: Value of full info game [RST15]

$$V_T(\mathcal{F}) = \sup_{x_1} \inf_{q_1 \in \Delta(\mathcal{Y})} \sup_{y_1} \mathbb{E}_{\widehat{y}_1 \sim q_1} \ldots \sup_{x_T} \inf_{q_T \in \Delta(\mathcal{Y})} \sup_{y_T} \mathbb{E}_{\widehat{y}_T \sim q_T} \left[\sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) - \min_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)\right]$$

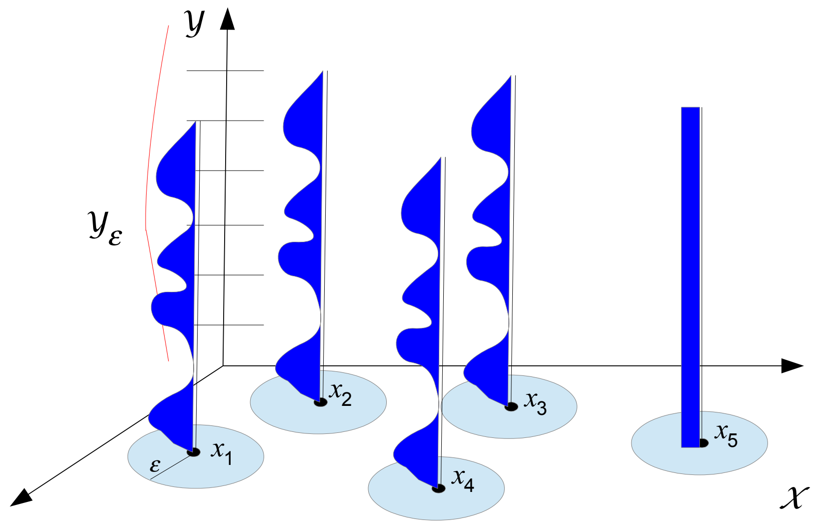
In particular:

$$\mathcal{F} = \{f : [0,1]^n \to [0,1], \quad f \text{ is 1-Lipschitz}\}$$

give

$$V_T(\mathcal{F}) = \begin{cases} \tilde{O}(T^{\frac{n-1}{n}}) & \text{if } n \geq 2 \\ \tilde{O}(\sqrt{T}) & \text{if } n = 1 \end{cases}$$

Contextual bandit game: a folk algorithm [K04,S14,...]

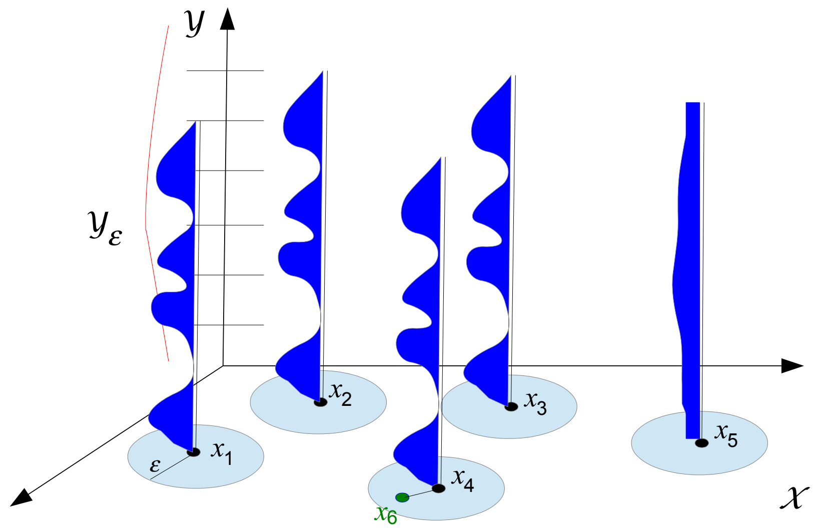


Each newly created ball centered in x_t hosts instance of EXP3 over discretized action space $\mathcal{Y}_\epsilon$

- If x_t outside any ball so far, create new ball centered on x_t
- Determine active EXP3 instance by past center x_s closest to x_t
- Draw action $\widehat{y}_t$ according to active EXP3 and update its weights only

Remark: No. balls never exceeds T

Contextual bandit game: a folk algorithm [K04,S14,...]



Each newly created ball centered in x_t hosts instance of EXP3 over discretized action space $\mathcal{Y}_\epsilon$

- If x_t outside any ball so far, create new ball centered on x_t
- Determine active EXP3 instance by past center x_s closest to x_t
- Draw action $\widehat{y}_t$ according to active EXP3 and update its weights only

Remark: No. balls never exceeds T

Contextual bandit game: regret bounds [K04,S14,...]

- $n =$ metric dimension of $\mathcal{X}$
- $1 =$ metric dimension of $\mathcal{Y}$

Then:

- Lipschitz losses : $\tilde{O}(T^{\frac{n+2}{n+3}})$ [folk alg]
- Convex losses : $\tilde{O}(T^{\frac{n+1}{n+2}})$ [folk alg + BEL16]
- Lower bound for $\underbrace{n = 0}_{no\ context}$: $\Omega(T^{\frac{2}{3}})$ [B+11]

In all cases:

- Exploit finite coverability of $\mathcal{X}$ and $\mathcal{Y}$
- Set radius ϵ appropriately

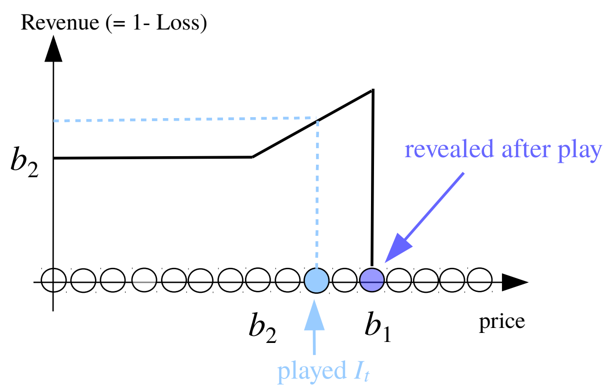
Very recent improvement in the finite action space case [FK18]

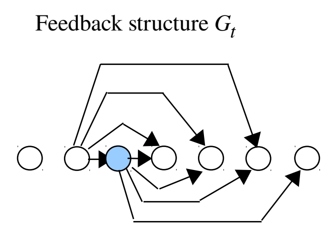
$$\tilde{O}(T^{\frac{n}{n+1}})$$

Contextual one-sided expert game/1 [CB+17]

Using Exp3-IX-like combined with folk alg on ϵ-balls over $\mathcal{X}$ yields regret

$$R_T(\mathcal{F}) \lesssim \sqrt{T \ln N_\epsilon} + T\epsilon \quad \text{if the } \ell_t \text{ are (one sided-)Lipschitz}$$
$$\lesssim T^{\frac{n+1}{n+2}} \quad \text{when optimizing on } \epsilon$$





Remark 1: No context ($n = 0$) case : $R_T(\mathcal{F}) = \tilde{\mathcal{O}}(\sqrt{T})$

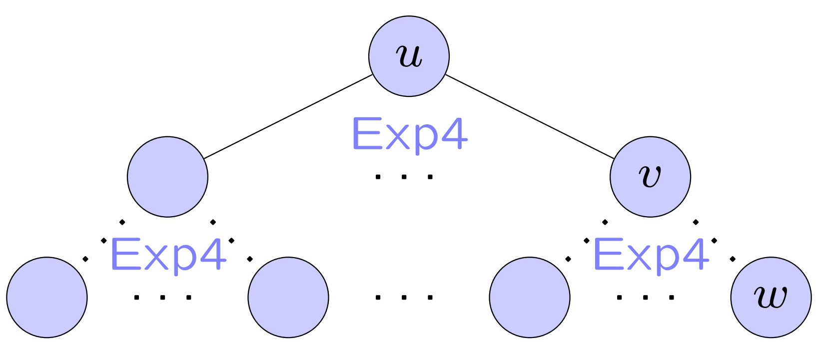
Remark 2: More general notions of one sided Lipschitz recently being used in online optimization (dispersion condition) and regret analysis in auction algs (Δ^0-Lipschitz) [F+18,B+18]

We can do better in the Lipschitz case

Contextual one-sided expert game/2: Chaining/1 [CB+17]

Ideas of the algorithm:
Hierarchical covering of $\mathcal{F}$ = tree whose nodes are functions in $\mathcal{F}$

- The nodes at each depth m define a (2^{-m})-covering of $\mathcal{F}$
- Any function $f^* \in \mathcal{F}$ is represented by unique path/chain in the tree
- Run an instance of Exp4 (adapted to one-sided expert feedback) on each node of tree
- Instance A_f at node f uses the predictions of child instances as expert advice



Level m $\rightsquigarrow$ 2^{-m} covering of $\mathcal{F}$

Level $m+1$ $\rightsquigarrow$ $2^{-(m+1)}$ covering

Level M (leaves) $\rightsquigarrow$ 2^{-M} covering

Contextual one-sided expert game/2: Chaining/2 [CB+17]

Key issues (Lipschitz losses):

- Small local ranges: losses associated with neighboring nodes are close
- Local version of Exp4 scaling with loss range: possible because of richer feedback
- Regret:

$$R_T(\mathcal{F}) \lesssim \gamma T + \int_\gamma^1 \sqrt{\frac{T}{\gamma} \ln N(\mathcal{F}, \epsilon)} d\epsilon \qquad \forall \gamma > 0$$

$$\lesssim T^{\frac{n}{n+1}} \quad \text{(when } \mathcal{F} \text{ are Lipschitz on } [0,1]^n\text{)}$$

- Improvements when $\mathcal{F} =$ Lipschitz functions on $[0,1]^n$
time efficient algorithm (wavelet-based approx.):
  - Improved regret rate $T^{\frac{n-1/3}{n+2/3}}$
  - Running time per round: $\approx T^\alpha$, $\alpha < 2$

Learning against Lipschitz policies

Bounds abound !

Exponents of T:

- Contextual bandits:
  - General Lipschitz losses: $\frac{n+2}{n+3}$
  - Convex losses: $\frac{n+1}{n+2}$
  - General Lipschitz but finite actions $\frac{n}{n+1}$ [FK18]
- Contextual one-sided:
  - General Lipschitz losses: $\frac{n}{n+1}$
  - One-sided Lipschitz losses: $\frac{n+1}{n+2}$
  - Rectangular context space and general Lipschitz losses ($n \geq 1$): $\frac{n-1/3}{n+2/3}$
- Contextual experts ($n \geq 2$): $\frac{n-1}{n}$ (tight) [RST15]

Conclusions and open questions

- Recent activity in nonstochastic bandits problems
- Several combinations are possible

Some open questions

In the composite anonymous feedback :

- Time-varying delay d
- fully adaptive adversaries (partially adaptive still possible)

In learning with Lipschitz policies :

- Tighter upper bounds with efficient alg:
  - folk approach need not capture complexity of $\mathcal{F}$
  - Covering $\mathcal{F}$ in function space does the job but algs. not efficient
- Lower bounds