Efficient learning by implicit exploration in bandit problems with side observations

Tomáš Kocák, Gergely Neu, Michal Valko, Rémi Munos
SequeL team, INRIA Lille - Nord Europe, France

SequeL – INRIA Lille

SequeL seminar

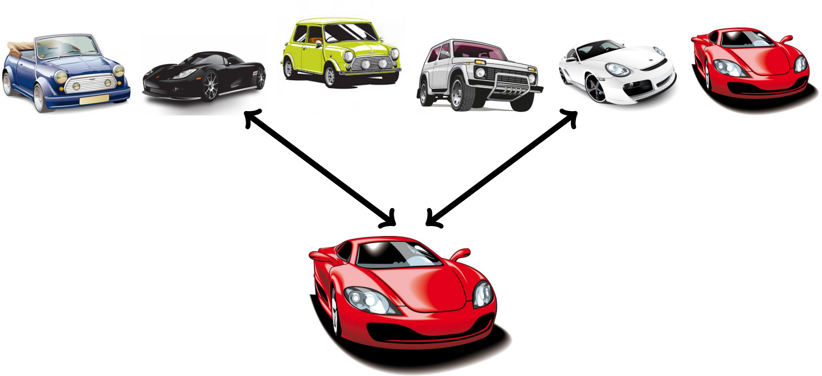
Lille, November 2014

Example 1

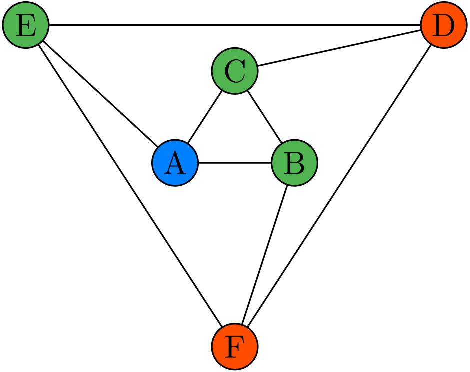
     

Example 1

Example 1



Example 1



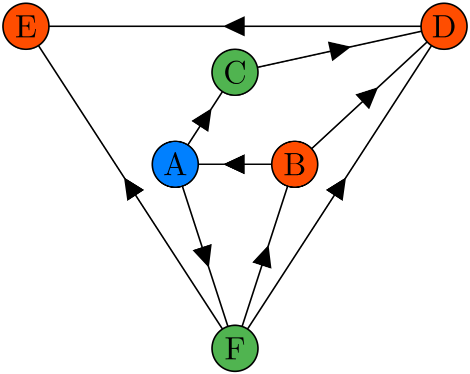
Example 2




Example 2

Example 2

Example 2



Learning setting

In each time step $t = 1, \ldots, T$

- **Environment (adversary):**
  - Privately assigns losses to actions
  - Generates an observation graph

Learning setting

In each time step $t = 1, \ldots, T$

- **Environment (adversary):**
  - Privately assigns losses to actions
  - Generates an observation graph
    - Undirected / Directed

Learning setting

In each time step $t = 1, \ldots, T$

- **Environment (adversary):**
  - Privately assigns losses to actions
  - Generates an observation graph
    - Undirected / Directed
    - Disclosed / Not disclosed

Learning setting

In each time step $t = 1, \ldots, T$

- **Environment (adversary):**
  - Privately assigns losses to actions
  - Generates an observation graph
    - Undirected / Directed
    - Disclosed / Not disclosed
- **Learner:**
  - Plays action $I_t \in [N]$
  - Obtain loss ℓ_{t,I_t} of action played
  - Observe losses of neighbors of I_t

Learning setting

In each time step $t = 1, \ldots, T$

- **Environment (adversary):**
  - Privately assigns losses to actions
  - Generates an observation graph
    - Undirected / Directed
    - Disclosed / Not disclosed
- **Learner:**
  - Plays action $I_t \in [N]$
  - Obtain loss ℓ_{t,I_t} of action played
  - Observe losses of neighbors of I_t
    - Graph: disclosed

Learning setting

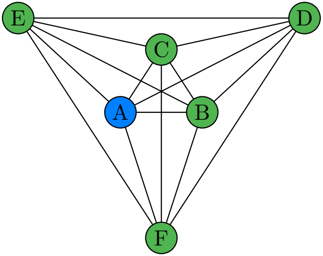
In each time step $t = 1, \ldots, T$

- **Environment (adversary):**
  - Privately assigns losses to actions
  - Generates an observation graph
    - Undirected / Directed
    - Disclosed / Not disclosed
- **Learner:**
  - Plays action $I_t \in [N]$
  - Obtain loss ℓ_{t,I_t} of action played
  - Observe losses of neighbors of I_t
    - Graph: disclosed
- **Performance measure:** Total expected regret

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

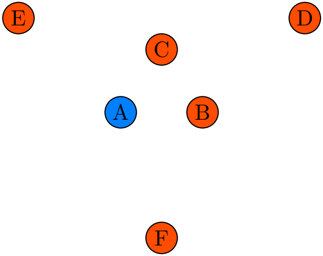
Full Information setting

- Pick an action (e.g. action A)
- Observe losses of all actions
- $R_T = \widetilde{\mathcal{O}}(\sqrt{T})$



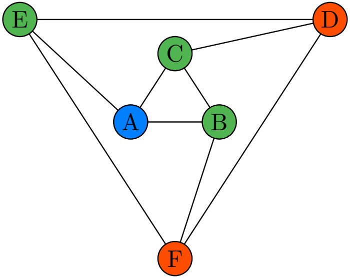
Bandit setting

- Pick an action (e.g. action A)
- Observe loss of a chosen action
- $R_T = \widetilde{\mathcal{O}}(\sqrt{NT})$



Side observation (Undirected case)

- Pick an action (e.g. action A)
- Observe losses of neighbors

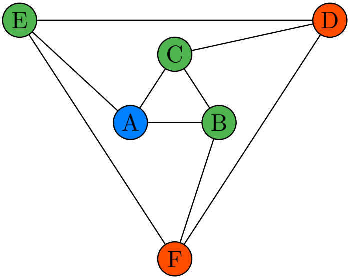


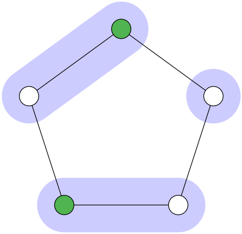
Side observation (Undirected case)

- Pick an action (e.g. action A)
- Observe losses of neighbors

Mannor and Shamir (ELP algorithm)

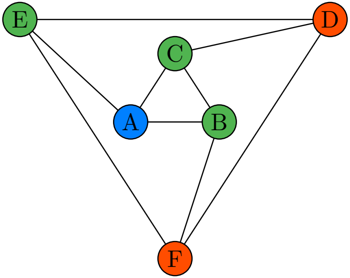
- Need to know graph
- Clique decomposition (c cliques)
- $R_T = \widetilde{\mathcal{O}}(\sqrt{cT})$





Side observation (Undirected case)

- Pick an action (e.g. action A)
- Observe losses of neighbors

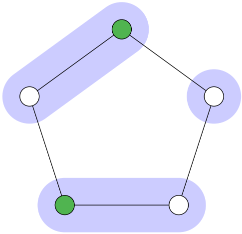


Mannor and Shamir (ELP algorithm)

- Need to know graph
- Clique decomposition (c cliques)
- $R_T = \widetilde{\mathcal{O}}(\sqrt{cT})$

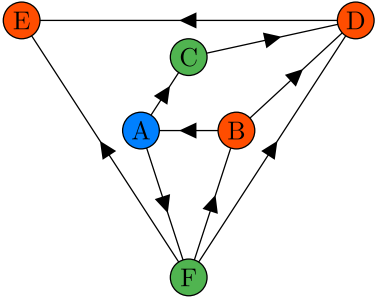
Alon, Cesa-Bianchi, Gentile, Mansour

- No need to know graph
- Independence set of α actions
- $R_T = \widetilde{\mathcal{O}}(\sqrt{\alpha T})$



Side observation (Directed case)

- Pick an action (e.g. action A)
- Observe losses of neighbors

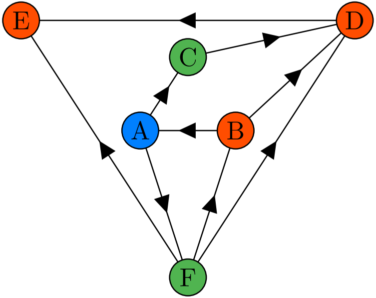


Side observation (Directed case)

- Pick an action (e.g. action A)
- Observe losses of neighbors

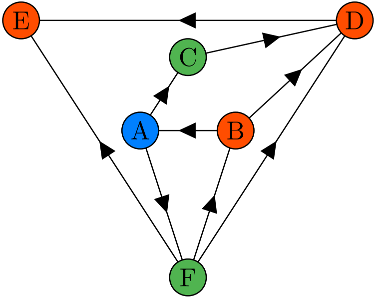
Alon, Cesa-Bianchi, Gentile, Mansour

- **Exp3-DOM**
- Need to know graph
- Need to find dominating set
- $R_T = \widetilde{\mathcal{O}}(\sqrt{\alpha T})$



Side observation (Directed case)

- Pick an action (e.g. action A)
- Observe losses of neighbors



Alon, Cesa-Bianchi, Gentile, Mansour

- **Exp3-DOM**
- Need to know graph
- Need to find dominating set
- $R_T = \widetilde{\mathcal{O}}(\sqrt{\alpha T})$

Our solution: Exp3-IX

- No need to know graph
- $R_T = \widetilde{\mathcal{O}}(\sqrt{\alpha T})$

Exp3 algorithms in general

- **Compute weights** using loss estimates $\hat{\ell}_{t,i}$.

$$w_{t,i} = \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,i}\right)$$

- **Play action** I_t such that

$$\mathbb{P}(I_t = i) = p_{t,i} = \frac{w_{t,i}}{W_t} = \frac{w_{t,i}}{\sum_{j=1}^{N} w_{t,j}}$$

- **Update loss estimates** (using observability graph)

Exp3 algorithms in general

- **Compute weights** using loss estimates $\hat{\ell}_{t,i}$.

$$w_{t,i} = \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,i}\right)$$

- **Play action** I_t such that

$$\mathbb{P}(I_t = i) = p_{t,i} = \frac{w_{t,i}}{W_t} = \frac{w_{t,i}}{\sum_{j=1}^{N} w_{t,j}}$$

- **Update loss estimates** (using observability graph)

How the algorithms approach to bias variance tradeoff?

Bias variance tradeoff approaches

- Approach of previous algorithms – Mixing
  - Bias sampling distribution $\mathbf{p}_t$ over actions
    - $\mathbf{p}'_t = (1-\gamma)\mathbf{p}_t + \gamma\mathbf{s}_t$ – mixed distribution
    - $\mathbf{s}_t$ – probability distribution which supports exploration
  - Loss estimates $\hat{\ell}_{t,i}$ are unbiased

- Approach of our algorithm – **I**mplicit e**X**ploration (**IX**)
  - Bias loss estimates $\hat{\ell}_{t,i}$
    - Biased loss estimates $\implies$ biased weights
    - Biased weights $\implies$ biased probability distribution
  - No need for mixing

Mannor and Shamir - ELP algorithm

- $\mathbb{E}[\hat{\ell}_{t,i}] = \ell_{t,i}$ – unbiased loss estimates
- $p'_{t,i} = (1-\gamma)p_{t,i} + \gamma s_{t,i}$ – bias by mixing
- $\mathbf{s}_t = \{s_{t,1}, \ldots, s_{t,N}\}$ – probability distribution over the action set

$$\mathbf{s}_t = \arg\max_{\mathbf{s}_t} \left[\min_{j\in[N]} \left(s_{t,j} + \sum_{k\in N_{t,j}} s_{t,k} \right) \right] = \arg\max_{\mathbf{s}_t} \left[\min_{j\in[N]} q_{t,j} \right]$$

- $q_{t,j}$ – probability that loss of j is observed according to $\mathbf{s}_t$

Mannor and Shamir - ELP algorithm

- $\mathbb{E}[\hat{\ell}_{t,i}] = \ell_{t,i}$ – unbiased loss estimates
- $p'_{t,i} = (1-\gamma)p_{t,i} + \gamma s_{t,i}$ – bias by mixing
- $\mathbf{s}_t = \{s_{t,1}, \ldots, s_{t,N}\}$ – probability distribution over the action set

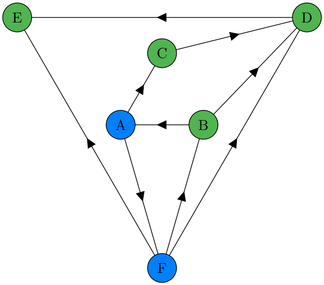
$$\mathbf{s}_t = \arg\max_{\mathbf{s}_t} \left[\min_{j\in[N]} \left(s_{t,j} + \sum_{k\in N_{t,j}} s_{t,k} \right) \right] = \arg\max_{\mathbf{s}_t} \left[\min_{j\in[N]} q_{t,j} \right]$$

- $q_{t,j}$ – probability that loss of j is observed according to $\mathbf{s}_t$
- **Computation of $\mathbf{s}_t$**
  - Graph needs to be disclosed
  - Solving simple linear program
- Needs to know graph before playing an action
- Graphs can be only undirected

Alon, Cesa-Bianchi, Gentile, Mansour - Exp3-DOM

- $\mathbb{E}[\hat{\ell}_{t,i}] = \ell_{t,i}$ – unbiased loss estimates
- $p'_{t,i} = (1-\gamma)p_{t,i} + \gamma s_{t,i}$ – bias by mixing
- $\mathbf{s}_t = \{s_{t,1}, \ldots, s_{t,N}\}$ – probability distribution over the action set

$$s_{t,i} = \begin{cases} \frac{1}{r} & \text{if } i \in R;\ |R| = r \\ 0 & \text{otherwise.} \end{cases}$$



- R – dominating set of r elements
- $\mathbf{s}_t$ – uniform distribution over R
- Needs to know graph beforehand
- Graphs can be directed

Previous algorithms - loss estimates

$$\hat{\ell}_{t,i} = \begin{cases} \ell_{t,i}/o_{t,i} & \text{if } \ell_{t,i} \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{E}[\hat{\ell}_{t,i}] = \frac{\ell_{t,i}}{o_{t,i}} o_{t,i} + 0(1 - o_{t,i}) = \ell_{t,i}$$

Previous algorithms - loss estimates

$$\hat{\ell}_{t,i} = \begin{cases} \ell_{t,i}/o_{t,i} & \text{if } \ell_{t,i} \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{E}[\hat{\ell}_{t,i}] = \frac{\ell_{t,i}}{o_{t,i}} o_{t,i} + 0(1 - o_{t,i}) = \ell_{t,i}$$

Exp3-IX - loss estimates

$$\hat{\ell}_{t,i} = \begin{cases} \ell_{t,i}/(o_{t,i} + \gamma) & \text{if } \ell_{t,i} \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{E}[\hat{\ell}_{t,i}] = \frac{\ell_{t,i}}{o_{t,i} + \gamma} o_{t,i} + 0(1 - o_{t,i}) = \ell_{t,i} - \ell_{t,i} \frac{\gamma}{o_{t,i} + \gamma} \leq \ell_{t,i}$$

- No mixing!

Analysis of Exp3 algorithms in general

- Evolution of W_{t+1}/W_t

$$\frac{1}{\eta}\log\frac{W_{t+1}}{W_t} = \frac{1}{\eta}\log\left(1-\eta\sum_{i=1}^{N}p_{t,i}\hat{\ell}_{t,i}+\frac{\eta^2}{2}\sum_{i=1}^{N}p_{t,i}(\hat{\ell}_{t,i})^2\right),$$

$$\sum_{i=1}^{N}p_{t,i}\hat{\ell}_{t,i} \le \left[\frac{\log W_t}{\eta}-\frac{\log W_{t+1}}{\eta}\right]+\frac{\eta}{2}\sum_{i=1}^{N}p_{t,i}(\hat{\ell}_{t,i})^2$$

- Taking expectation and summing over time

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}p_{t,i}\hat{\ell}_{t,i}\right]-\mathbb{E}\left[\sum_{t=1}^{T}\hat{\ell}_{t,k}\right] \le \mathbb{E}\left[\frac{\log N}{\eta}\right]+\mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}p_{t,i}(\hat{\ell}_{t,i})^2\right]$$

Regret bound of Exp3-IX

$$\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\hat{\ell}_{t,i}\right]}_{A} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\hat{\ell}_{t,k}\right]}_{B} \le \mathbb{E}\left[\frac{\log N}{\eta}\right] + \underbrace{\mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}(\hat{\ell}_{t,i})^2\right]}_{C}$$

Lower bound of A (using definition of loss estimates)

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\hat{\ell}_{t,i}\right] \ge \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\ell_{t,i}\right] - \mathbb{E}\left[\gamma\sum_{t=1}^{T}\sum_{i=1}^{N} \frac{p_{t,i}}{o_{t,i}+\gamma}\right]$$

Lower bound of B (optimistic loss estimates: $\mathbb{E}[\hat{\ell}] < \mathbb{E}[\ell]$)

$$-\mathbb{E}\left[\sum_{t=1}^{T}\hat{\ell}_{t,k}\right] \ge -\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,k}\right]$$

Upper bound of C (using definition of loss estimates)

$$\mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}(\hat{\ell}_{t,i})^2\right] \le \mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{N} \frac{p_{t,i}}{o_{t,i}+\gamma}\right]$$

Regret bound of Exp3-IX

$$R_T \le \frac{\log N}{\eta} + \left(\frac{\eta}{2} + \gamma\right) \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i=1}^{N} \frac{p_{t,i}}{o_{t,i} + \gamma}\right]$$

$$R_T \approx \mathcal{O}\left(\sqrt{\log N \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i=1}^{N} \frac{p_{t,i}}{o_{t,i} + \gamma}\right]}\right)$$

Regret bound of Exp3-IX

$$R_T \leq \frac{\log N}{\eta} + \left(\frac{\eta}{2} + \gamma\right) \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i=1}^{N} \frac{p_{t,i}}{o_{t,i} + \gamma}\right]$$

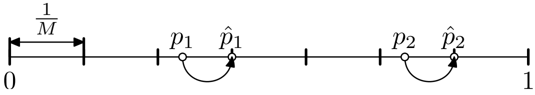
$$R_T \approx \mathcal{O}\left(\sqrt{\log N \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i=1}^{N} \frac{p_{t,i}}{o_{t,i} + \gamma}\right]}\right)$$

Graph lemma

- Graph G with $V(G) = \{1, \ldots, N\}$
- d_i^- – in-degree of vertex i
- α – independence set of G
- Turán's Theorem + induction

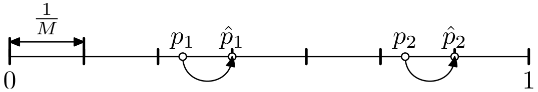
$$\sum_{i=1}^{N} \frac{1}{1 + d_i^-} \leq 2\alpha \log\left(1 + \frac{N}{\alpha}\right)$$

Discretization



$$\sum_{i=1}^{N} \frac{p_{t,i}}{o_{t,i} + \gamma} = \sum_{i=1}^{N} \frac{p_{t,i}}{p_{t,i} + \sum_{j \in N_i^-} p_{t,j} + \gamma} \leq \sum_{i=1}^{N} \frac{\hat{p}_{t,i}}{\hat{p}_{t,i} + \sum_{j \in N_i^-} \hat{p}_{t,j}} + 2$$

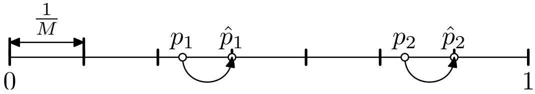
Discretization



$$\sum_{i=1}^{N} \frac{p_{t,i}}{o_{t,i}+\gamma} = \sum_{i=1}^{N} \frac{p_{t,i}}{p_{t,i}+\sum_{j\in N_i^-} p_{t,j}+\gamma} \leq \sum_{i=1}^{N} \frac{\hat{p}_{t,i}}{\hat{p}_{t,i}+\sum_{j\in N_i^-} \hat{p}_{t,j}} + 2$$

Note: we set $M = \lceil N^2/\gamma \rceil$

Discretization



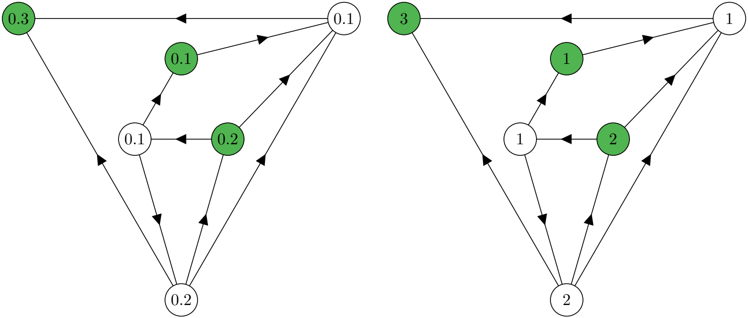
$$\sum_{i=1}^{N} \frac{p_{t,i}}{o_{t,i} + \gamma} = \sum_{i=1}^{N} \frac{p_{t,i}}{p_{t,i} + \sum_{j \in N_i^-} p_{t,j} + \gamma} \leq \sum_{i=1}^{N} \frac{\hat{p}_{t,i}}{\hat{p}_{t,i} + \sum_{j \in N_i^-} \hat{p}_{t,j}} + 2$$

Note: we set $M = \lceil N^2/\gamma \rceil$

$$\sum_{i=1}^{N} \frac{\hat{p}_{t,i}}{\hat{p}_{t,i} + \sum_{j \in N_i^-} \hat{p}_{t,j}}$$

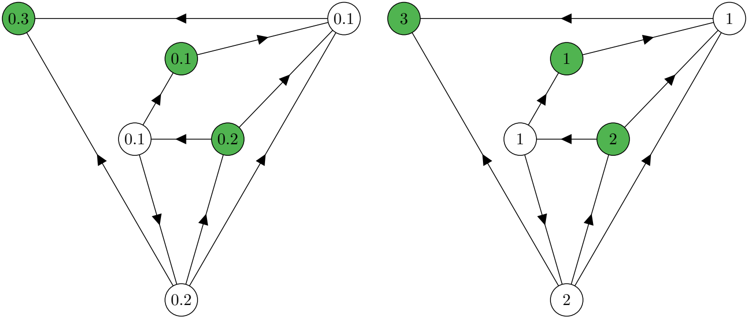
$$\sum_{i=1}^{N} \frac{\hat{p}_{t,i}}{\hat{p}_{t,i} + \sum_{j \in N_i^-} \hat{p}_{t,j}}$$

Example: let $M = 10$



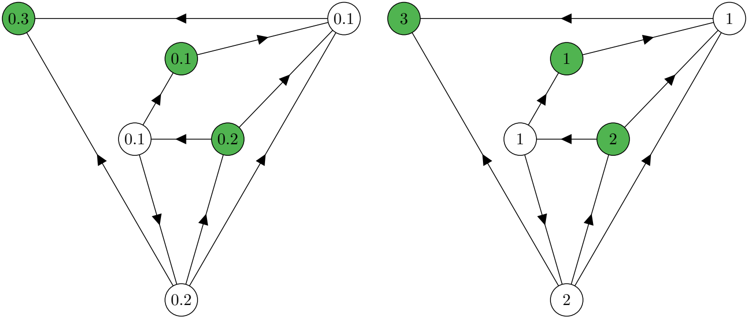
$$\sum_{i=1}^{N} \frac{M\hat{p}_{t,i}}{M\hat{p}_{t,i} + \sum_{j \in N_i^-} M\hat{p}_{t,j}}$$

Example: let $M = 10$



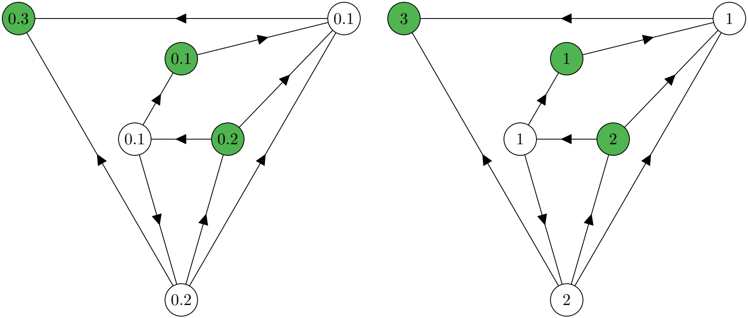
$$\sum_{i=1}^{N} \frac{M\hat{p}_{t,i}}{M\hat{p}_{t,i} + \sum_{j \in N_i^-} M\hat{p}_{t,j}} = \sum_{i=1}^{N} \sum_{k \in C_i} \frac{1}{1 + d_k^-}$$

Example: let $M = 10$



$$\sum_{i=1}^{N} \frac{M\hat{p}_{t,i}}{M\hat{p}_{t,i} + \sum_{j \in N_i^-} M\hat{p}_{t,j}} = \sum_{i=1}^{N} \sum_{k \in C_i} \frac{1}{1 + d_k^-} \leq 2\alpha \log\left(1 + \frac{M+N}{\alpha}\right)$$

Example: let $M = 10$



Exp3-IX regret bound

$$R_T \leq \frac{\log N}{\eta} + \left(\frac{\eta}{2} + \gamma\right) \sum_{t=1}^{T} \mathbb{E}\left[2\alpha_t \log\left(1 + \frac{\lceil N^2/\gamma \rceil + N}{\alpha_t}\right) + 2\right]$$

$$R_T = \widetilde{\mathcal{O}}\left(\sqrt{\overline{\alpha} T \log(N)}\right)$$

Exp3-IX regret bound

$$R_T \leq \frac{\log N}{\eta} + \left(\frac{\eta}{2} + \gamma\right) \sum_{t=1}^{T} \mathbb{E}\left[2\alpha_t \log\left(1 + \frac{\lceil N^2/\gamma \rceil + N}{\alpha_t}\right) + 2\right]$$

$$R_T = \widetilde{\mathcal{O}}\left(\sqrt{\overline{\alpha} T \log(N)}\right)$$

Next step

Exp3-IX regret bound

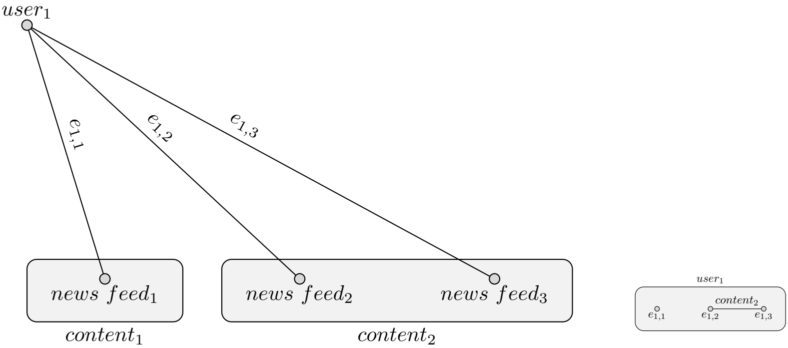
$$R_T \leq \frac{\log N}{\eta} + \left(\frac{\eta}{2} + \gamma\right) \sum_{t=1}^{T} \mathbb{E}\left[2\alpha_t \log\left(1 + \frac{\lceil N^2/\gamma \rceil + N}{\alpha_t}\right) + 2\right]$$

$$R_T = \widetilde{\mathcal{O}}\left(\sqrt{\overline{\alpha} T \log(N)}\right)$$

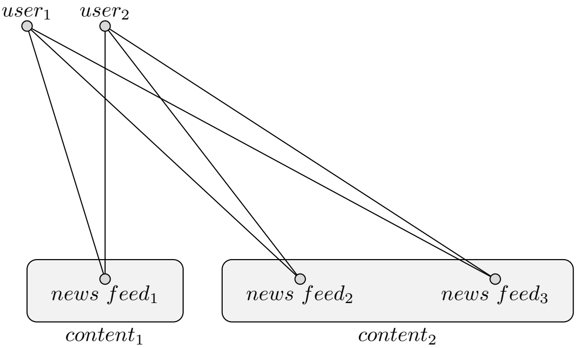
Next step

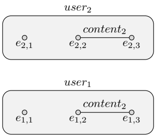
Generalization of the setting

Example

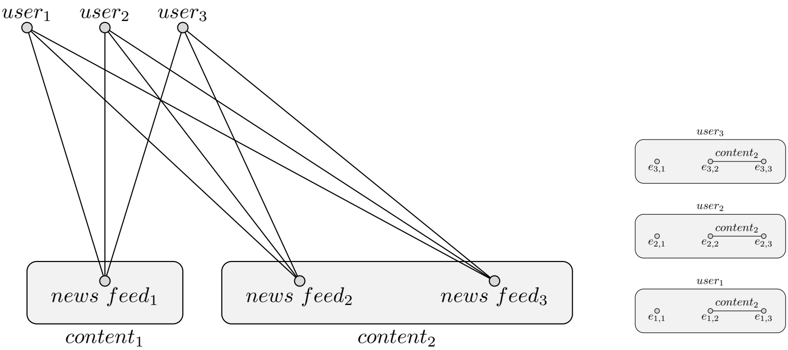

$user_1$
$e_{1,1}$
$e_{1,2}$
$e_{1,3}$
$news\ feed_1$
$news\ feed_2$
$news\ feed_3$
$content_1$
$content_2$
$user_1$
$content_2$
$e_{1,1}$
$e_{1,2}$
$e_{1,3}$

Example

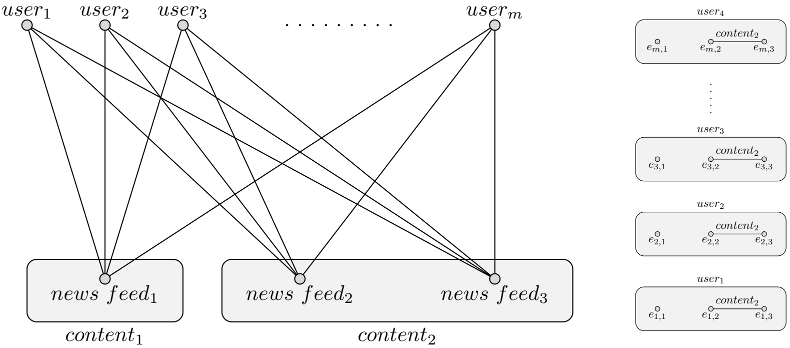

$user_1$
$user_2$
$news\ feed_1$
$news\ feed_2$
$news\ feed_3$
$content_1$
$content_2$


$user_2$
$content_2$
$e_{2,1}$
$e_{2,2}$
$e_{2,3}$
$user_1$
$content_2$
$e_{1,1}$
$e_{1,2}$
$e_{1,3}$

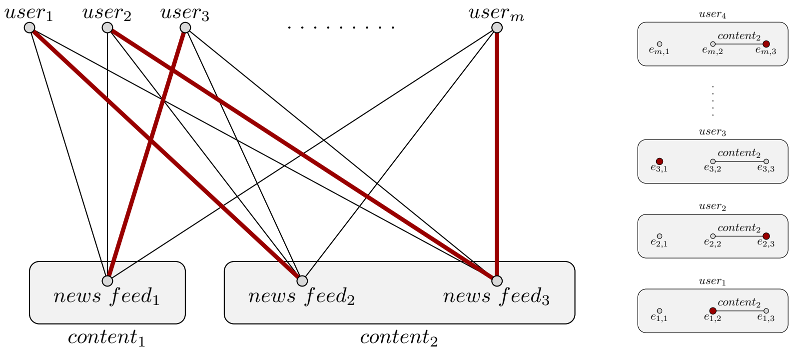
Example


$user_1$
$user_2$
$user_3$
$news\ feed_1$
$content_1$
$news\ feed_2$
$news\ feed_3$
$content_2$
$user_3$
$content_2$
$e_{3,1}$
$e_{3,2}$
$e_{3,3}$
$user_2$
$content_2$
$e_{2,1}$
$e_{2,2}$
$e_{2,3}$
$user_1$
$content_2$
$e_{1,1}$
$e_{1,2}$
$e_{1,3}$

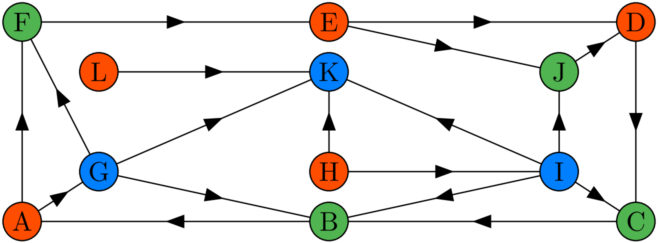
Example


$user_1$
$user_2$
$user_3$
$user_m$
$news\ feed_1$
$news\ feed_2$
$news\ feed_3$
$content_1$
$content_2$
$user_4$
$content_2$
$e_{m,1}$ $e_{m,2}$ $e_{m,3}$
$user_3$
$content_2$
$e_{3,1}$ $e_{3,2}$ $e_{3,3}$
$user_2$
$content_2$
$e_{2,1}$ $e_{2,2}$ $e_{2,3}$
$user_1$
$content_2$
$e_{1,1}$ $e_{1,2}$ $e_{1,3}$

Example



- Play m out of N nodes (combinatorial structure)
- Obtain losses of all played nodes
- Observe losses of all neighbors of played nodes



- Play action $\mathbf{V}_t \in S \subset \{0,1\}^N$, $\|\mathbf{v}\|_1 \leq m$ fro all $\mathbf{v} \in S$
- Obtain losses $\mathbf{V}_t^\top \boldsymbol{\ell}_t$
- Observe additional losses according to the graph

FPL-IX algorithm

- **Draw perturbation** $Z_{t,i} \sim \text{Exp}(1)$ for all $i \in [N]$

- **Play "the best" action** $\mathbf{V}_t$ according to total loss estimate $\widehat{\mathbf{L}}_{t-1}$ and perturbation $\mathbf{Z}_t$

$$\mathbf{V}_t = \arg\min_{\mathbf{v} \in \mathcal{S}} \mathbf{v}^\top \left(\eta_t \widehat{\mathbf{L}}_{t-1} - \mathbf{Z}_t \right)$$

- **Compute loss estimates**

$$\hat{\ell}_{t,i} = \ell_{t,i} K_{t,i} \mathbb{1}\{\ell_{t,i} \text{ is observed}\}$$

- $K_{t,i}$: geometric random variable with

$$\mathbb{E}\left[K_{t,i}\right] = \frac{1}{o_{t,i} + (1 - o_{t,i})\gamma}$$

FPL-IX - regret bound

$$R_T = \widetilde{\mathcal{O}}\left(m^{3/2}\sqrt{\sum_{t=1}^{T}\alpha_t}\right) = \widetilde{\mathcal{O}}\left(m^{3/2}\sqrt{\overline{\alpha}T}\right)$$

Conclusion

- Introduction of **I**mplicit e**X**ploration idea
- **New algorithm for simple actions**
  - Using implicit exploration idea
  - Same regret bound as previous algorithm
  - No need to know graph before an action is played
  - Computationally efficient
- **New combinatorial setting with side observations**
- **Algorithm for combinatorial setting**
  - Using implicit exploration idea
  - No need to know graph before an action is played
  - Computationally efficient

Thank you!

Tomáš Kocák
tomas.kocak@inria.fr
sequel.lille.inria.fr