Ordonnancement d'objets par bandits unimodaux sur des graphes paramétriques

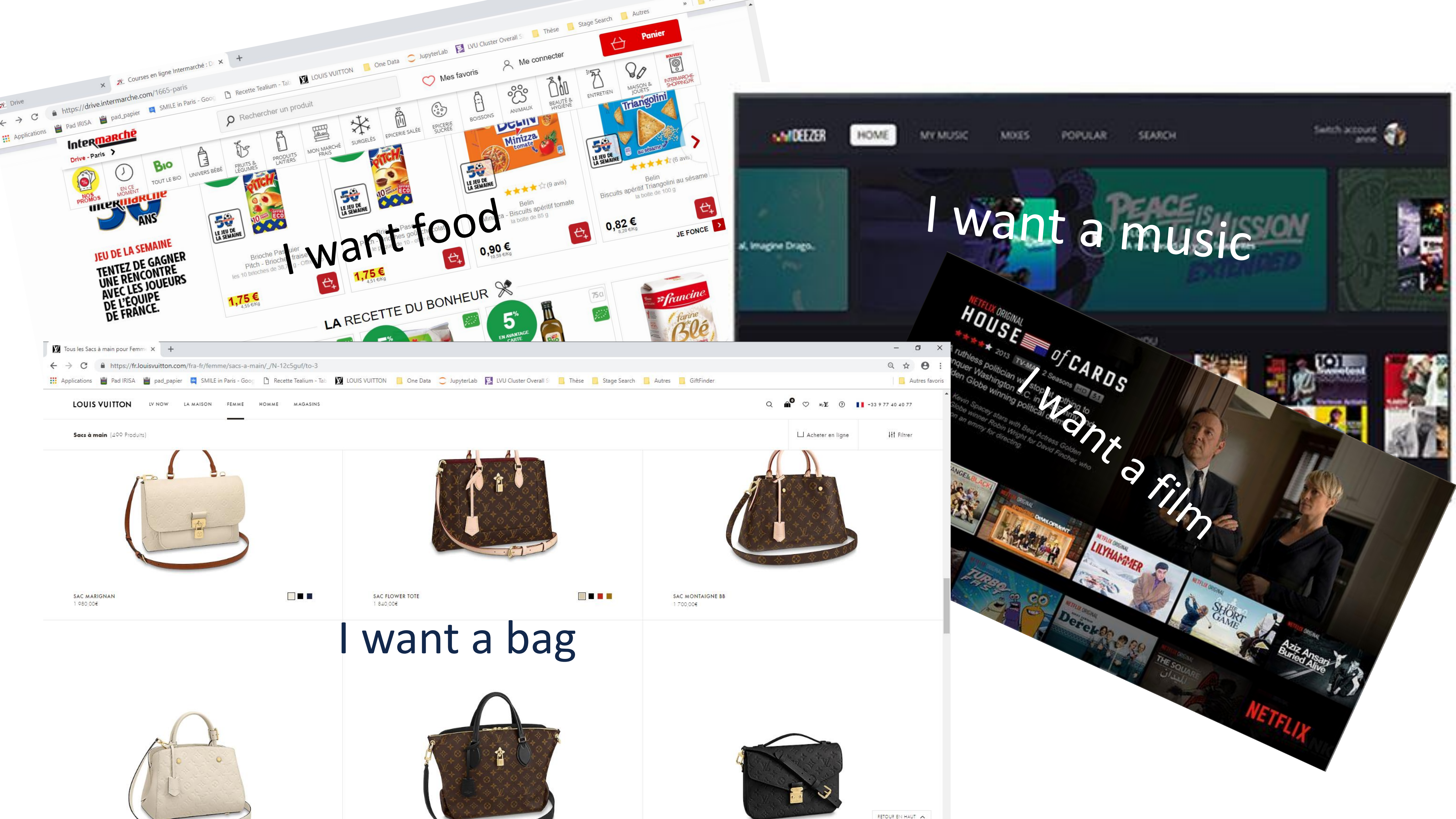
(CAp 2021 – ICML 2021)

LOUIS VUITTON

UMR IRISA

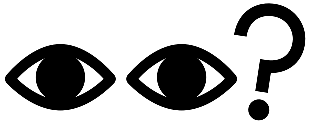
CREST
CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

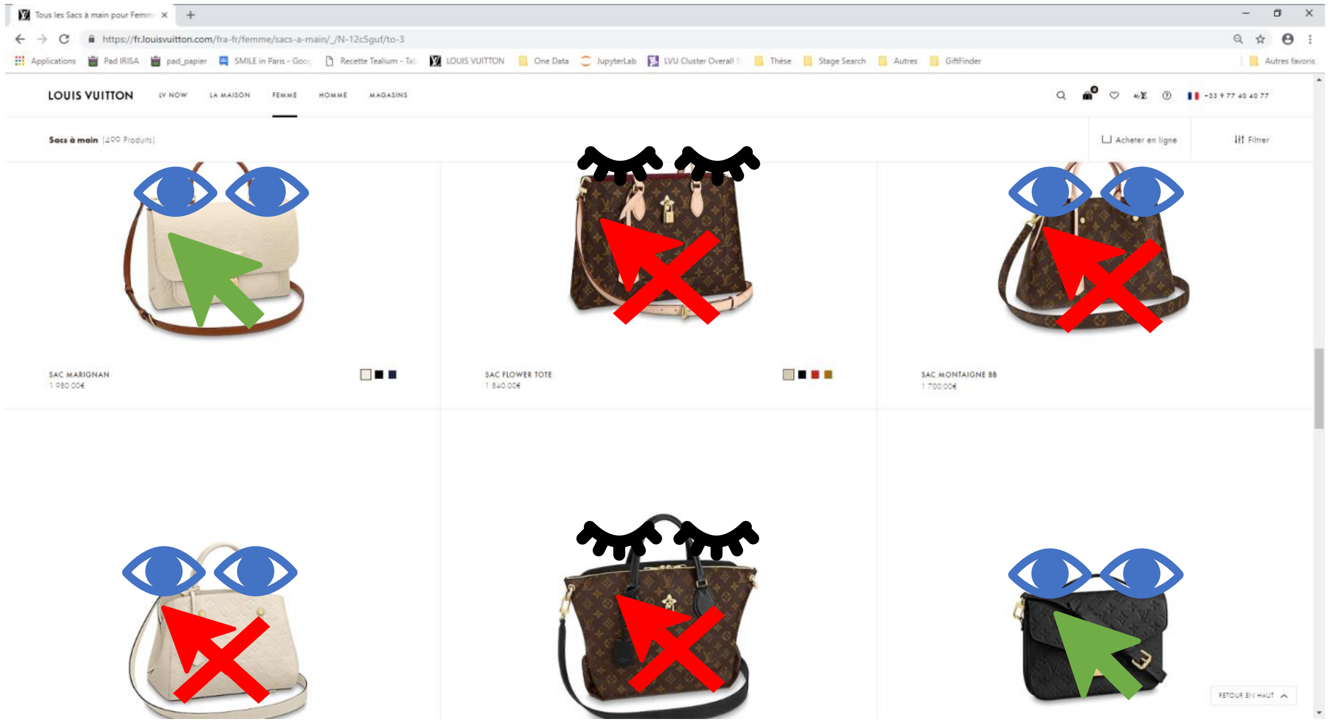
Camille-Sovanneary Gauthier, Romaric Gaudel, Elisa Fromont, Boammani Aser Lompo


I want food
I want a music
I want a film
I want a bag

Multiple way to read

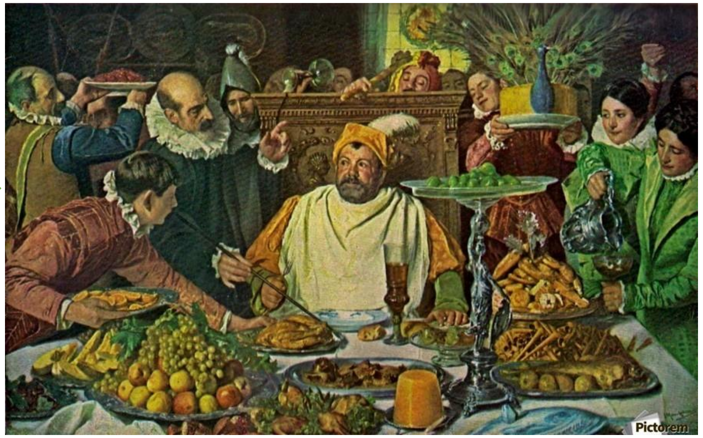
Have you looked at my reco ?

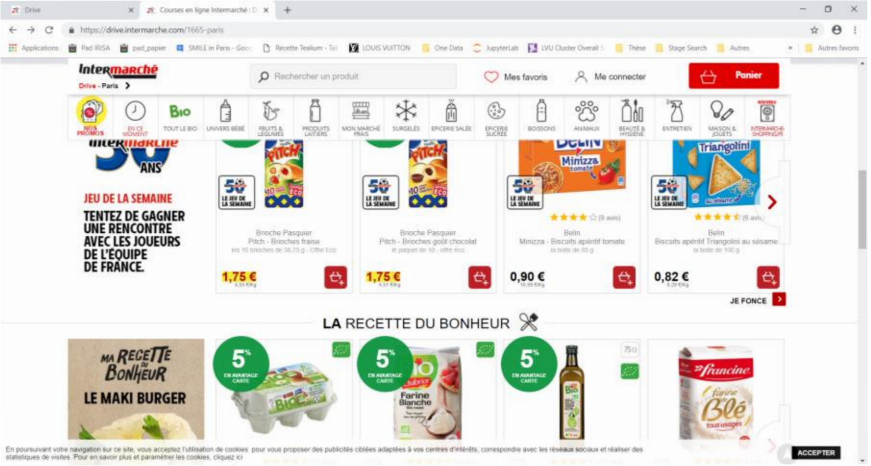




Rewards have to be defined

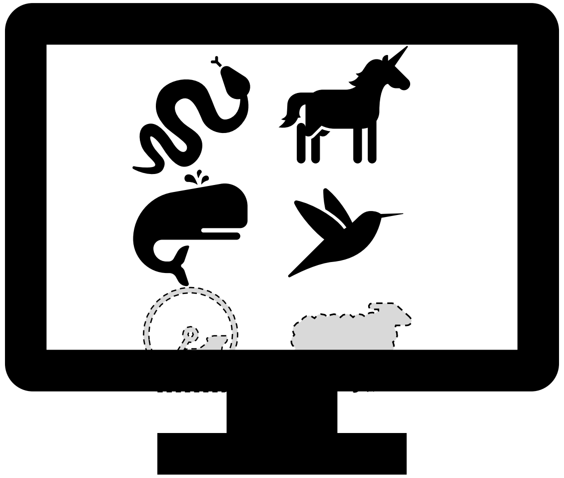
Multi clicks and multi propositions: giving THE best or maximise click rate





Recommendations are complex tools

Multi propositions



Importance of items' order

Partial attention

Reward to interpret

State of the art

Position-based model

Other models: Cascading Model [2]; Dependent Click Model...

<u>Position-based model </u>[1] : user click is motivated independently by the position and the item

Setting: {L items ; K positions} , at time t :

Notation: $\boldsymbol{\kappa_l}$ ***view's probability*** of position l $\in [K]$;

$\boldsymbol{\theta_i}$ ***click's probability*** of the item i $\in [L]$.

$$Y_k(t) \sim Ber(\kappa_k)$$ [user consideration of position k]

$$X_i(t) \sim Ber(\theta_i)$$ [user feedback on item i]

$$\boldsymbol{Z_{k,i}}(t) \sim X_i(t) * Y_k(t)$$ [***the observation***]

In other word:

observed → $Z_{k,i}(t) \sim Ber(\theta_i * \kappa_k)$ ← unknown

From Information Retrieval to Bandits

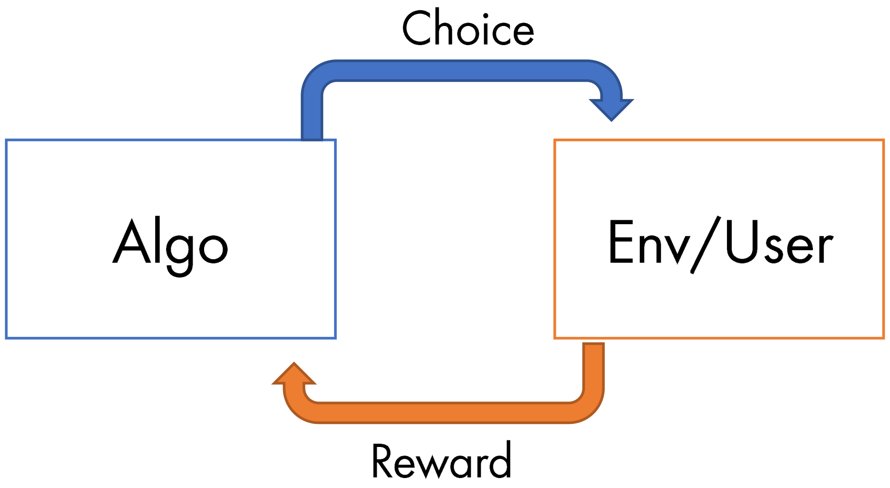
Aim: give the right list of answers to the right person.

From Information Retrieval (up to ~2010; WWW, WSDM, SIGIR,...) ...

- From collected data
- Find the right model
- Infer the parameters of the model

...to Bandits Theory (starting in 2015; ICML, NIPS,...)

- Account for the data collection process
- Infer parameters AND handle parameters « uncertainty »
- => Exploration/Exploitation dilemma



Main Bandit approaches

Main Articles

- Multiple-play bandits in the Position-based model by P. Lagrée, C. Vernade, and O. Cappe, 2016, NeurIPS [3]:
  - Multiple Play (PBM)
  - Several approaches: TS, UCB, Pie
  - Assume κ known

- Position-based multiple-play bandit problem with unknown position bias by J. Komiyama, J. Honda, and A. Takeda, 2017, NeurIPS [4]:
  - Multiple Play (PBM)
  - Permutation exploration / Non convex optimization

- Bandit Algorithm for Both Unknown Best Position and Best Item Display on Web Pages by C.-S. Gauthier, R. Gaudel and E. Fromont, 2021, IDA [5]:
  - Multiple Play (PBM)
  - No assumption on parameters

Main Bandit approaches

Main Articles

Open fields:

- Linear bandit: contextual representation for numerous products
- Diversity
- Lack for PBM

Us:

Lack for provable approaches on PBM with κ unknown

- M… C. Vernade, and O. Cappe, 2016, NeurIPS [3]:
- Pos… on bias by J. Ko…
- Ba… Display on Web Pages by C.-S. Gauthier, R. Gaudel and E. Fromont, 2021, IDA [5]:
  - No assumption on parameters

Our Contribution

New ***bandit algorithm***, GRAB, learns online a graph of permutations (of recommendations)

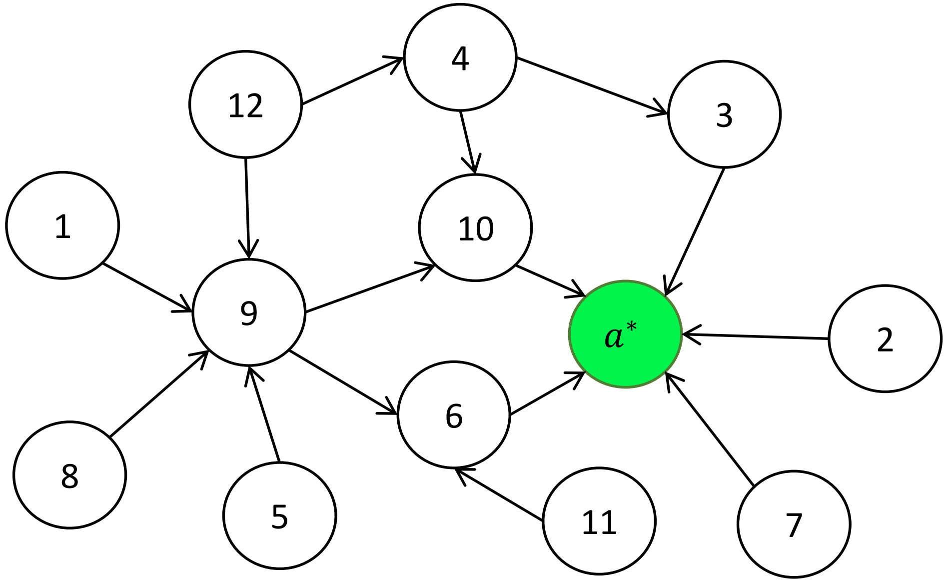
- simple to implement and efficient in terms of computation time;
- handles the PBM bandit setting without any knowledge on the impact of positions (contrarily to many competitors);
- empirically exhibits a regret on par with other theoretically proven algorithms on both artificial and real datasets.
- ***$O(L/\Delta \log T)$ regret upper-bound*** (see cumulative regret below)

Unimodality

Unimodality (Definition [6]) :

Let A be a ***set of arms*** and $(\nu_a)_{a\in A}$ a set of rewards distribution of respective expectations $(\mu_a)_{a\in A}$. G=(V,E) be a graph with vertices V = A and edges E. The set of expected rewards $(\mu_a)_{a\in A}$ is unimodal w.r.t G, if and only if :

1) there is a unique best arm, $argmax_a \mu_a = a^*$
2) for any $a \neq a^*$, there exists a path ($a^0 = a, a^1, \ldots, a^n = a^*$) and , for all i ∈ $[n], \mu_{a^i} > \mu_{a^{i-1}}$ and $a^i \in N(a^{i-1})$

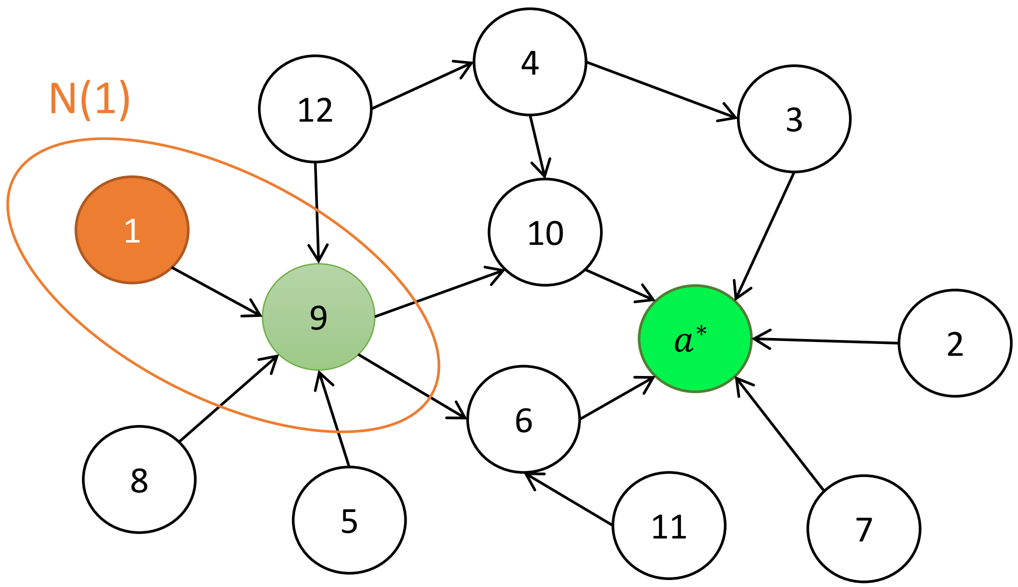


Unimodality

Unimodality (Definition [6]) :

Let A be a ***set of arms*** and $(\nu_a)_{a \in A}$ a set of rewards distribution of respective expectations $(\mu_a)_{a \in A}$. G=(V,E) be a graph with vertices V = A and edges E. The set of expected rewards $(\mu_a)_{a \in A}$ is unimodal w.r.t G, if and only if :

1) there is a unique best arm, $argmax_a \mu_a = a^*$
2) for any $a \neq a^*$, there exists a path ($a^0 = a$, a^1,… , $a^n = a^*$) and , for all i ∈ $[n]$, $\mu_{a^i} > \mu_{a^{i-1}}$ and $a^i \in N(a^{i-1})$

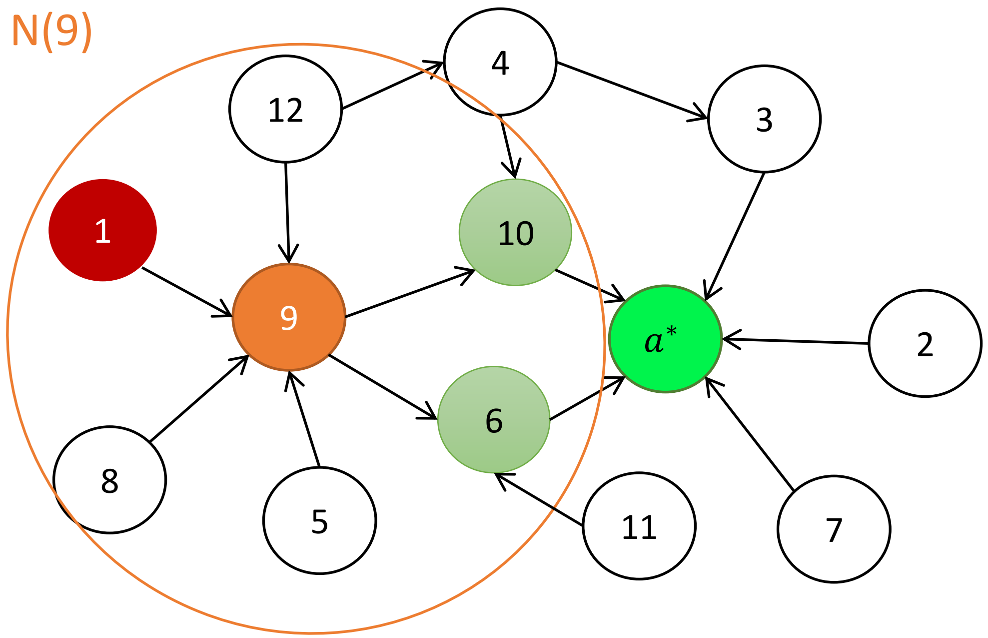


Unimodality

Unimodality (Definition [6]) :

Let A be a ***set of arms*** and $(\nu_a)_{a \in A}$ a set of rewards distribution of respective expectations $(\mu_a)_{a \in A}$. G=(V,E) be a graph with vertices V = A and edges E. The set of expected rewards $(\mu_a)_{a \in A}$ is unimodal w.r.t G, if and only if :

1) there is a unique best arm, $argmax_a \mu_a = a^*$
2) for any $a \neq a^*$, there exists a path ($a^0 = a$, a^1,... , $a^n = a^*$) and , for all i ∈ $[n]$, $\mu_{a^i} > \mu_{a^{i-1}}$ and $a^i \in N(a^{i-1})$

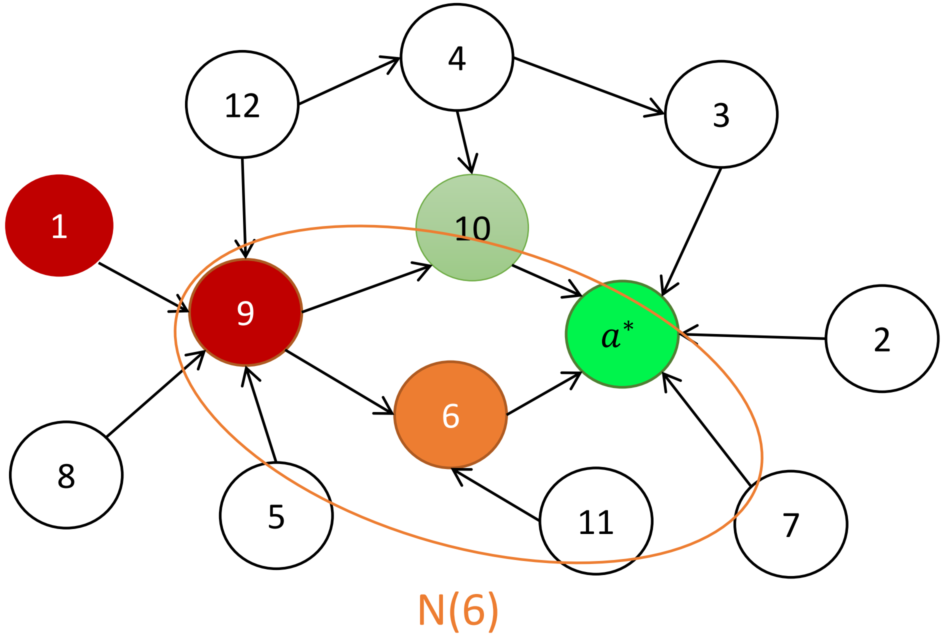


Unimodality

Unimodality (Definition [6]) :

Let A be a ***set of arms*** and $(\nu_a)_{a \in A}$ a set of rewards distribution of respective expectations $(\mu_a)_{a \in A}$. G=(V,E) be a graph with vertices V = A and edges E. The set of expected rewards $(\mu_a)_{a \in A}$ is unimodal w.r.t G, if and only if :

1) there is a unique best arm, $argmax_a \mu_a = a^*$
2) for any $a \neq a^*$, there exists a path ($a^0 = a, a^1, \dots, a^n = a^*$) and , for all i ∈ $[n], \mu_{a^i} > \mu_{a^{i-1}}$ and $a^i \in N(a^{i-1})$

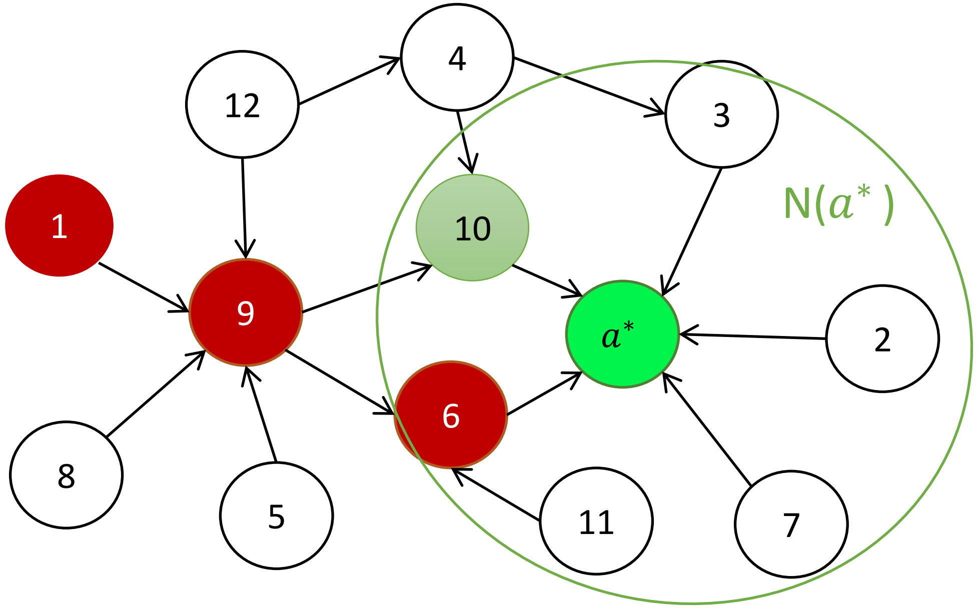


Unimodality

Unimodality (Definition [6]) :

Let A be a ***set of arms*** and $(\nu_a)_{a\in A}$ a set of rewards distribution of respective expectations $(\mu_a)_{a\in A}$. G=(V,E) be a graph with vertices V = A and edges E. The set of expected rewards $(\mu_a)_{a\in A}$ is unimodal w.r.t G, if and only if :

1) there is a unique best arm, $argmax_a \mu_a = a^*$
2) for any $a \neq a^*$, there exists a path ($a^0 = a$, a^1,... , $a^n = a^*$) and , for all i ∈ $[n]$, $\mu_{a^i} > \mu_{a^{i-1}}$ and $a^i \in N(a^{i-1})$



Notes :
- Unimodal Bandit OSUB [6] can jump from node to node
- Its regret R(T) depends on γ:= max degree of the graph. $R(T) = O(\frac{\gamma}{\Delta}\log T)$

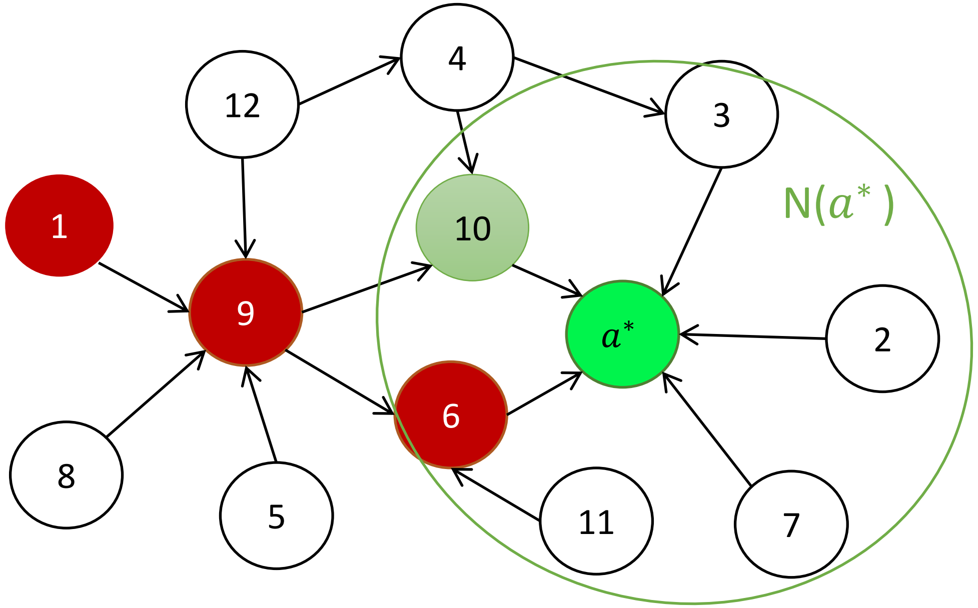
Unimodality

Unimodality (Definition [6]) :

Let A be a ***set of arms*** and $(\nu_a)_{a\in A}$ a set of rewards distribution of respective expectations $(\mu_a)_{a\in A}$. G=(V,E) be a graph with vertices V = A and edges E. The set of expected rewards $(\mu_a)_{a\in A}$ is unimodal w.r.t G, if and only if :

1) there is a unique best arm, $argmax_a \mu_a = a^*$

2) for any $a \neq a^*$, there exists a path ($a^0 = a, a^1,\ldots, a^n = a^*$) and , for all i ∈ $[n], \mu_{a^i} > \mu_{a^{i-1}}$ and $a^i \in N(a^{i-1})$

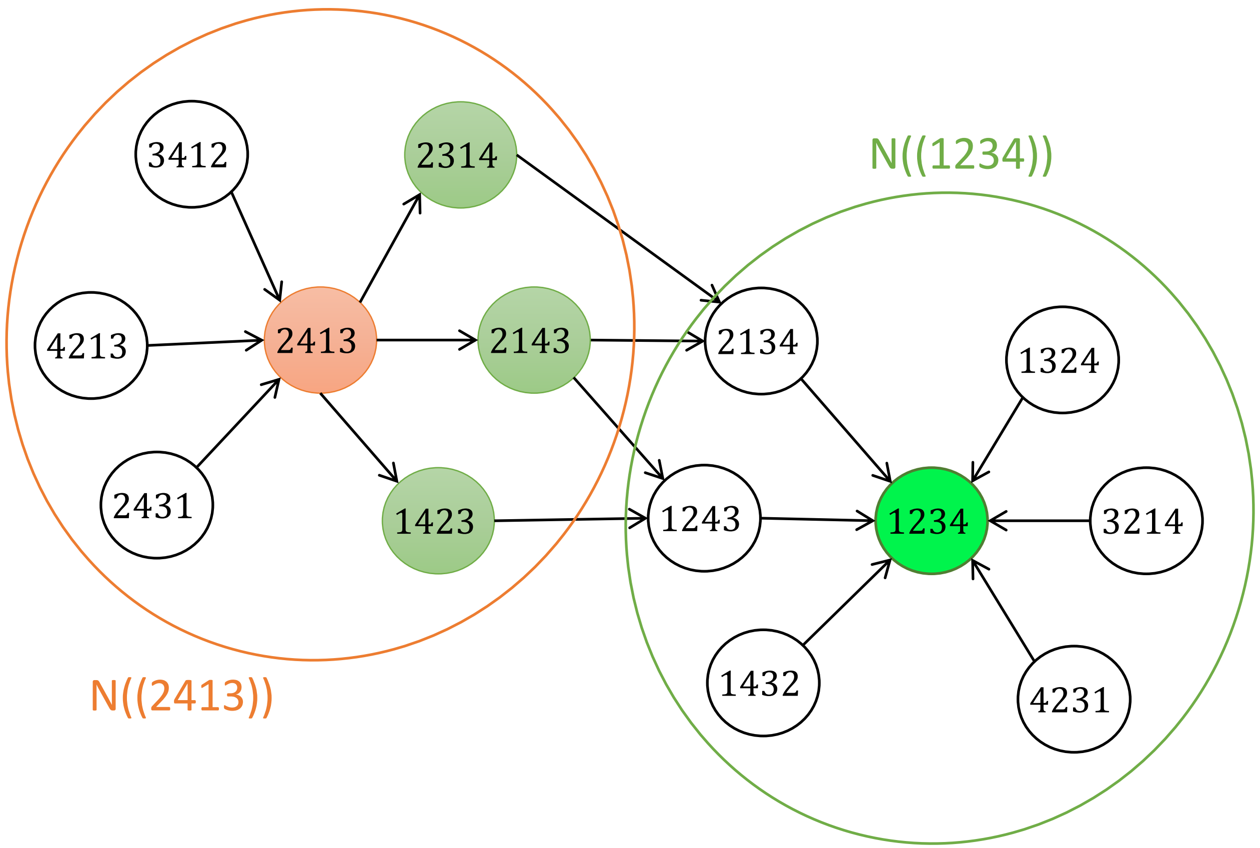


Notes :
- Unimodal Band… can jump from node to nod…
- Its regret … ax degree of the g…

Our nodes are lists of items
=> Find the good structure
to keep unimodality but
reduce the graph's degree

Our approach

Unimodal bandit for PBM recommendation : S-GRAB



We explore this graph in order to get the higher μ.

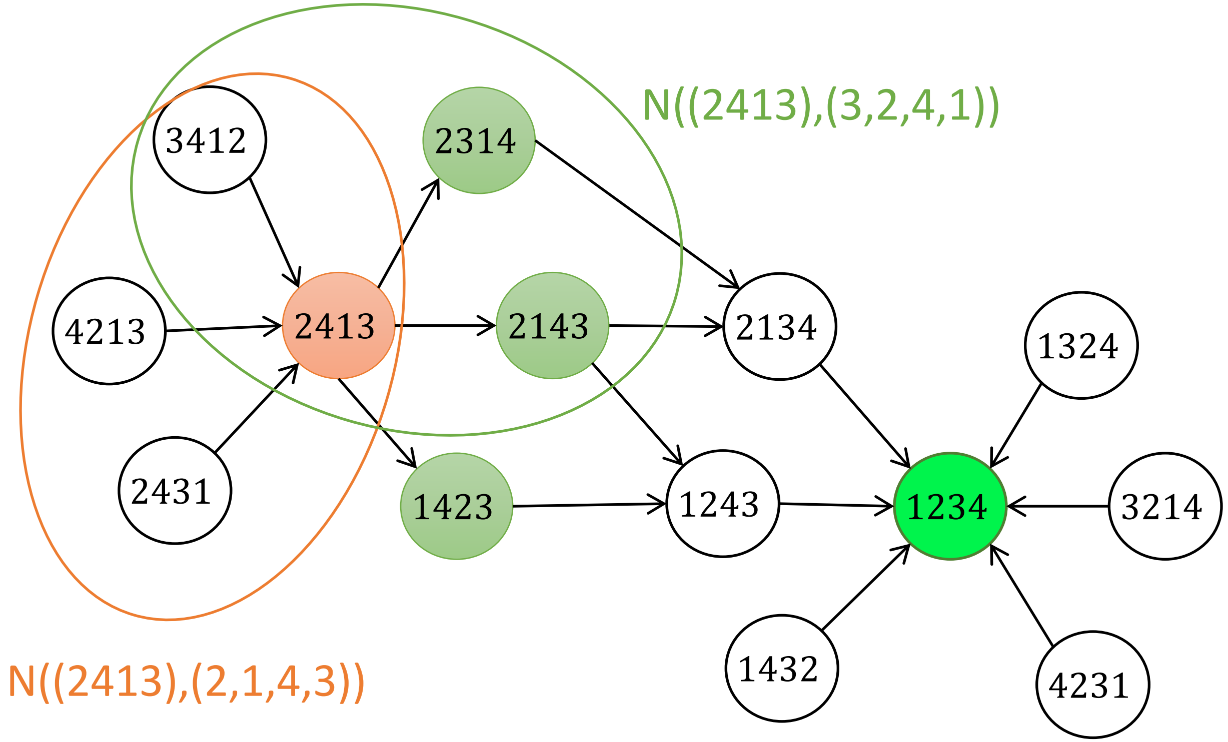
The expected reward μ increases when you exchange two items such that the most attractive one gets in the most looked position.

We have $\mu_{[2413]} - \mu_{[2143]} = (\kappa_2 - \kappa_3)(\theta_4 - \theta_1)$

With $\mathrm{N}(\boldsymbol{a}) = \{\boldsymbol{a} \circ (l, l') : l, l' \in [L]^2, l > l'\}$

We get $\mathrm{R(T)} = \mathrm{O}(\frac{LK}{\Delta}\mathrm{logT})$ (same as TopRank [8])

Unimodal bandit for PBM recommendation :GRAB

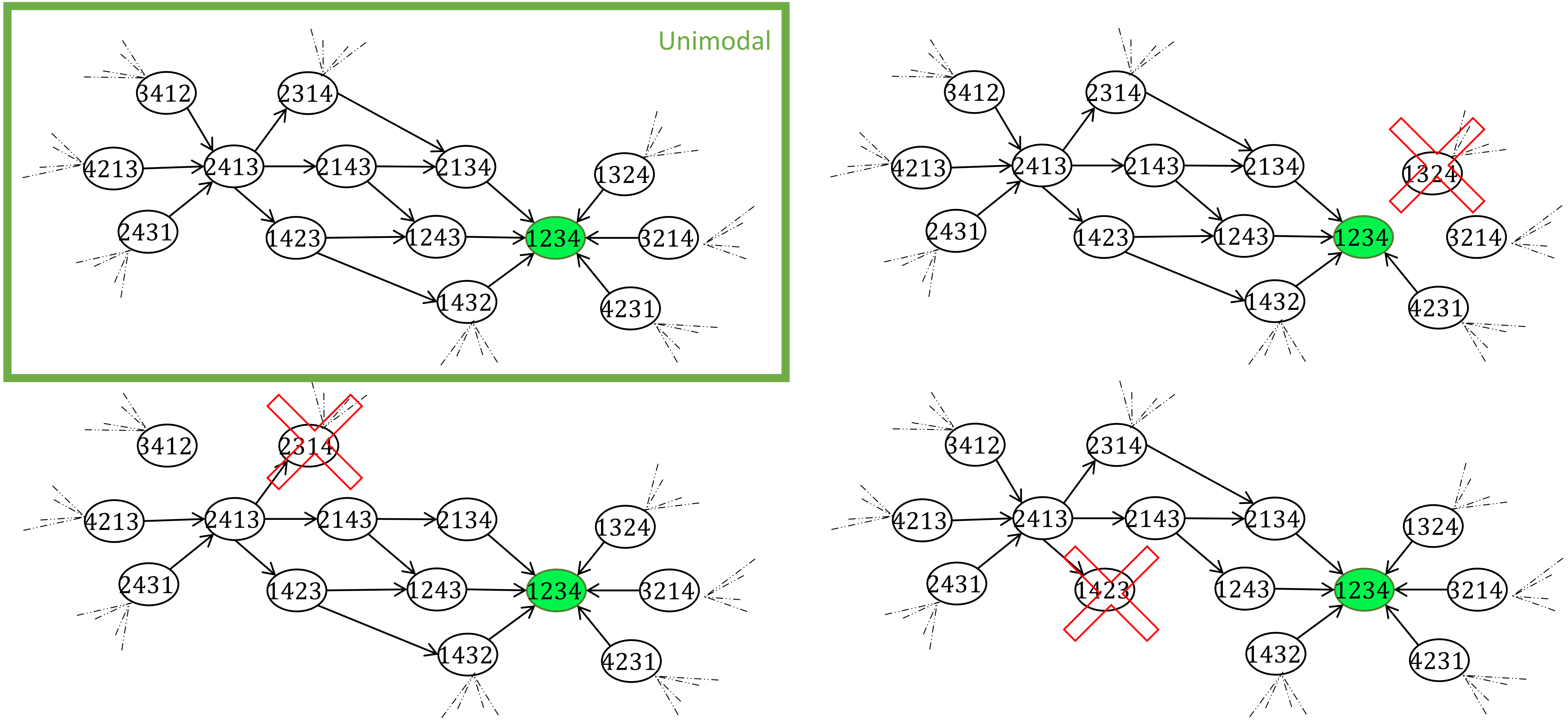


With the right order on positions, you may limit the number of transpositions explored.

with $\mathrm{N}(\boldsymbol{a}, \pi\ ) = \{\boldsymbol{a} \circ (\pi_{a_k}, \pi_{a_{k+1}}) \colon k\ \in [K-1]\}$

Unimodal bandit for PBM recommendation :GRAB

We get $\mathbf{R(T) = O(\frac{L}{\Delta}\log T)}$



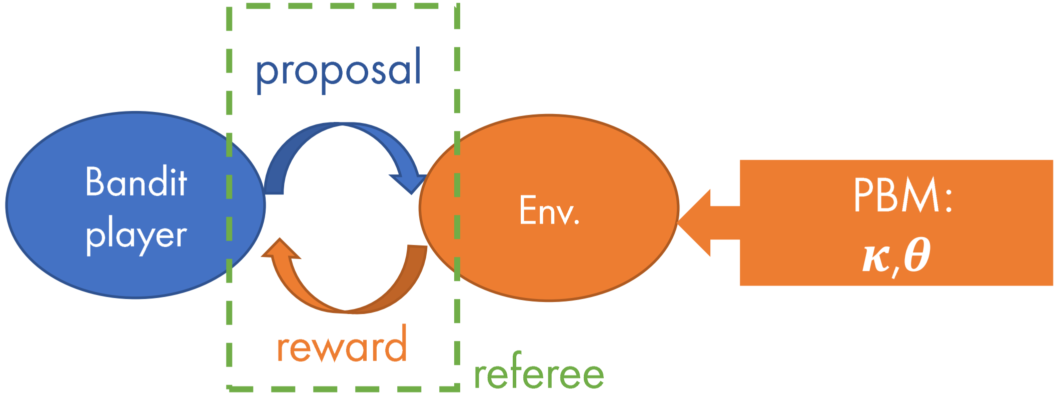
Best regret upper bound

Algorithm	Handled Behavioral Model	Regret
GRAB	PBM	$O(\frac{L}{\Delta}\log(T))$
CombUCB1 [9]	PBM	$O(\frac{LK^2}{\Delta}\log(T))$
PBM-PIE [3]	PBM with κ known	$O(\frac{(L-K)}{\Delta}\log(T))$
PMED-Hinge [4]	PBM with $\kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_K$	$O(c^*(\boldsymbol{\theta}, \boldsymbol{\kappa})\log(T))$
TopRank [8]	PBM with $\kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_K$	$O(\frac{LK}{\Delta}\log(T))$
OSUB [6]	Unimodal	$O(\frac{\gamma}{\Delta}\log(T))$
PB-MHB [5]	PBM	$\emptyset$

Experimental Setting

Opponents:

- GRAB
- S-GRAB
- ε-Greedy
- PMED [4]
- PB_MHB [5]
- TopRank [8]
- KL-CombUCB [9]



Measure:

Cumulative pseudo regret

Data:

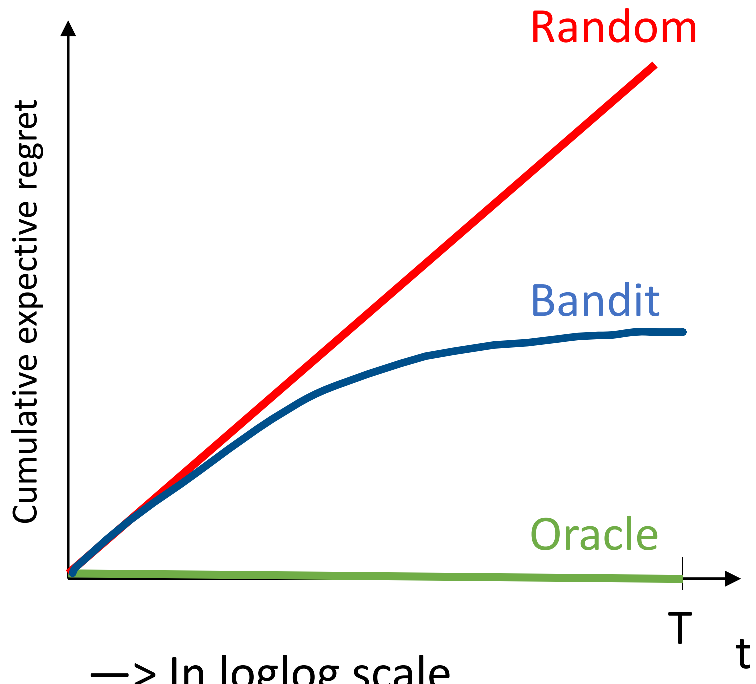
- purely simulated ($\boldsymbol{\kappa}, \boldsymbol{\theta}$ set by us)
- $\boldsymbol{\kappa}, \boldsymbol{\theta}$ inferred Yandex's logs => 10 selected queries [7] (Pyclic module)

Measure

Cumulative pseudo regret:

$$R_T = \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{E}[\boldsymbol{r}_k(t)|\boldsymbol{i}_k^*] - \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{E}[\boldsymbol{r}_k(t)|i_k\ (t)]$$

$$R_T = \mu^* T - \sum_{t=1}^{T}\sum_{k=1}^{K} \theta_{i_k(t)}\kappa_k$$



—> In loglog scale

Want to minimize it

Best provable algorithm

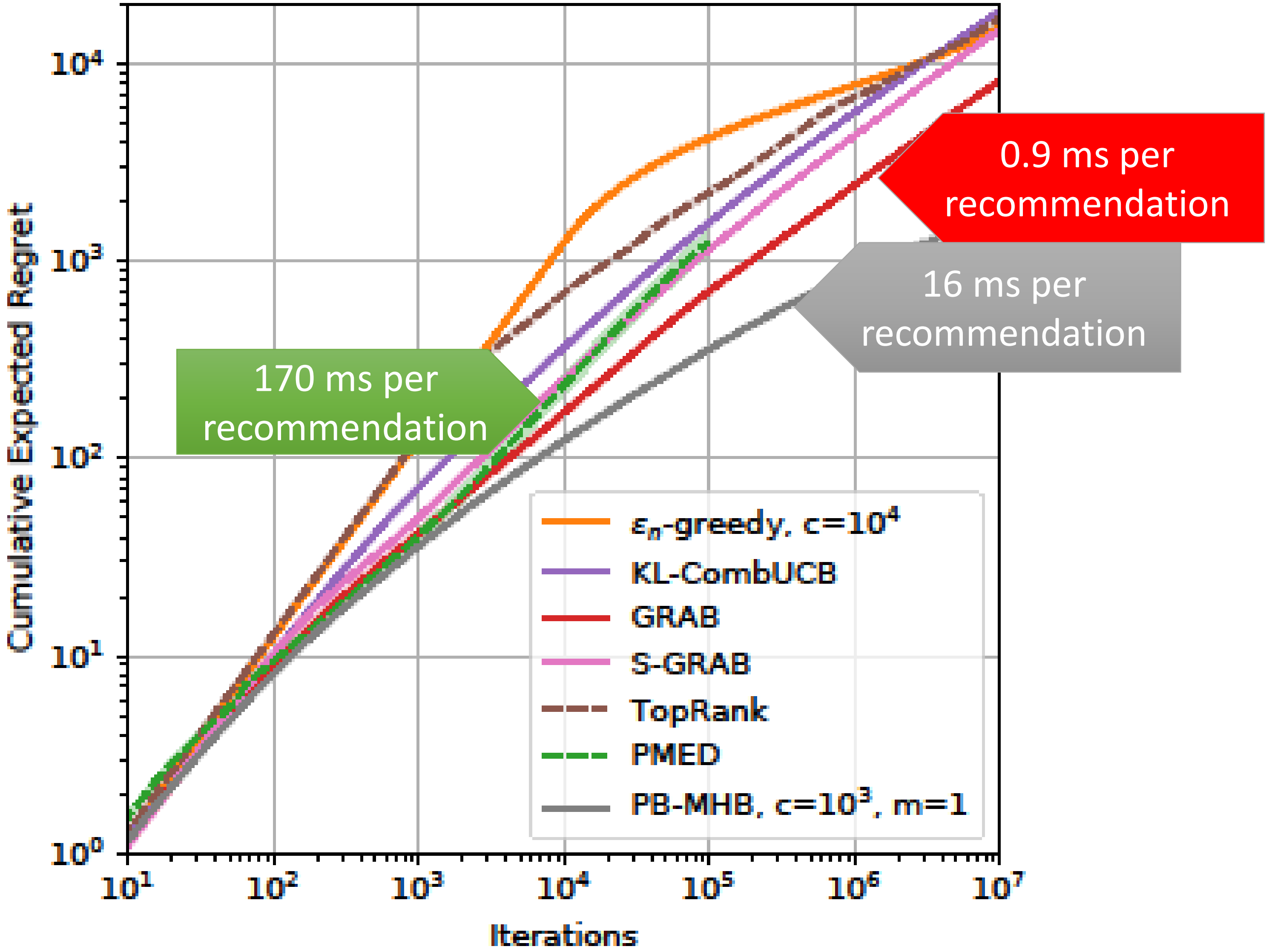
<u>Environment:</u>

$\boldsymbol{\kappa}, \boldsymbol{\theta}$ inferred from Yandex

<u>Referee:</u>

Average on 200 runs of 10^7 trials with L=10; K=5

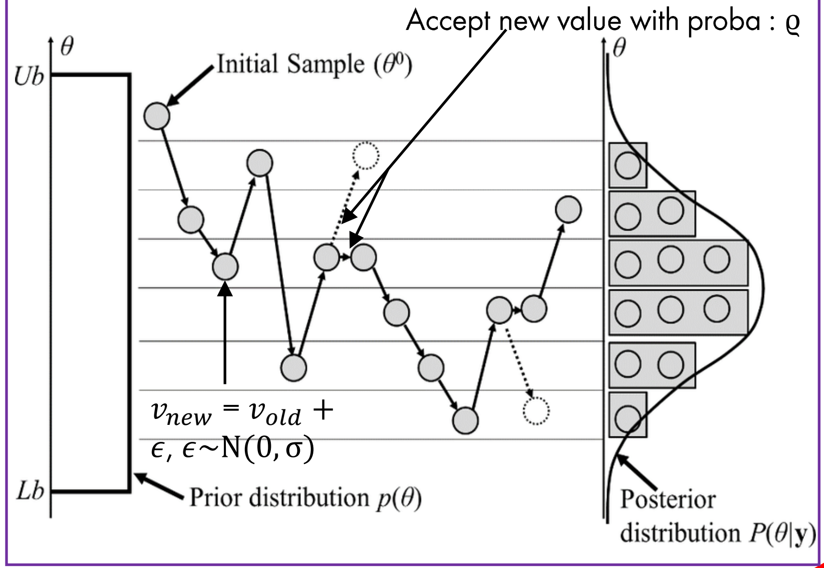
(= 20 games on each of the 10 queries selected)

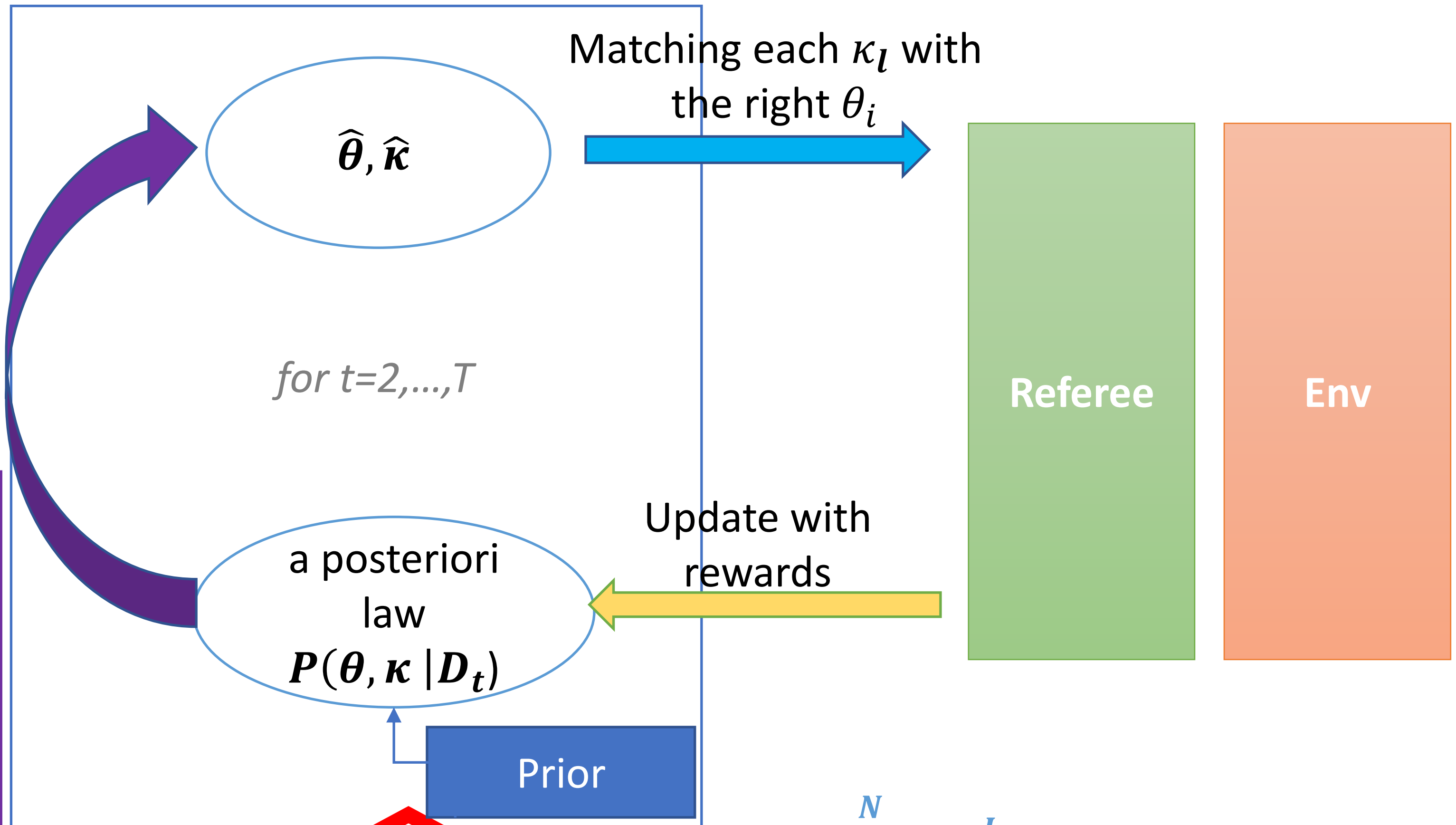


PB-MHB [5]

THOMPSON SAMPLING

Draw thanks to **Metropolis Hasting**





$$P(\theta, \kappa \,|D_t) \propto \prod_{i=1}^{N} \prod_{l=1}^{L} (\theta_i \kappa_l)^{S_{i,l}(t-1)} (1 - \theta_i \kappa_l)^{F_{i,l}(t-1)}$$

Not usual law (Beta product) =>
Not easy to draw from this law

PB-MHB [5]

Based on Thompson sampling.

As rewards come from (1) with a bayesian point of view, we can:

- Set a uniform prior on $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$
- Update through a Beta likelihood
- Target the following posterior:

$$\boldsymbol{P}(\boldsymbol{\theta}, \boldsymbol{\kappa} \mid \boldsymbol{D_t}) \propto \prod_{i=1}^{N} \prod_{l=1}^{L} (\boldsymbol{\theta_i \kappa_l})^{S_{i,l}(t-1)} (\mathbf{1} - \boldsymbol{\theta_i \kappa_l})^{F_{i,l}(t-1)}, \quad (2)$$

with D_t, collected data = list of items and reward at each time from 0 to t-1,

$S_{i,l}(t-1) = \sum_{s=1}^{t-1} \mathbb{I}(i_l(s) = i)\mathbb{I}(r_l(s) = 1)$ = number of time i has been clicked while being displayed in position l;

$F_{i,l}(t-1) = \sum_{s=1}^{t-1} \mathbb{I}(i_l(s) = i)\mathbb{I}(r_l(s) = 0)$ = number of time i has been clicked while being displayed in position l;

- Matching parameters

Split to draw

Split the formula (independence + Gibbs):

$$P(\theta_i|\boldsymbol{\kappa}, \boldsymbol{D}) = \alpha \prod_{l=1}^{L} \theta_i^{S_{i,l}(t-1)} (1-\theta_i\kappa_l)^{F_{i,l}(t-1)}, \text{ for } \theta_i \text{ in } \boldsymbol{\theta} \quad (3)$$

$$P(\kappa_l|\boldsymbol{\theta}, \boldsymbol{D}) = \beta \prod_{i=1}^{M} \kappa_l^{S_{i,l}(t-1)} (1-\theta_i\kappa_l)^{F_{i,l}(t-1)}, \text{ for } \kappa_l \text{ in } \boldsymbol{\kappa} \quad (4)$$

Draw thanks to Monte Carlo Markov Chain (Metropolis-Hasting with Gaussian random walk kernel per parameter)

Take home

Setting adopted: List recommendation with multiple rewards with full unknown PBM setting

Our approach : Transpose PBM into a unimodal graph

Our (empirical) ***result***: Better regret with less information.

Possible search areas:

- Extend to other behavioural setting
- Contextual Bandits
- ...

Questions ?

Camille-Sovanneary GAUTHIER : camille-sovanneary.gauthier@louisvuitton.com

Romaric GAUDEL : romaric.gaudel@ensai.fr

Elisa FROMONT : elisa.fromont@irisa.fr

References

[1] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: Estimating the click-through rate for new ads, 2007, WWW’07

[2] B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan. Cascading bandits: Learning to rank in the cascade model, 2015, ICML

[3] P. Lagrée, C. Vernade, and O. Cappe, Multiple-play bandits in the Position-based model, 2016, NeurIPS

[4] J. Komiyama, J. Honda, A. Takeda : Position-based multiple-play bandit problem with unknown position bias, 2017, NeurIPS

[5] C.-S. Gauthier, R. Gaudel and E. Fromont, Bandit Algorithm for Both Unknown Best Position and Best Item Display on Web Pages, 2021, IDA

[6] Combes, R. and Proutière, A. Unimodal bandits: Regret lower bounds and optimal algorithms, 2014, ICML’14,

[7] T. Lattimore, B. Kveton, S. Li, C. Szepesvari: TopRank: A practical algorithm for online stochastic ranking, 2018, NeurIPS

[8] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models, 2008, in Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM.

[9] Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits, B. Kveton, Z. Wen, A. Ashkan and C. Szepesvari, 2015, AISTATS'15

[10] A. Chuklin, I. Markov, M. de Rijke. Click Models for Web Search. Morgan & Claypool Publishers, 2015.

[11] J.Komiyama, J Honda and H. Nakagawa 2015. Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays, 2015, ICML’15.

[12] Trinh C., Kaufmann E., Vernade C. and Combes R. Solving Bernoulli Rank-One Bandits with Unimodal Thompson Sampling, arXiv:1912.03074v1 [stat.ML] 6 Dec 2019