Learning to Rank under Multinomial Logit Choice

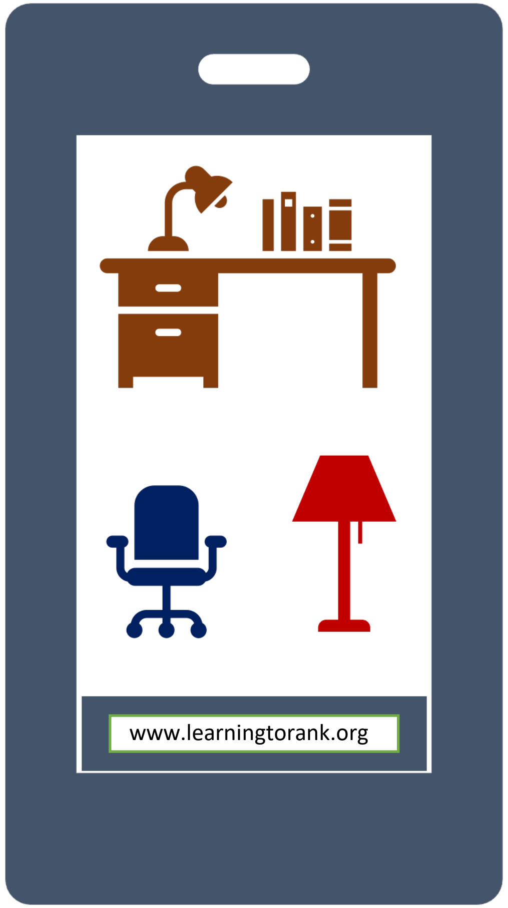
James A Grant (he/him) **-** Lancaster University
Joint work with David S Leslie

EcoSta 2022 - Sunday 5th June

j.grant@lancaster.ac.uk - @james_a_grant

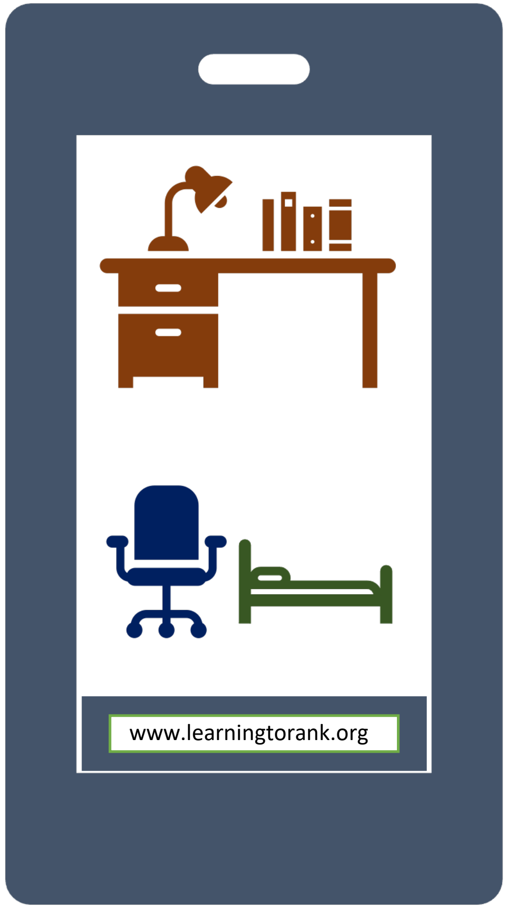
High Level Idea

Determining an optimal selection and positioning of website content to maximise the number of clicked items over time.



High Level Idea

Determining an optimal **selection** and positioning of website content to maximise the number of clicked items over time.



High Level Idea

Determining an optimal **selection** and **positioning** of website content to maximise the number of clicked items over time.

High Level Idea

Determining an optimal **selection** and **positioning** of website content to maximise the number of **clicked items over time**.

High Level Idea

Determining an optimal **selection** and **positioning** of website content to maximise the number of **clicked items over time**.

- Novelty in a click model which allows simultaneous consideration with prominence weighting

(Online) Learning to Rank

Ranking content for user satisfaction has roots in information retrieval

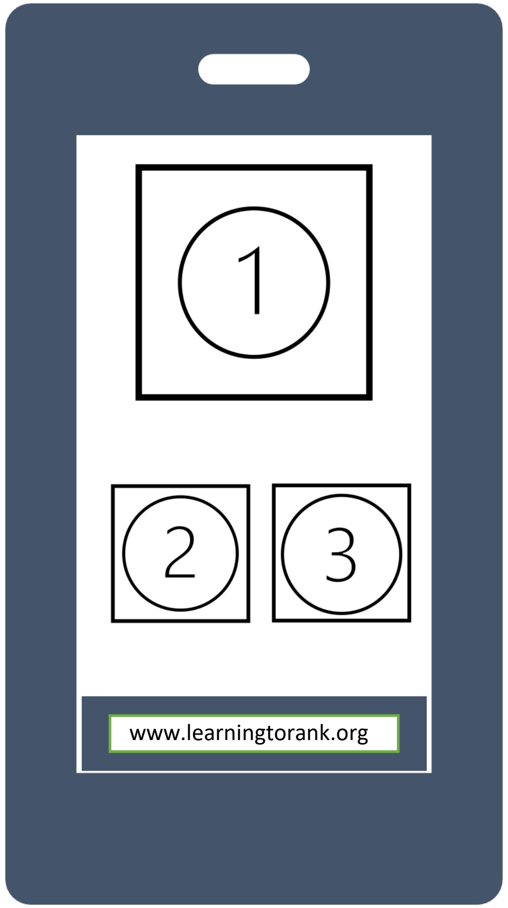
- Search engine optimisation
- Rank by perceived relevance

More recently: learn the optimal ranking through sequential recommendations and feedback (*learning to rank*)

- Underlying attractiveness unknown
- Display a set of items
- Observe click or no click (click model differentiates approach – Chuklin et al. (2015))
- Update estimates of item attractiveness and repeat

Click Model

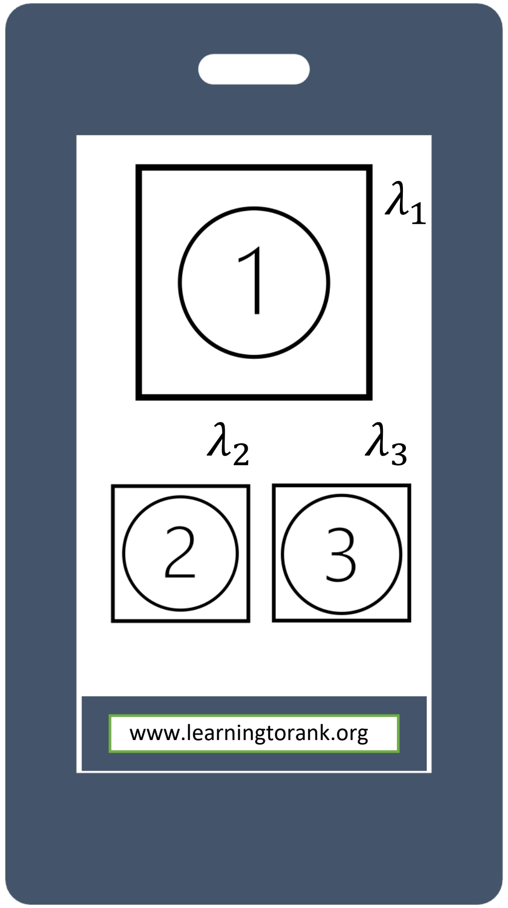
We formulate a multinomial logit choice model with position effects.



Click Model

We formulate a multinomial logit choice model with position effects.

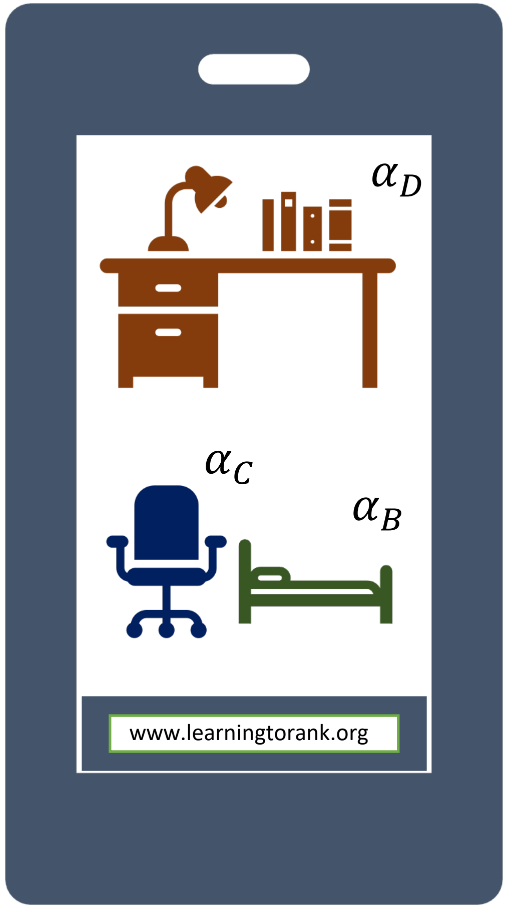
Each position i has an associated weight $\lambda_i \in (0,1]$,



Click Model

We formulate a multinomial logit choice model with position effects.

Each position i has an associated weight $\lambda_i \in (0,1]$, and item j has an attractiveness $\alpha_j \in (0,1]$.



Click Model

We formulate a multinomial logit choice model with position effects.

Each position i has an associated weight $\lambda_i \in (0,1]$, and item j has an attractiveness $\alpha_j \in (0,1]$.

A no-click option is endowed with dummy weights $\lambda_0 = 1, \alpha_0 = 1$.

Click Model

A click indicator C is modelled as a random variable on $\{0,1,2,..,K\}$ with distribution dependent on the ordered item list $\boldsymbol{a} = (a_1, \dots, a_K)$.

$$P(C = k \mid \boldsymbol{a}) = \frac{\alpha_{a_k}\lambda_k}{1 + \sum_{j=1}^{K} \alpha_{a_j} \lambda_j}$$

$$P(C = 0 \mid \boldsymbol{a}) = \frac{1}{1 + \sum_{j=1}^{K} \alpha_{a_j} \lambda_j}$$

Learning to Rank with Multinomial Logit Choice

Aim is to design an effective algorithm to select lists of items $\boldsymbol{a} = (a_1, \dots, a_K)$ from a set $\mathcal{A}$ without initial knowledge of $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$.

Objective to maximise expected clicks over T sets of recommendations – or equivalently minimise **regret**

$$\min_{\boldsymbol{a}_1,\dots,\boldsymbol{a}_T \subset \mathcal{A}} \left(\sum_{t=1}^{T} \max_{\boldsymbol{a} \in \mathcal{A}} P(C_t \neq 0 \mid \boldsymbol{a}) - P(C_t \neq 0 \mid \boldsymbol{a}_t) \right)$$

Requires a balance between **exploration** and **exploitation**.

Balancing Exploration and Exploitation

Optimism in the face of uncertainty is widely deployed technique for online learning

Underlying optimisation problem: $\max_{a \in \mathcal{A}} E(Reward(a))$

Balancing Exploration and Exploitation

Optimism in the face of uncertainty is widely deployed technique for online learning

Underlying optimisation problem: $\max_{a \in \mathcal{A}} \mu_a$ select largest mean

Balancing Exploration and Exploitation

Optimism in the face of uncertainty is widely deployed technique for online learning

Underlying optimisation problem: $\max_{a \in \mathcal{A}} \mu_a$ select largest mean

At time t naïve approach is: $\max_{a \in \mathcal{A}} \hat{\mu}_{a,t}$ largest sample mean

Balancing Exploration and Exploitation

Optimism in the face of uncertainty is widely deployed technique for online learning

Underlying optimisation problem: $\max_{a \in \mathcal{A}} \mu_a$ select largest mean

At time t naïve approach is: $\max_{a \in \mathcal{A}} \hat{\mu}_{a,t}$ largest sample mean

Can go badly wrong if initial data is atypical – under-exploration

Balancing Exploration and Exploitation

Optimism in the face of uncertainty is widely deployed technique for online learning

Underlying optimisation problem: $\max_{a \in \mathcal{A}} \mu_a$ select largest mean

At time t naïve approach is: $\max_{a \in \mathcal{A}} \hat{\mu}_{a,t}$ largest sample mean

Can go badly wrong if initial data is atypical – under-exploration

Optimistic approach is: $\max_{a \in \mathcal{A}} \hat{\mu}_{a,t} + B_{a,t}$ largest **optimistic** value

Optimism in the Face of Uncertainty

Optimistic approach is: $\max_{a \in \mathcal{A}} \hat{\mu}_{a,t} + B_{a,t}$ largest **optimistic** value

Key decision is the form of the $B_{a,t}$ term.

Optimism in the Face of Uncertainty

Optimistic approach is: $\max_{a \in \mathcal{A}} \hat{\mu}_{a,t} + B_{a,t}$ largest **optimistic** value

Key decision is the form of the $B_{a,t}$ term.

When rewards are independent across actions we derive $B_{a,t}$ from simple finite-time concentration results, e.g. Chernoff-Hoeffding bound.

Optimism in the Face of Uncertainty

Optimistic approach is: $\max_{a \in \mathcal{A}} \hat{\mu}_{a,t} + B_{a,t}$ largest **optimistic** value

Key decision is the form of the $B_{a,t}$ term.

When rewards are independent across actions we derive $B_{a,t}$ from simple finite-time concentration results, e.g. Chernoff-Hoeffding bound.

- $P\left(|\hat{\mu}_{a,t} - \mu| > B\right) \leq \exp\left(-\frac{2B^2}{\sum_s \mathbb{I}\{A_s = a\}}\right)$
- Choosing $B_{a,t} = \sqrt{2\log(t) / \sum_s \mathbb{I}\{A_s = a\}}$ guarantees optimal regret scaling (details not for today).

OFU for Multinomial Logit

Design of a suitable optimistic approach is more complex where estimates of unknown parameters are not just sample averages.

Likelihood for our MNL choice model is

$$\mathcal{L}(C_1, \dots, C_t; \alpha_1, \dots, \alpha_J, \lambda_1, \dots, \lambda_K) = \prod_{s=1}^{t} \left(\frac{1}{1 + \sum_{k=1}^{K} \alpha_{a_t(k)} \lambda_k} \right)^{\mathbb{I}\{C_t = 0\}} \prod_{k=1}^{K} \left(\frac{\alpha_{a_t(k)} \lambda_k}{1 + \sum_{k=1}^{K} \alpha_{a_t(k)} \lambda_k} \right)^{\mathbb{I}\{C_t = k\}}$$

Sufficiently complex that we have no closed form for MLEs and estimate them via an EM algorithm - finite-time concentration inequalities are elusive.

OFU for Multinomial Logit

We can learn λ parameters relatively easily since each slot is used. When $J > K$ we want to ensure appropriate exploration of items. We want a (non-asymptotic) result like

$$P\left(\left|\hat{\alpha}_{j,t}^{MLE} - \alpha_j\right| > B \mid \boldsymbol{a_1}, \dots, \boldsymbol{a_t}\right) \leq f(B, \boldsymbol{a_1}, \dots, \boldsymbol{a_t}).$$

OFU for Multinomial Logit

We can learn λ parameters relatively easily since each slot is used. When $J > K$ we want to ensure appropriate exploration of items. We want a (non-asymptotic) result like

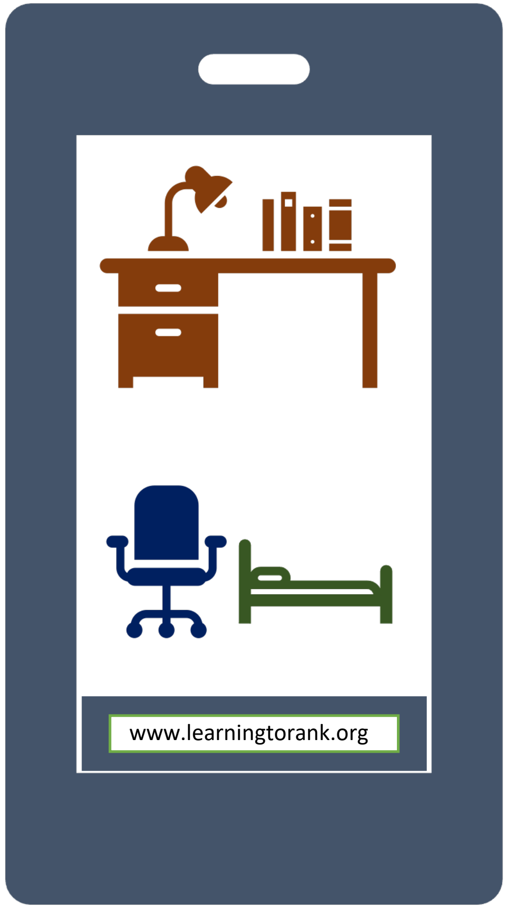
$$P\left(\left|\hat{\alpha}_{j,t}^{MLE} - \alpha_j\right| > B \mid \boldsymbol{a_1}, \dots, \boldsymbol{a_t}\right) \leq f(B, \boldsymbol{a_1}, \dots, \boldsymbol{a_t}).$$

Combine two ideas:

- Batched decision-making (Agrawal et al. (2017, 2019))
- Functional concentration inequalities (Bobkov and Ledoux (1998), and Joulin and Privault (2004))

Batched Decision Making

In the setting where $\lambda_1 = \cdots = \lambda_K = 1$, a pattern of repeatedly displaying the same item set until a no-click is observed is used (Agrawal et al. 2017, 2019).

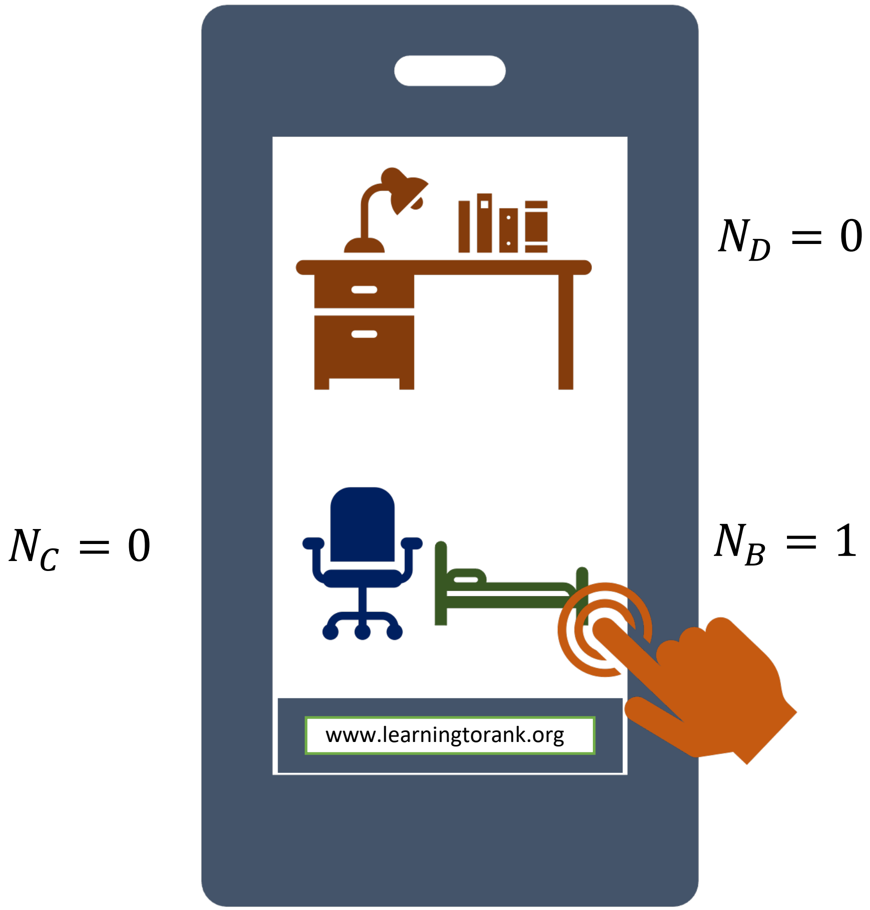


Batched Decision Making

In the setting where $\lambda_1 = \cdots = \lambda_K = 1$, a pattern of repeatedly displaying the same item set until a no-click is observed is used (Agrawal et al. 2017, 2019).

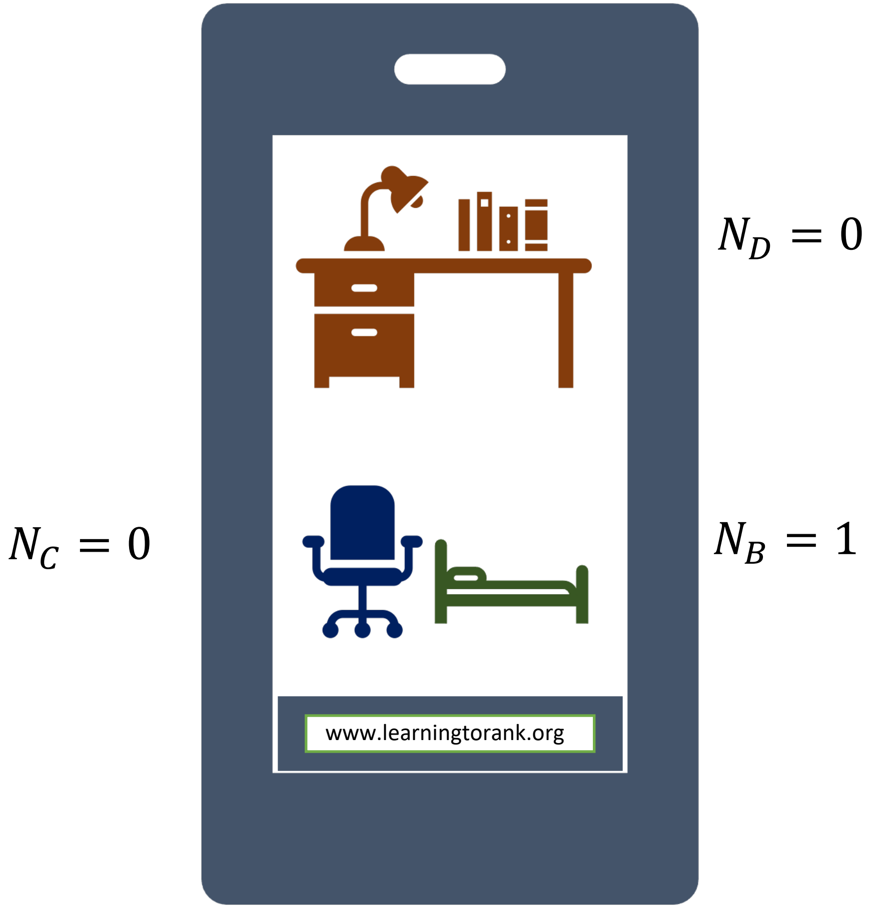
Batched Decision Making

In the setting where $\lambda_1 = \cdots = \lambda_K = 1$, a pattern of repeatedly displaying the same item set until a no-click is observed is used (Agrawal et al. 2017, 2019).



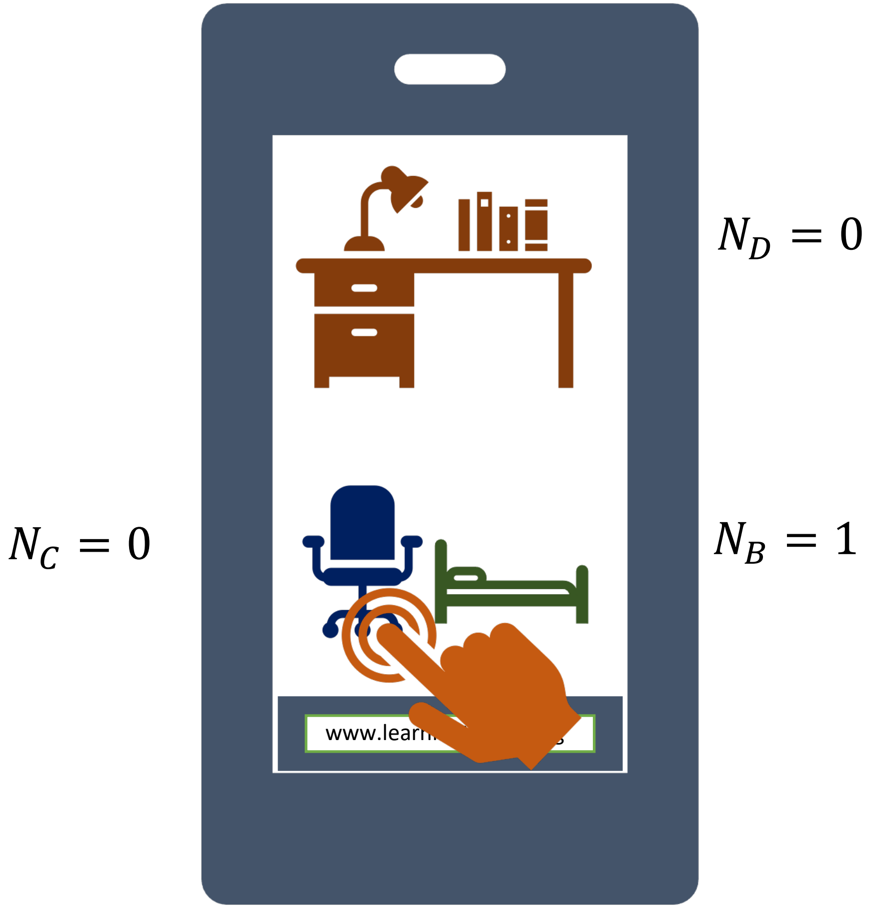
Batched Decision Making

In the setting where $\lambda_1 = \cdots = \lambda_K = 1$, a pattern of repeatedly displaying the same item set until a no-click is observed is used (Agrawal et al. 2017, 2019).



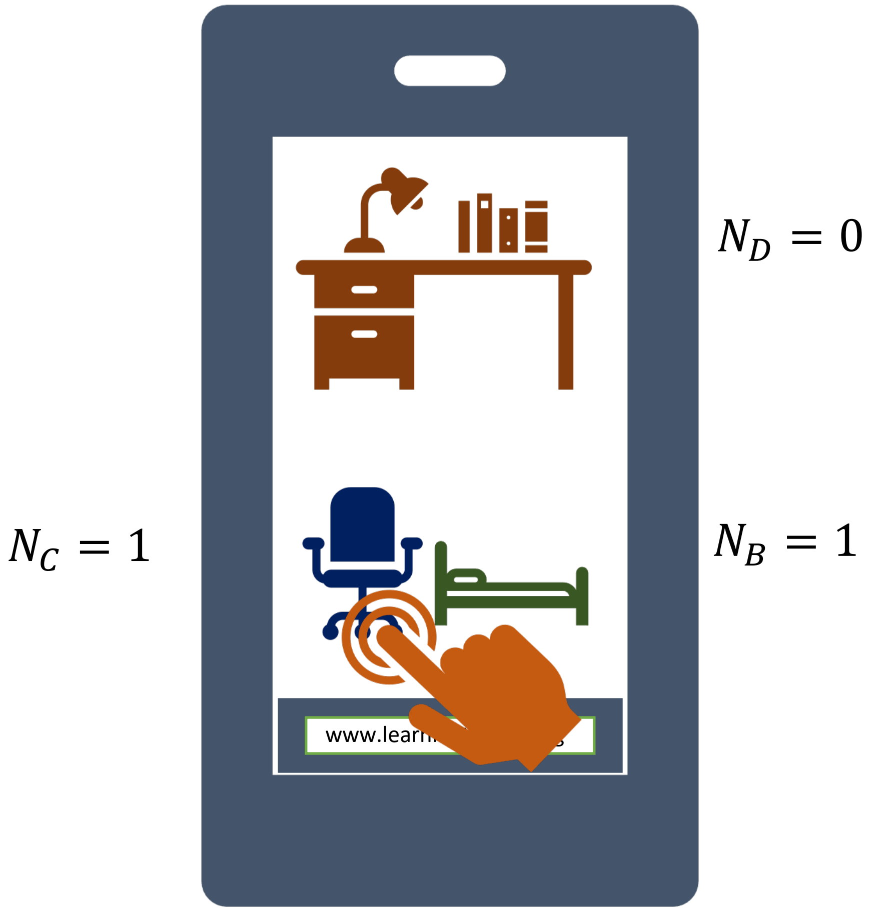
Batched Decision Making

In the setting where $\lambda_1 = \cdots = \lambda_K = 1$, a pattern of repeatedly displaying the same item set until a no-click is observed is used (Agrawal et al. 2017, 2019).



Batched Decision Making

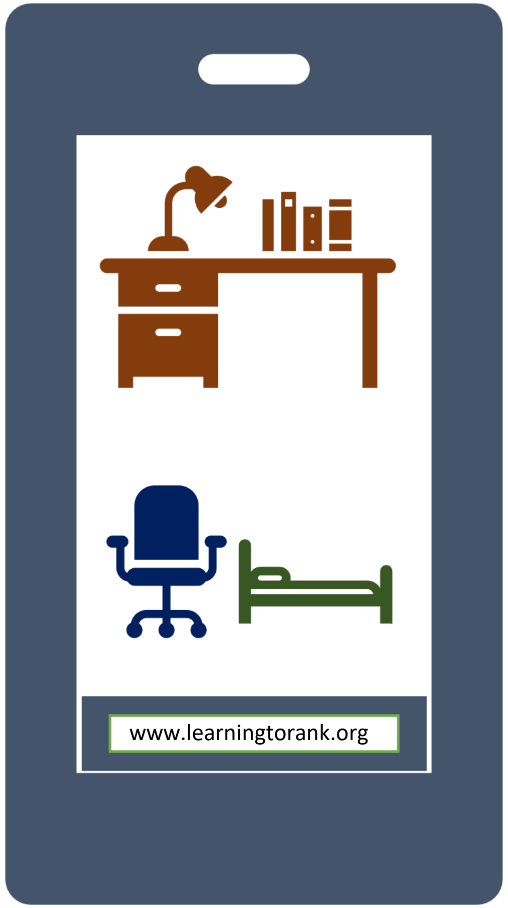
In the setting where $\lambda_1 = \cdots = \lambda_K = 1$, a pattern of repeatedly displaying the same item set until a no-click is observed is used (Agrawal et al. 2017, 2019).



Batched Decision Making

In the setting where $\lambda_1 = \cdots = \lambda_K = 1$, a pattern of repeatedly displaying the same item set until a no-click is observed is used (Agrawal et al. 2017, 2019).

$N_C = 1$

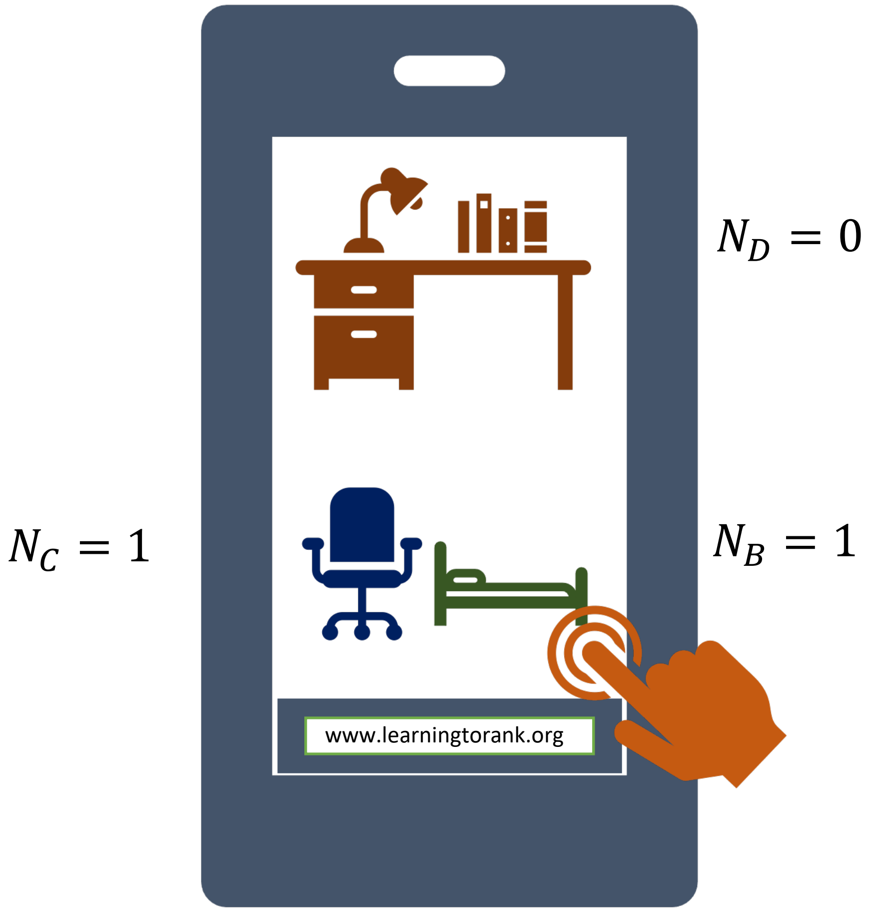


$N_D = 0$

$N_B = 1$

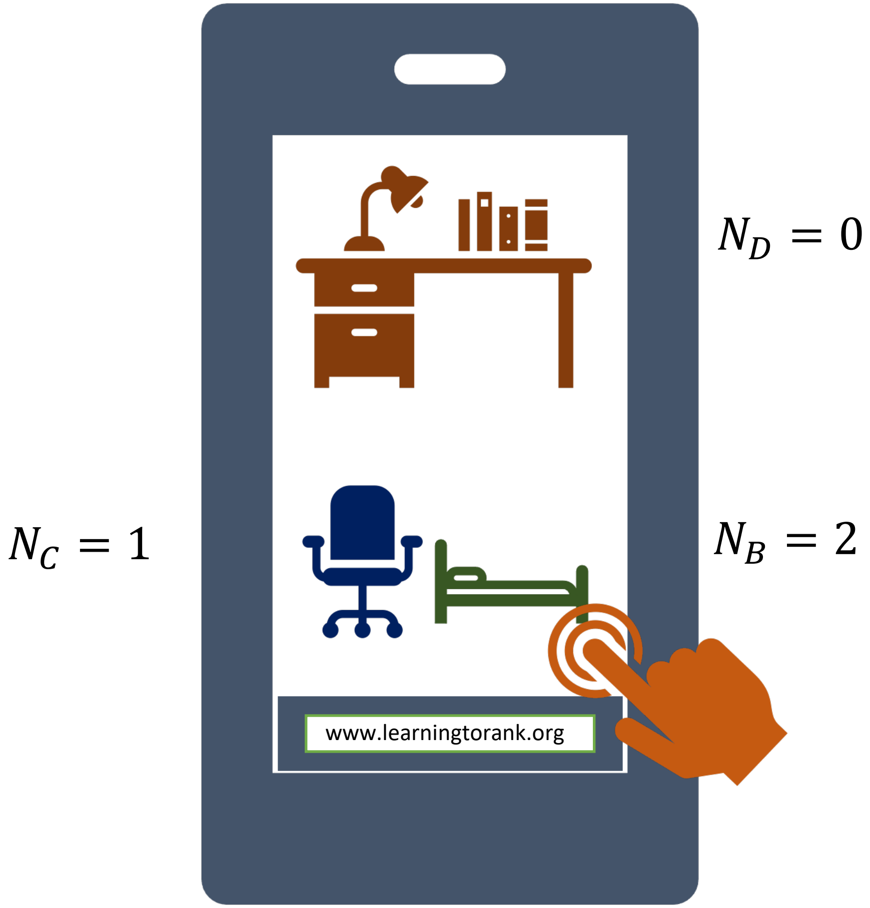
Batched Decision Making

In the setting where $\lambda_1 = \cdots = \lambda_K = 1$, a pattern of repeatedly displaying the same item set until a no-click is observed is used (Agrawal et al. 2017, 2019).



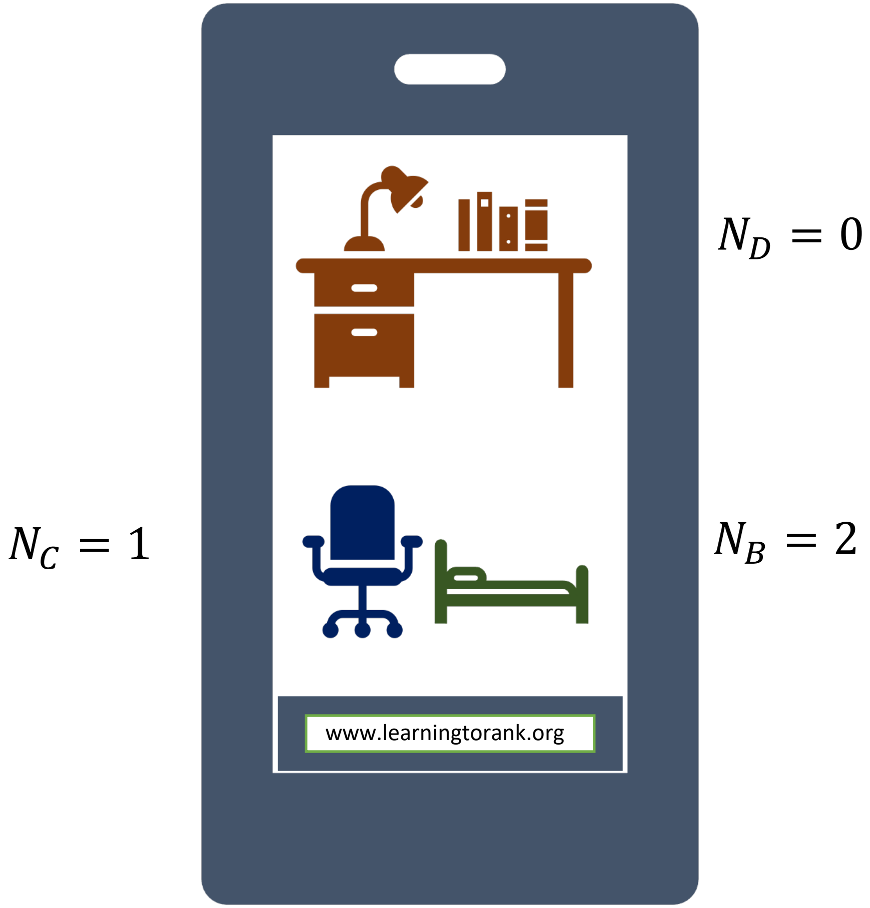
Batched Decision Making

In the setting where $\lambda_1 = \cdots = \lambda_K = 1$, a pattern of repeatedly displaying the same item set until a no-click is observed is used (Agrawal et al. 2017, 2019).



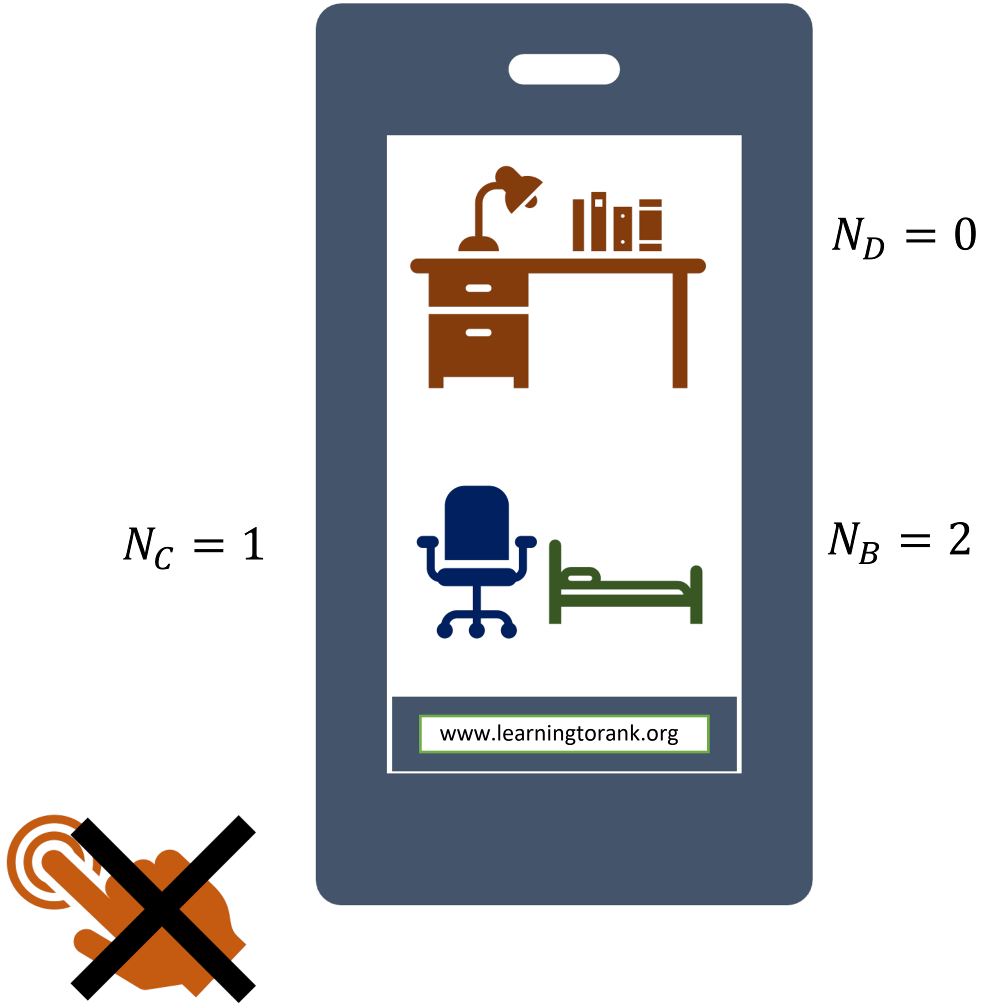
Batched Decision Making

In the setting where $\lambda_1 = \cdots = \lambda_K = 1$, a pattern of repeatedly displaying the same item set until a no-click is observed is used (Agrawal et al. 2017, 2019).



Batched Decision Making

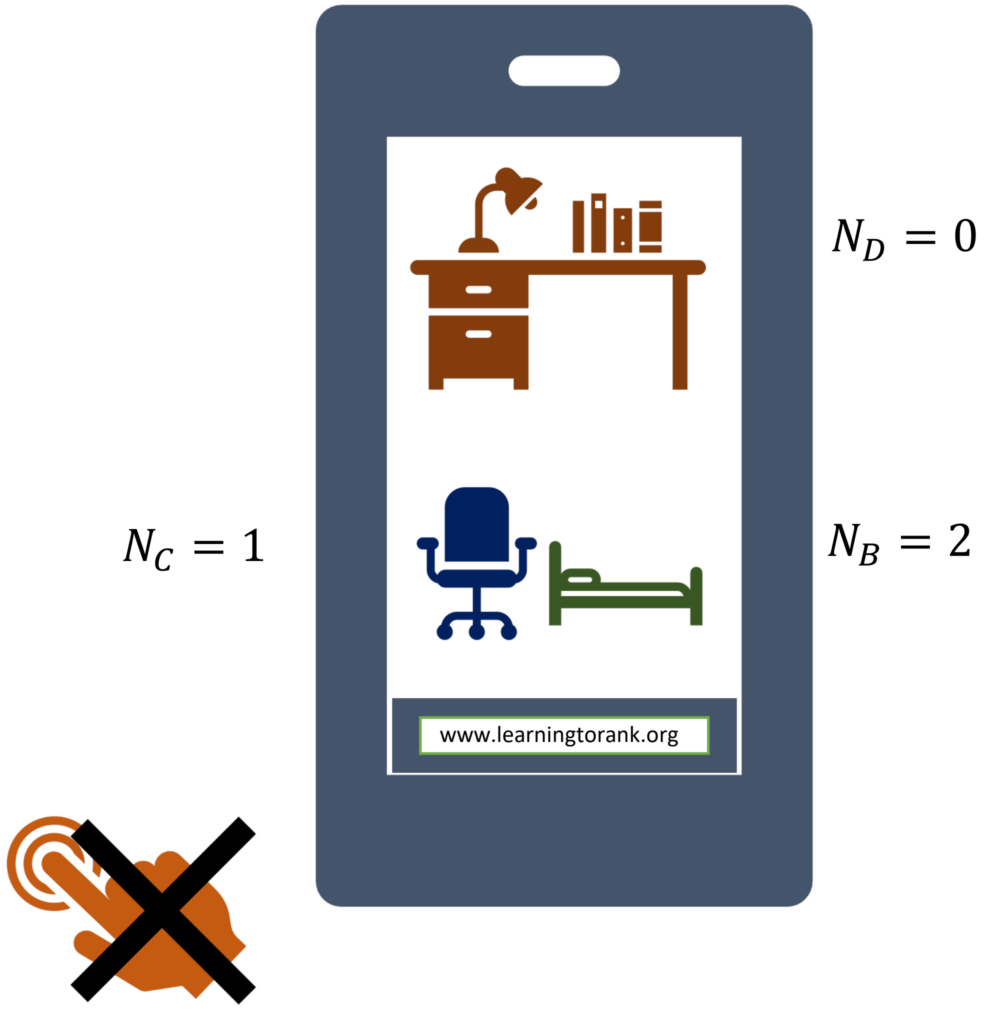
In the setting where $\lambda_1 = \cdots = \lambda_K = 1$, a pattern of repeatedly displaying the same item set until a no-click is observed is used (Agrawal et al. 2017, 2019).



Batched Decision Making

In the setting where $\lambda_1 = \cdots = \lambda_K = 1$, a pattern of repeatedly displaying the same item set until a no-click is observed is used (Agrawal et al. 2017, 2019).

Benefit is that N_B, N_C, N_D are then Geometric r.v.s conditional on $\boldsymbol{a}$.

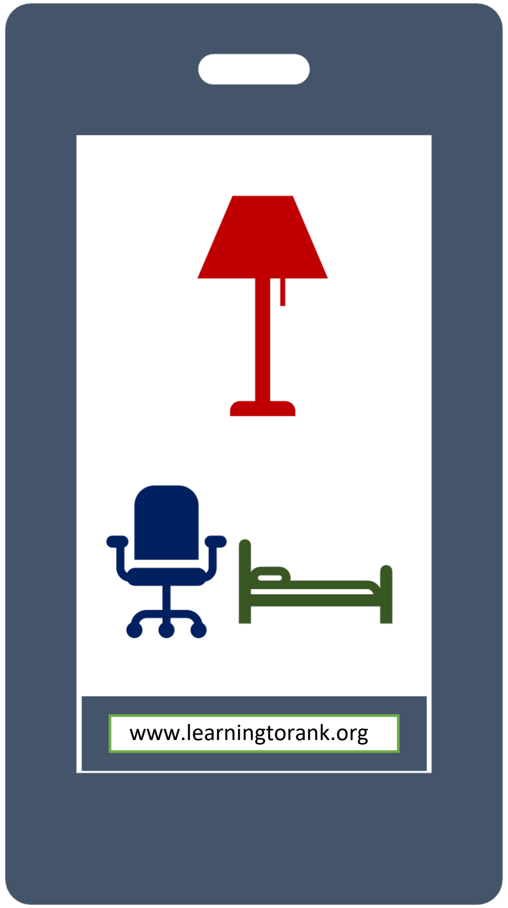


Batched Decision Making

In the setting where $\lambda_1 = \cdots = \lambda_K = 1$, a pattern of repeatedly displaying the same item set until a no-click is observed is used (Agrawal et al. 2017, 2019).

Benefit is that N_B, N_C, N_D are then Geometric r.v.s conditional on $\boldsymbol{a}$.

- Allows for MLEs which are amenable to the derivation of concentration results.



Functional Concentration Inequalities

In our setting, the batched decision making gives independent Geometric r.v.s at the item-slot combination level, but no closed form MLEs. We instead view the MLEs as functions of sums of N_{kj}.

Using logarithmic Sobolev inequalities (Joulin and Privault (2004)) we derive functional confidence intervals specific to the click model

$$P\left(\left|\hat{\alpha}_{j,L}^{MLE} - \alpha_j\right| > \sqrt{36\beta_{j,L}\log(JL)}\right) < 2/JL^2$$

where $\beta_{j,L}$ is a sum of finite difference gradients of α^{MLE} viewed as a function of the N_{kj}s.

Optimistic Algorithm

In epoch $l = 1,2, \ldots$

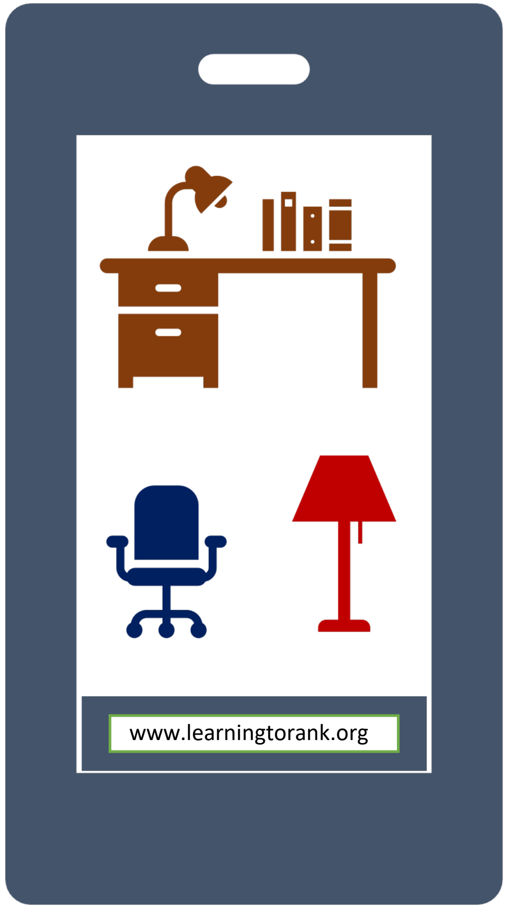
- Compute MLEs $\hat{\alpha}_{j,l}^{MLE}$ $\forall j$, and $\hat{\lambda}_{k,l}^{MLE}$
- For $j = 1, \ldots, J$ and $k = 1, \ldots, K$
  - Compute MLEs with $N_{jk,l} \leftarrow N_{jk,l} + 1, \quad \tilde{\alpha}_{j.jk,l}^{MLE}$
- Compute $\beta_{j,l}$ $\forall j$ using the $\tilde{\alpha}_{j.jk,l}^{MLE}$s
- Compute UCBs $\bar{\alpha}_{j,l} = \hat{\alpha}_{j,l} + \sqrt{36\beta_{j,l}\log(Jl)}$
- Choose an action which maximises the optimistic reward by pairing the largest $\bar{\alpha}_{j,l}$ items with the most valuable slots until all slots are filled.

Conclusions

Online learning optimal selections with MNL choice + position effects.

Future work:

- User Personalisation
    - Covariates/latent factors
- Optimism for intractable MLEs



Key References & Contact

- Agrawal, S., Avadhanula, V., Goyal, V., Zeevi, A. (2019). MNL-Bandit: A Dynamic Learning Approach to Assortment Selection. *Operations Research*
- Chuklin, A., Markov, I., and Rijke, M. d. (2015). Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*
- Grant, J.A., Leslie D.S. (2020). Learning to Rank Under Multinomial Logit Choice. In Submission, arXiv:2009.03207.
- Joulin, A., Privault, N. (2004). Functional Inequalities for Discrete Gradients and an Application to the Geometric Distribution. *ESAIM: Probability and Statistics.*

j.grant@lancaster.ac.uk @james_a_grant