Improved Online Confidence Bounds for Multinomial Logistic Bandits

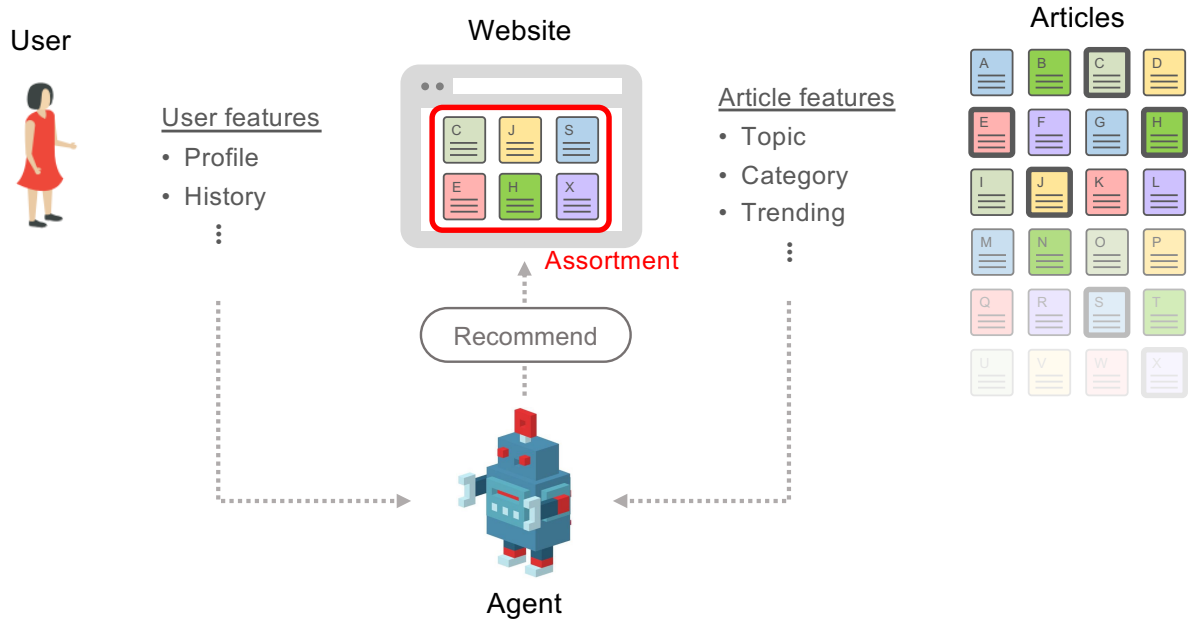
(ICML 2025)

Joongkyu Lee & Min-hwan Oh

Seoul National University

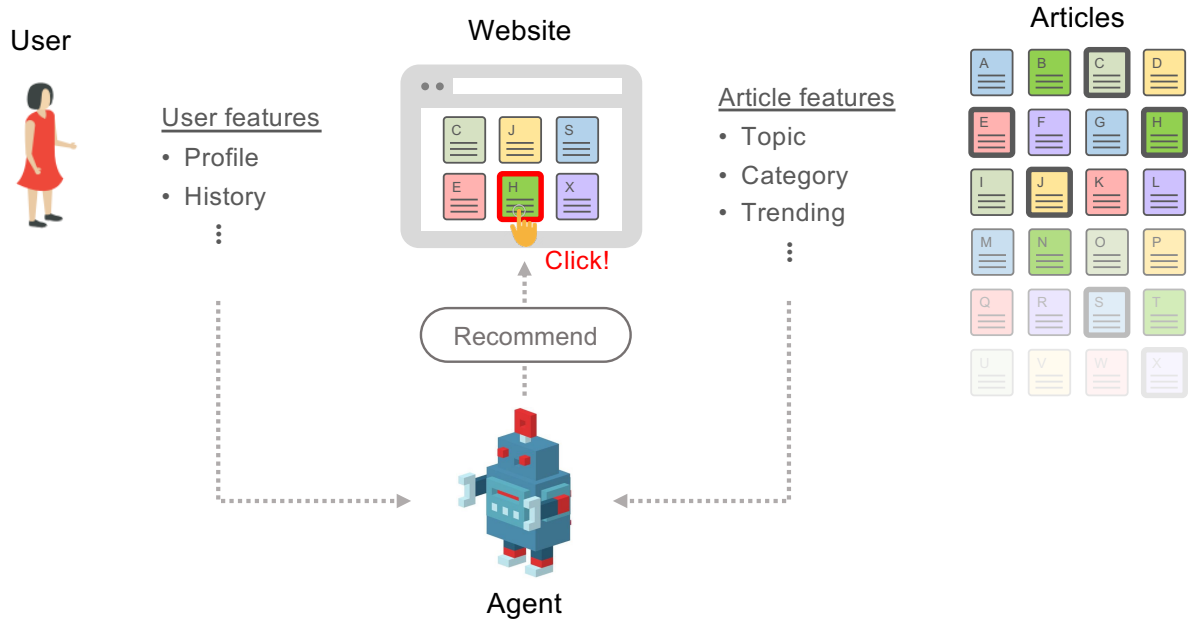
SEOUL NATIONAL UNIVERSITY

Sequential Assortment Selection Problem



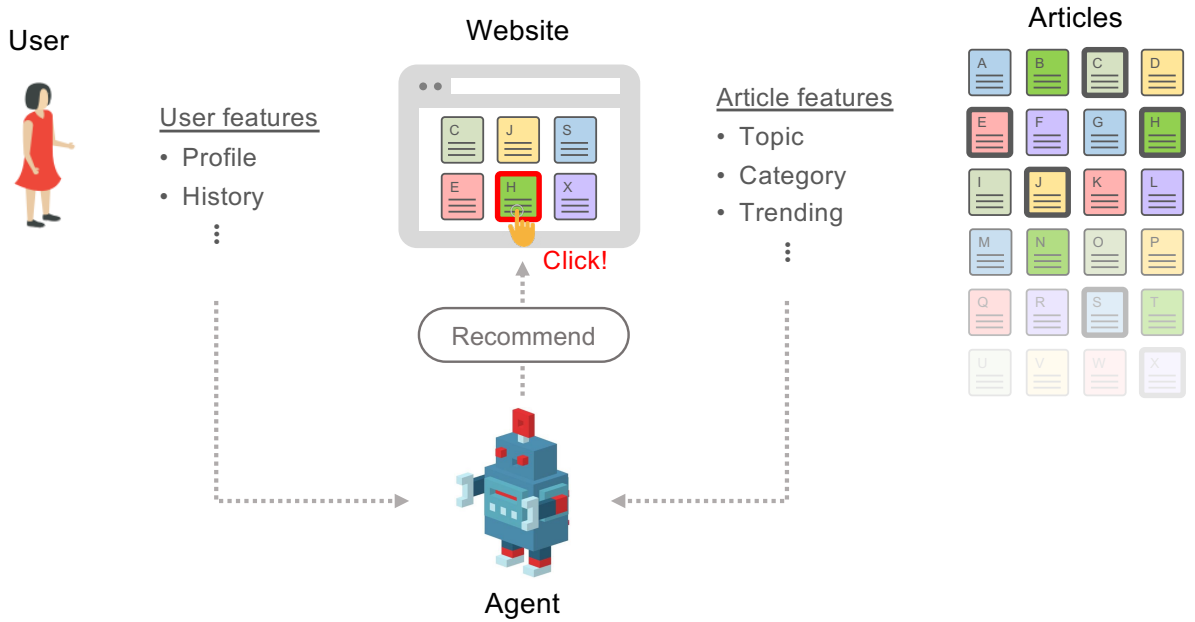
- Agent recommends an assortment (a set of items)

Sequential Assortment Selection Problem



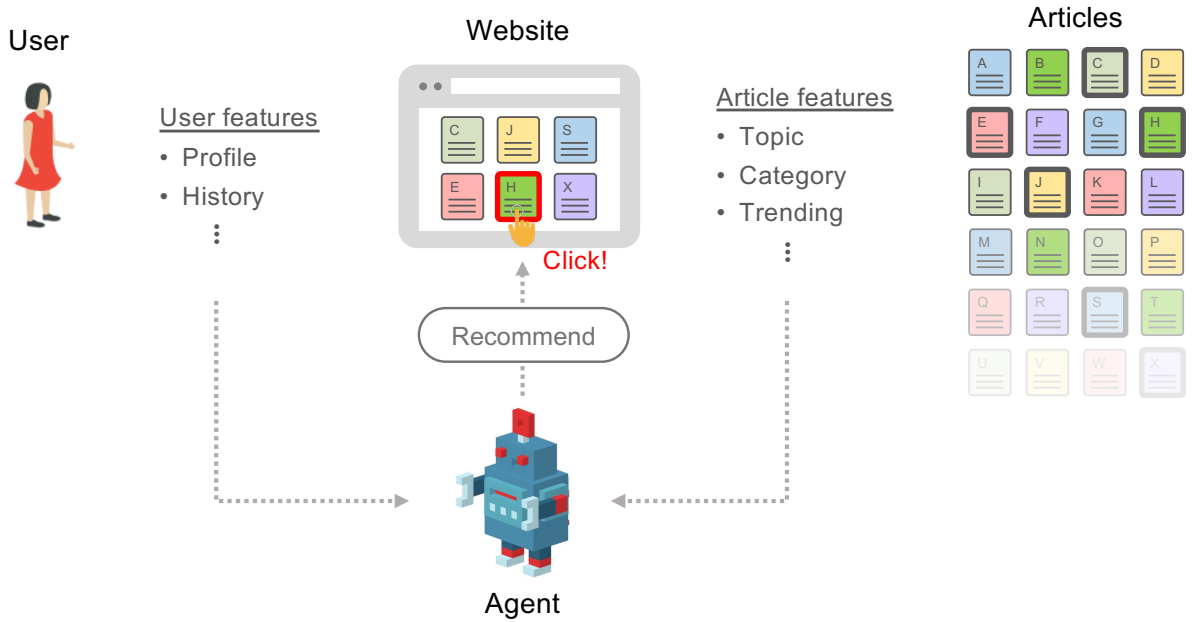
- Agent recommends an assortment (a set of items)
- User chooses one item from offered multiple options

Sequential Assortment Selection Problem



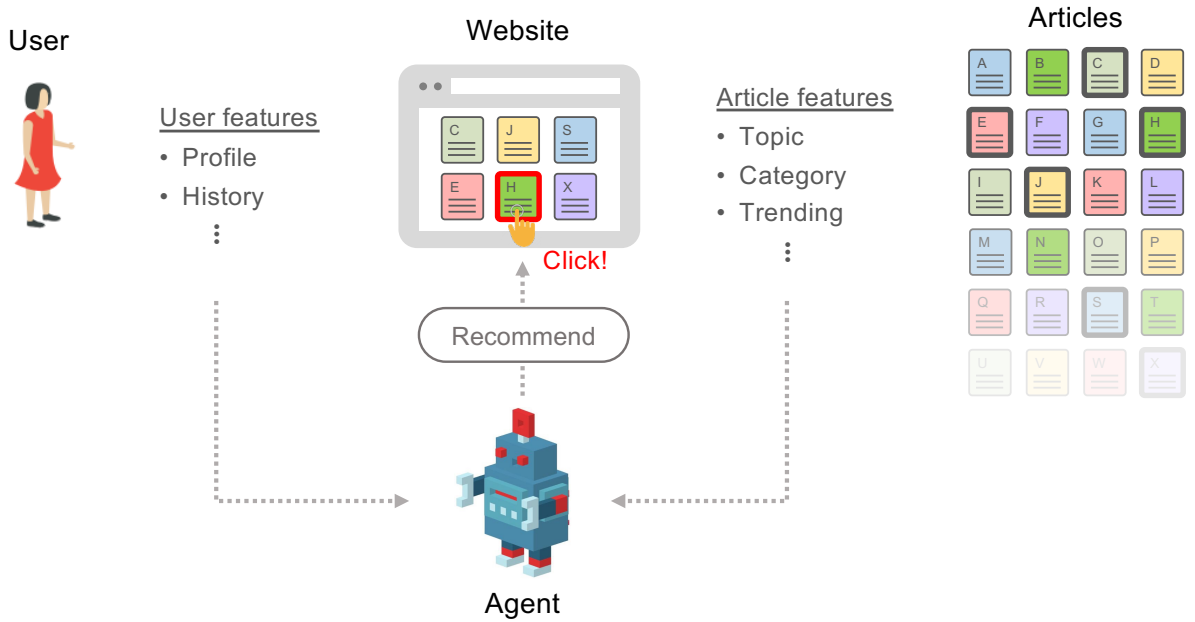
- Agent recommends an assortment (a set of items)
- User chooses one item from offered multiple options
- For every round $t = 1, \ldots, T$:
  1. Observe contexts $x_{ti} \in \mathbb{R}^d$ and rewards $r_{ti} \in [0, 1]$ for every item $i \in [N]$

Sequential Assortment Selection Problem



- Agent recommends an assortment (a set of items)
- User chooses one item from offered multiple options
- For every round $t = 1, \ldots, T$:
  1. Observe contexts $x_{ti} \in \mathbb{R}^d$ and rewards $r_{ti} \in [0, 1]$ for every item $i \in [N]$
  2. Offer an assortment $S_t = \{i_1, \ldots, i_m\}$ such that $m \leq K$

Sequential Assortment Selection Problem



- Agent recommends an assortment (a set of items)
- User chooses one item from offered multiple options
- For every round $t = 1, \ldots, T$:
  1. Observe contexts $x_{ti} \in \mathbb{R}^d$ and rewards $r_{ti} \in [0, 1]$ for every item $i \in [N]$
  2. Offer an assortment $S_t = \{i_1, \ldots, i_m\}$ such that $m \leq K$
  3. Observe the user click decision $c_t \in S_t \cup \{0\}$ ("0": outside option)

Multinomial Logit (MNL) Choice Model (McFadden, 1977)

- Probability of choosing any item i in assortment S_t:

$$p_t(i|S_t, \mathbf{w}^\star) := \frac{\exp(x_{ti}^\top \mathbf{w}^\star)}{1+\sum_{j\in S_t} \exp(x_{tj}^\top \mathbf{w}^\star)}$$

  - Here, $\mathbf{w}^\star \in \mathbb{R}^d$ is an *unknown* parameter
  - Bounded assumption: $\|\mathbf{w}^\star\|_2 \leq B$

Multinomial Logit (MNL) Choice Model (McFadden, 1977)

- Probability of choosing any item i in assortment S_t:

$$p_t(i|S_t, \mathbf{w}^\star) := \frac{\exp(x_{ti}^\top \mathbf{w}^\star)}{1+\sum_{j\in S_t} \exp(x_{tj}^\top \mathbf{w}^\star)}$$

  - Here, $\mathbf{w}^\star \in \mathbb{R}^d$ is an *unknown* parameter
  - Bounded assumption: $\|\mathbf{w}^\star\|_2 \le B$

- Expected revenue of the assortment S:

$$R_t(S, \mathbf{w}^\star) := \sum_{i\in S} p_t(i|S, \mathbf{w}^\star) r_{ti} = \frac{\sum_{i\in S} \exp(x_{ti}^\top \mathbf{w}^\star) r_{ti}}{1+\sum_{j\in S} \exp(x_{tj}^\top \mathbf{w}^\star)}$$

  - WLOG, let $r_{ti} \in [0, 1]$.
  - We say the rewards are *uniform* when $r_{ti} \equiv 1$ for all t and i.

Multinomial Logit (MNL) Choice Model (McFadden, 1977)

- Probability of choosing any item i in assortment S_t:

$$p_t(i|S_t, \mathbf{w}^\star) := \frac{\exp(x_{ti}^\top \mathbf{w}^\star)}{1+\sum_{j\in S_t} \exp(x_{tj}^\top \mathbf{w}^\star)}$$

  - Here, $\mathbf{w}^\star \in \mathbb{R}^d$ is an *unknown* parameter
  - Bounded assumption: $\|\mathbf{w}^\star\|_2 \le B$

- Expected revenue of the assortment S:

$$R_t(S, \mathbf{w}^\star) := \sum_{i\in S} p_t(i|S, \mathbf{w}^\star) r_{ti} = \frac{\sum_{i\in S} \exp(x_{ti}^\top \mathbf{w}^\star) r_{ti}}{1+\sum_{j\in S} \exp(x_{tj}^\top \mathbf{w}^\star)}$$

  - WLOG, let $r_{ti} \in [0, 1]$.
  - We say the rewards are *uniform* when $r_{ti} \equiv 1$ for all t and i.

- Optimal assortment: $S_t^\star = \arg\max_{S\in\mathcal{S}} R_t(S, \mathbf{w}^\star)$

Multinomial Logit (MNL) Choice Model (McFadden, 1977)

- Probability of choosing any item i in assortment S_t:

$$p_t(i|S_t, \mathbf{w}^\star) := \frac{\exp(x_{ti}^\top \mathbf{w}^\star)}{1+\sum_{j\in S_t} \exp(x_{tj}^\top \mathbf{w}^\star)}$$

  - Here, $\mathbf{w}^\star \in \mathbb{R}^d$ is an *unknown* parameter
  - Bounded assumption: $\|\mathbf{w}^\star\|_2 \le B$

- Expected revenue of the assortment S:

$$R_t(S, \mathbf{w}^\star) := \sum_{i\in S} p_t(i|S, \mathbf{w}^\star) r_{ti} = \frac{\sum_{i\in S} \exp(x_{ti}^\top \mathbf{w}^\star) r_{ti}}{1+\sum_{j\in S} \exp(x_{tj}^\top \mathbf{w}^\star)}$$

  - WLOG, let $r_{ti} \in [0, 1]$.
  - We say the rewards are *uniform* when $r_{ti} \equiv 1$ for all t and i.

- Optimal assortment: $S_t^\star = \arg\max_{S\in\mathcal{S}} R_t(S, \mathbf{w}^\star)$
- Goal: Minimize $\mathbf{Reg}_T(\mathbf{w}^\star) = \sum_{t=1}^T R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star)$

Previous Works & Main Contributions

Table. T: total rounds, d: feature dimension, K: maximum assortment size, B: bound on the parameter norm, $1/\kappa = \mathcal{O}(K^2 e^B)$, $\kappa_t^\star = \sum_{i \in S_t^\star} p_t(i|S_t^\star, \mathbf{w}^\star) p_t(0|S_t^\star, \mathbf{w}^\star) \leq 1$, $\sigma_t^2 \leq 1$: reward variance at round t.

	Regret	Rewards	Comput. per Round
Chen et al. (2020) (MLE-UCB)	$\mathcal{O}\Big(Bd\log(KT)\sqrt{T}\Big)$	$r_{ti} \in [0,1]$	Intractable
Oh and Iyengar (2021) (UCB-MNL)	$\mathcal{O}\Big(\frac{1}{\kappa} d \log T \sqrt{T}\Big) = \mathcal{O}\Big(K^2 e^B d \log T \sqrt{T}\Big)$	$r_{ti} \in [0,1]$	$\mathcal{O}(t)$
Perivier and Goyal (2022) (OFU-MNL)	$\mathcal{O}\Big(BKd\log(KT)\sqrt{\sum_{t=1}^T \kappa_t^\star}\Big)$	$r_{ti} \equiv 1$	Intractable
Lee and Oh (2024) (OFU-MNL+)	$\mathcal{O}\Big(B^{3/2} d \log K (\log T)^{3/2} \sqrt{T}\Big)$	$r_{ti} \in [0,1]$	$\mathcal{O}(1)$

1. No results free of $\mathrm{poly}(B)$ dependence
2. No results completely independent of K
3. No variance-dependent bounds for arbitrary rewards $r_{ti} \in [0,1]$

Previous Works & Main Contributions

Table. T: total rounds, d: feature dimension, K: maximum assortment size, B: bound on the parameter norm, $1/\kappa = \mathcal{O}(K^2 e^B)$, $\kappa_t^\star = \sum_{i \in S_t^\star} p_t(i|S_t^\star, \mathbf{w}^\star) p_t(0|S_t^\star, \mathbf{w}^\star) \leq 1$, $\sigma_t^2 \leq 1$: reward variance at round t.

	Regret	Rewards	Comput. per Round
Chen et al. (2020) (MLE-UCB)	$\mathcal{O}\Big(Bd\log(KT)\sqrt{T}\Big)$	$r_{ti} \in [0,1]$	Intractable
Oh and Iyengar (2021) (UCB-MNL)	$\mathcal{O}\Big(\frac{1}{\kappa} d \log T \sqrt{T}\Big) = \mathcal{O}\Big(K^2 e^B d \log T \sqrt{T}\Big)$	$r_{ti} \in [0,1]$	$\mathcal{O}(t)$
Perivier and Goyal (2022) (OFU-MNL)	$\mathcal{O}\Big(BKd\log(KT)\sqrt{\sum_{t=1}^T \kappa_t^\star}\Big)$	$r_{ti} \equiv 1$	Intractable
Lee and Oh (2024) (OFU-MNL+)	$\mathcal{O}\Big(B^{3/2} d \log K (\log T)^{3/2} \sqrt{T}\Big)$	$r_{ti} \in [0,1]$	$\mathcal{O}(1)$
This work **(OFU-MNL++)**	$\mathcal{O}\Big(d \log T \sqrt{\sum_{t=1}^T \sigma_t^2}\Big)$, for large t	$r_{ti} \in [0,1]$	$\mathcal{O}(1)$

1. No results free of $\mathrm{poly}(B)$ dependence
2. No results completely independent of K
3. No variance-dependent bounds for arbitrary rewards $r_{ti} \in [0,1]$

First $\mathrm{poly}(B), K$-free, and variance-dependent regret bound in MNL bandits!

Parameter Update using Online Mirror Descent (OMD)

- **MNL loss function**: For the choice response $y_{ti} = \mathbb{1}(c_t = i)$, we define:

$$\ell_t(\mathbf{w}) := -\sum_{i \in S_t} y_{ti} \log p_t(i|S_t, \mathbf{w}).$$

- **Online parameter estimation**: Estimate $\mathbf{w}^\star$ by an online mirror descent (Orabona, 2019).

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w} \in \mathcal{W}} \langle \nabla \ell_t(\mathbf{w}_t), \mathbf{w} \rangle + \frac{1}{2\eta} \|\mathbf{w} - \mathbf{w}_t\|^2_{\widetilde{H}_t}, \quad \forall t \geq 1.$$

  - Step-size parameter: $\eta > 0$
  - Hessian matrices: $H_t := \lambda \mathbf{I}_d + \sum_{s=1}^{t-1} \nabla^2 \ell_s(\mathbf{w}_{s+1})$ and $\widetilde{H}_t := H_t + \eta \ell_t(\mathbf{w}_t)$

- Computation cost: $\mathcal{O}(Kd^3)$
- Storage cost: $\mathcal{O}(d^2)$

Result 1: Improved Online Confidence Bound

Theorem 1. Improved online confidence bound

Let $\lambda > 0$ and $\mathcal{T} \subseteq [T]$ denote the set of update rounds. For all $t \in \mathcal{T}$, we assume the following update conditions hold:

$$\sup_{\mathbf{w} \in \mathcal{W}_t} |x_{ti}^\top (\mathbf{w} - \mathbf{w}^\star)| \leq \alpha, \quad \forall i \in S_t,$$

where $\mathcal{W}_t$ is a compact convex set. Then, with high probability, we have:

$$\|\mathbf{w}_t - \mathbf{w}^\star\|_{H_t} = \mathcal{O}\left(\alpha\sqrt{d\log(t/\delta)} + \alpha\sqrt{\lambda}\right).$$

- Updating at every round ($\mathcal{T} = [T]$) yields $\mathcal{O}(B\sqrt{d\log t})$.
  - Improves over Lee and Oh (2024): $\mathcal{O}(B\sqrt{d}\log t \log K)$

Result 1: Improved Online Confidence Bound

Theorem 1. Improved online confidence bound

Let $\lambda > 0$ and $\mathcal{T} \subseteq [T]$ denote the set of update rounds. For all $t \in \mathcal{T}$, we assume the following update conditions hold:

$$\sup_{\mathbf{w} \in \mathcal{W}_t} |x_{ti}^\top (\mathbf{w} - \mathbf{w}^\star)| \leq \alpha, \quad \forall i \in S_t,$$

where $\mathcal{W}_t$ is a compact convex set. Then, with high probability, we have:

$$\|\mathbf{w}_t - \mathbf{w}^\star\|_{H_t} = \mathcal{O}\left(\alpha\sqrt{d\log(t/\delta)} + \alpha\sqrt{\lambda}\right).$$

- Updating at every round ($\mathcal{T} = [T]$) yields $\mathcal{O}(B\sqrt{d\log t})$.
  - Improves over Lee and Oh (2024): $\mathcal{O}(B\sqrt{d}\log t \log K)$
- **Key implication**: If $\mathcal{W}_t$ is constructed so that α stays small, and updates occur only in this case, we obtain $\mathcal{O}(\sqrt{d\log t})$ for large t.
  - No dependence on B or K!

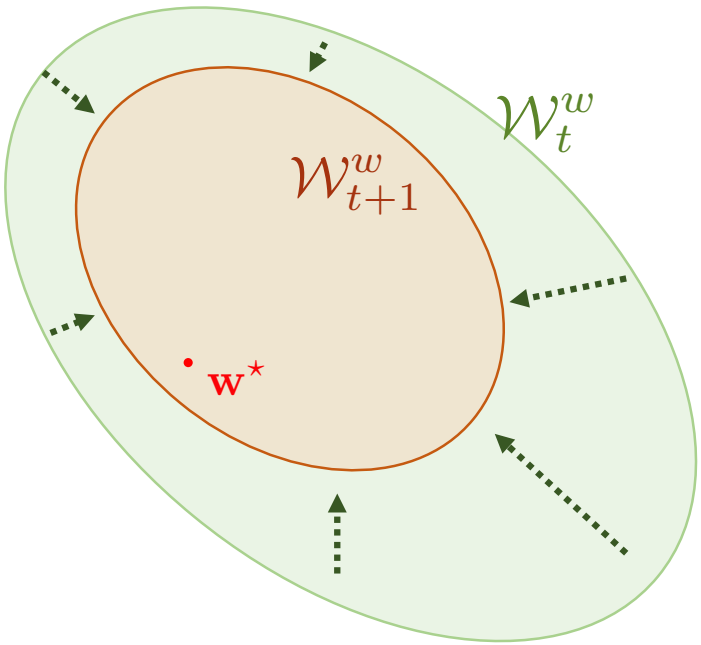
Algorithm

Algorithm OFU-MNL++

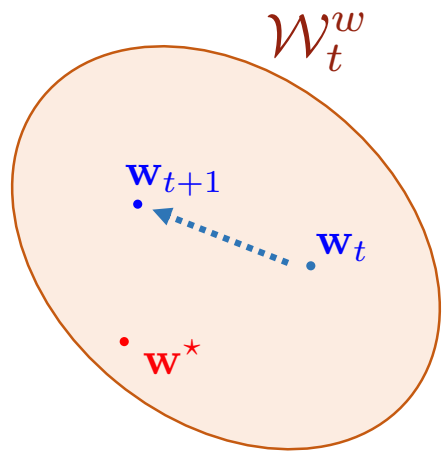
Initialize: $\mathcal{W}_1^w(\delta) = \mathcal{W}$, $H_1 = \lambda \mathbf{I}_d$, $H_1^w = \lambda_1^w \mathbf{I}_d$, $\mathbf{w}_1, \mathbf{w}_1^w \in \mathcal{W}$
for round $t = 1, \ldots, T$ **do**
  Observe feature set $\mathcal{X}_t = \{x_{ti}\}_{i=1}^N$ and rewards $\{r_{ti}\}_{i=1}^N$
  if $\max_{x \in \mathcal{X}_t} \|x\|_{(H_t^w)^{-1}} \geq \frac{1}{6\sqrt{2}\zeta_t(\delta)}$ **then** ▷ *Large* α
    Offer $S_t = \{i_t\}$, where $x_{ti_t} = \arg\max_{x \in \mathcal{X}_t} \|x\|^2_{(H_t^w)^{-1}}$
    Update $(\mathbf{w}_{t+1}^w, H_{t+1}^w) \leftarrow \texttt{OMD}(\mathcal{W}, \ell_t, H_t^w, \mathbf{w}_t^w, \eta^w)$
    Calculate $\mathcal{W}_{t+1}^w(\delta) \leftarrow \left\{\mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w} - \mathbf{w}_{t+1}^w\|_{H_{t+1}^w} \leq \zeta_{t+1}(\delta)\right\}$
    Update $H_{t+1} \leftarrow H_t$ and $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t$.
  else ▷ *Small* α
    Offer $S_t = \arg\max_{S \in \mathcal{S}} \widetilde{R}_t(S)$
    Update $(\mathbf{w}_{t+1}, H_{t+1}) \leftarrow \texttt{OMD}(\mathcal{W}_t^w(\delta), \ell_t, H_t, \mathbf{w}_t, \eta)$
    Update $H_{t+1}^w \leftarrow H_t^w$, $\mathbf{w}_{t+1}^w \leftarrow \mathbf{w}_t^w$, and $\mathcal{W}_{t+1}^w(\delta) \leftarrow \mathcal{W}_t^w(\delta)$
  end if
end for

Algorithm

i) Large α: update search space



ii) Small α: update search space



Result 2: Regret Bound

Theorem 2. $\mathrm{poly}(B), K$**-free variance-dependent regret bound**

With high probability, the regret of `OFU-MNL++` satisfies:

$$\mathbf{Reg}_T(\mathbf{w}^\star) = \mathcal{O}\left(\left(d\log T + Bd\sqrt{\log T}\right)\sqrt{\sum_{t=1}^{T}\sigma_t^2}\right).$$

Result 2: Regret Bound

Theorem 2. $\mathrm{poly}(B), K$**-free variance-dependent regret bound**

With high probability, the regret of `OFU-MNL++` satisfies:

$$\mathbf{Reg}_T(\mathbf{w}^\star) = \mathcal{O}\left(\left(d\log T + Bd\sqrt{\log T}\right)\sqrt{\sum_{t=1}^{T}\sigma_t^2}\right).$$

- Variance-dependent regret achieved via refined analysis.

Result 2: Regret Bound

Theorem 2. $\mathrm{poly}(B), K$**-free variance-dependent regret bound**

With high probability, the regret of `OFU-MNL++` satisfies:

$$\mathbf{Reg}_T(\mathbf{w}^\star) = \mathcal{O}\left(\left(d\log T + Bd\sqrt{\log T}\right)\sqrt{\sum_{t=1}^{T}\sigma_t^2}\right).$$

- Variance-dependent regret achieved via refined analysis.
- For sufficiently large T: $\mathcal{O}\left(d\log T\sqrt{\sum_{t=1}^{T}\sigma_t^2}\right)$.

Result 2: Regret Bound

Theorem 2. $\mathrm{poly}(B), K$-free variance-dependent regret bound

With high probability, the regret of `OFU-MNL++` satisfies:

$$\mathbf{Reg}_T(\mathbf{w}^\star) = \mathcal{O}\left(\left(d\log T + Bd\sqrt{\log T}\right)\sqrt{\sum_{t=1}^{T}\sigma_t^2}\right).$$

- Variance-dependent regret achieved via refined analysis.
- For sufficiently large T: $\mathcal{O}\left(d\log T\sqrt{\sum_{t=1}^{T}\sigma_t^2}\right)$.
  - First variance-dependent and $\mathrm{poly}(B), K$-free regret
  - Compared to Lee and Oh (2024): $\mathcal{O}\left(B^{3/2}d\log K(\log T)^{3/2}\sqrt{T}\right)$
    $\Longrightarrow$ ours improves by a factor of $\mathcal{O}\left(B^{3/2}\log K\sqrt{\log T}\right)$.

Result 2: Regret Bound

Theorem 2. $\mathrm{poly}(B), K$-free variance-dependent regret bound

With high probability, the regret of `OFU-MNL++` satisfies:

$$\mathbf{Reg}_T(\mathbf{w}^\star) = \mathcal{O}\left(\left(d\log T + Bd\sqrt{\log T}\right)\sqrt{\sum_{t=1}^{T}\sigma_t^2}\right).$$

- Variance-dependent regret achieved via refined analysis.
- For sufficiently large T: $\mathcal{O}\left(d\log T\sqrt{\sum_{t=1}^{T}\sigma_t^2}\right)$.
  - First variance-dependent and $\mathrm{poly}(B), K$-free regret
  - Compared to Lee and Oh (2024): $\mathcal{O}\left(B^{3/2}d\log K(\log T)^{3/2}\sqrt{T}\right)$
    $\Longrightarrow$ ours improves by a factor of $\mathcal{O}\left(B^{3/2}\log K\sqrt{\log T}\right)$.
  - Since $\sigma_t^2 \le 1$, this represents a strict improvement over $\sqrt{T}$.

References I

Chen, X., Wang, Y., and Zhou, Y. (2020). Dynamic assortment optimization with changing contextual information. The Journal of Machine Learning Research, 21(1):8918–8961.

Lee, J. and Oh, M.-h. (2024). Nearly minimax optimal regret for multinomial logistic bandit. In The Thirty-eighth Annual Conference on Neural Information Processing Systems.

McFadden, D. (1977). Modelling the choice of residential location.

Oh, M.-h. and Iyengar, G. (2021). Multinomial logit contextual bandits: Provable optimality and practicality. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 9205–9213.

Orabona, F. (2019). A modern introduction to online learning. arXiv preprint arXiv:1912.13213.

Perivier, N. and Goyal, V. (2022). Dynamic pricing and assortment under a contextual mnl demand. Advances in Neural Information Processing Systems, 35:3461–3474.