Thompson Sampling for Multinomial Logit Contextual Bandits

Min-hwan Oh and **Garud Iyengar**

m.oh@columbia.edu

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

Introduction

- Which **set of items (assortment)** should you recommend?

- Most common form of recommendations in practice
  - Online retail: Amazon, Walmart, eBay, etc.
  - Video streaming services: Netflex, Youtube, etc.
  - News websites/feeds, web searches and many more
- Contextual information is readily available
  - User profile, search keywords
  - Features of items to be recommended

Multinomial Logit Contextual Bandits

"Combinatorial Contextual Bandit **with User Choice**"

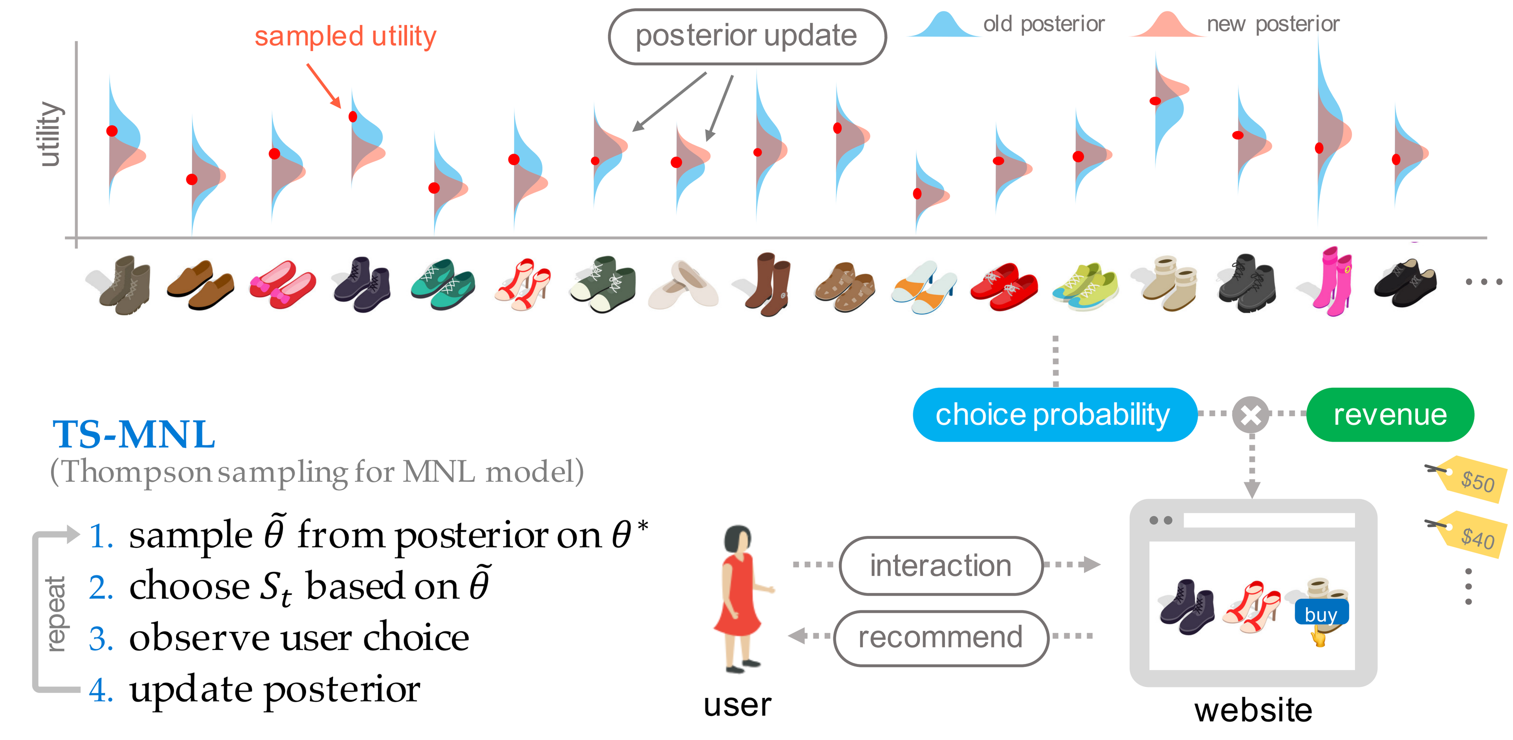
- For each round $t = 1, \dots, T$:
  1. Context $x_{ti} \in \mathbb{R}^d$ and revenue r_{ti} revealed for all items $i \in [N]$
  2. Agent selects assortment $S_t \subset [N]$ (with $|S_t| \le K$)
  3. Agent observes user choice $y_t \in \{0,1\}^{|S_t|}$
- Choice given by **multinomial logit** (MNL) model $p_i(S_t, \theta^*)$
  - Probability that user chooses $i \in S_t$ [3]:

$$p(i|S_t, \theta^*) = \frac{\exp(\overbrace{x_{ti}^\top \theta^*}^{\text{utility}})}{1 + \sum_{j \in S_t} \exp(x_{tj}^\top \theta^*)}$$

  - $\theta^* \in \mathbb{R}^d$ unknown true parameter
- Expected revenue for assortment S_t: $R_t(S_t, \theta^*) = \sum_{i \in S_t} r_{ti} p(i|S_t, \theta^*)$
- Goal: minimize total regret

$$\text{Regret}(T) = \mathbb{E}\Big[\underbrace{\sum_{t=1}^{T} R_t(S_t^*, \theta^*)}_{\text{optimal total revenue}} - \underbrace{\sum_{t=1}^{T} R_t(S_t, \theta^*)}_{\text{agent's total revenue}} \Big]$$

where $S_t^* = \arg\max_S R_t(S, \theta^*)$



TS-MNL
(Thompson sampling for MNL model)

repeat
1. sample $\tilde{\theta}$ from posterior on θ^*
2. choose S_t based on $\tilde{\theta}$
3. observe user choice
4. update posterior

Theorem (Bayesian Regret)

The Bayesian regret of TS-MNL is: $\text{BayesRegret}(T) = \widetilde{\mathcal{O}}(d\sqrt{T})$

- But, can we show the worst-case regret?

Challenges in Worst-Case Regret Analysis

- Decomposing worst-case immediate regret:

$$\text{Regret}(t) = \underbrace{\mathbb{E}[R_t(S_t^*, \theta^*) - R_t(S_t, \widetilde{\theta}_t)]}_{(a)} + \underbrace{\mathbb{E}[R_t(S_t, \widetilde{\theta}_t) - R_t(S_t, \theta^*)]}_{(b)}$$

- (b) controlled by concentration of $\widetilde{\theta}_t$
- (a) controlled by ensuring optimism of sampled $\widetilde{\theta}_t$
  - In Bayesian regret, $(a) = 0$ since $\widetilde{\theta}_t$ and θ^* are iid
  - Probability each utility is optimistic: exponentially small in K

TS-MNL with Optimistic Sampling

- Sample from Gaussian distribution
  - TS as generic randomized algorithm based on MLE $\hat{\theta}_t$
  - Sample $\widetilde{\theta}_t \sim \mathcal{N}(\hat{\theta}_t, \alpha_t^2 V_t^{-1})$: α_t = conf. radius, $V_t = \sum_{\tau} \sum_{i \in S_\tau} x_{\tau i} x_{\tau i}^\top$
- Optimistic sampling
  - Draw M samples $\{\widetilde{\theta}_t^{(j)}\}_{j=1}^M$ from $\mathcal{N}(\hat{\theta}_t, \alpha_t^2 V_t^{-1})$
  - Define optimistic utility: $\widetilde{u}_{ti} = \max_{1 \le j \le M} \{x_{ti}^\top \widetilde{\theta}_t^{(j)}\}$
  - Expected revenue of assortment S based on $\widetilde{u}_{ti}$:

$$\widetilde{R}_t(S) = \frac{\sum_{i \in S} r_{ti} \exp\{\widetilde{u}_{ti}\}}{1 + \sum_{j \in S} \exp\{\widetilde{u}_{tj}\}}$$

Lemma (Ensuring Optimism)

Let $\alpha_t = \mathcal{O}\big(\sqrt{2d \log(1 + t/d)}\big)$ and take $M = \lceil 1 + C \log K \rceil$ samples for some constant C. Then $\mathbb{P}\big(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta^*) \mid \mathcal{F}_t\big) \ge \frac{1}{4\sqrt{e\pi}}$.

- Choose optimistic assortment at least with a constant frequency
- Cumulative regret due to random sampling can be bounded

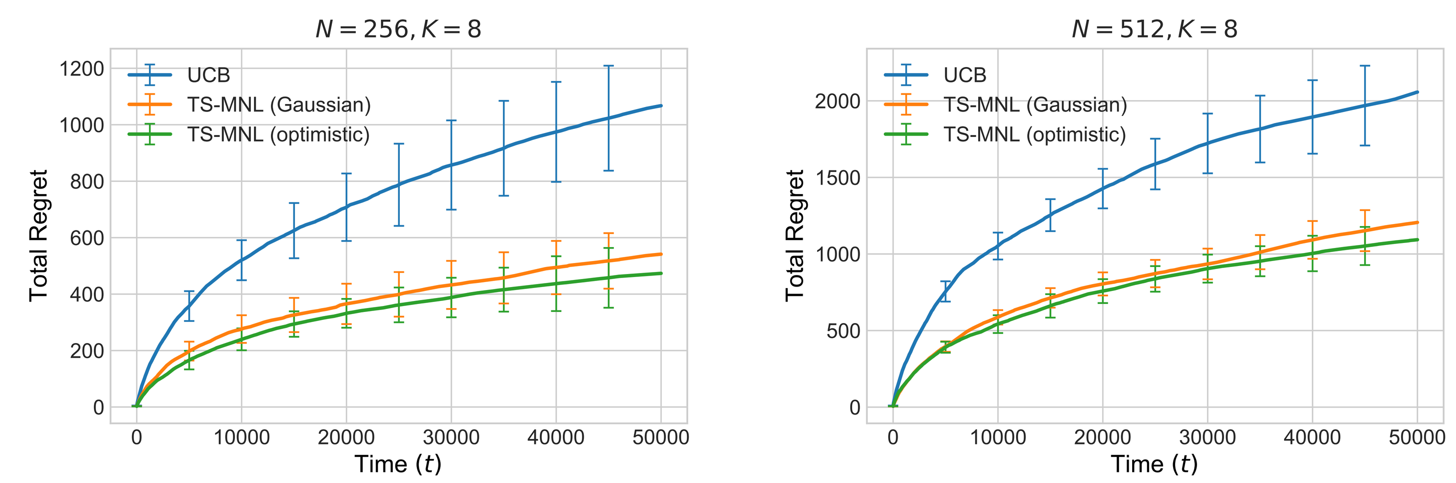
Theorem (Worst-Case Regret)

The **worst-case** regret of TS-MNL + **optimistic sampling** with $M = \lceil 1 + C \log K \rceil$ samples is: $\text{Regret}(T) = \widetilde{\mathcal{O}}\big(d^{3/2}\sqrt{T}\big)$

- Matches regret bound for linear TS bandits [1]
- Additional $\sqrt{d}$ factor vs Bayesian regret: deviation of random sampling addressed in worst-case regret analysis
- In case of a finite number of items (actions), i.e., $N < e^d$, $\mathcal{O}(d\sqrt{T \log N} \log T)$ worst-case regret
- First worst-case regret guarantee of Thompson sampling for combinatorial contextual bandit

Numerical Experiments

- Dataset: MovieLens 1M dataset (https://movielens.org)
- 1M ratings of 4000 movies by 6000 users: ratings on 1-5 scale
- Comparison with UCB method [2] and TS-MNL variants



References

[1] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

[2] Xi Chen, Yining Wang, and Yuan Zhou. Dynamic assortment optimization with changing contextual information. *arXiv preprint arXiv:1810.13069*, 2018.

[3] Daniel McFadden. Modeling the choice of residential location. *Transportation Research Record*, (673), 1978.