Thompson Sampling for Multinomial Logit Contextual Bandits

Min-hwan Oh and Garud Iyengar
m.oh@columbia.edu

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: Netflix, Youtube, etc.
  - News websites/feeds, web searches and many more
- Contextual information 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 $t = 1,...,T$
  1. Context $x_t \in \mathbb{R}^d$ and revenue $r_t$ revealed for all items $i \in [N]$
  2. Agent selects assortment $S_t \subset [N]$ (with $|S_t| \leq K$)
  3. Agent observes user choice $y_t \in \{0,1\}^{|S_t|}$
- Choice given by multinomial logit (MNL) model $p(S_t, \theta^*)$
  - Probability that user chooses $i \in S_t$ is $\frac{\text{utility}(i)}{1 + \sum_{j \in S_t} \exp(x_t^j \theta^*)}$
  - $\theta^* \in \mathbb{R}^d$ unknown true parameter
- Expected revenue for assortment $S_t$: $R_t(S_t, \theta^*) = \sum_{i \in S_t} r_t \cdot p(i|S_t, \theta^*)$
- Goal: minimize total regret

$$\text{Regret}(T) = \mathbb{E} \left[ \sum_{t=1}^T R_t(S_t, \theta^*) - \sum_{t=1}^T R_t(S_t, \theta^*) \right]$$

where $S_t^* = \arg \max_{S_t} R_t(S_t, \theta^*)$

TS-MNL with Optimistic Sampling

- Sample from Gaussian distribution
  - TS as generic randomized algorithm based on MLE $\hat{\theta}_t$
  - $\hat{\theta}_t \sim \mathcal{N} (\hat{\theta}_t, \alpha^{-1} V_t^{-1})$
- Sample $\tilde{\theta}_t \sim \mathcal{N} (\hat{\theta}_t, \alpha^{-1} V_t^{-1})$ and $\bar{\theta}_t = \max \{ \tilde{\theta}_t, \theta^* \}$
- Expected revenue of assortment $S$ based on $\bar{\theta}_t$:

$$\tilde{R}_t(S) = \sum_{i \in S} r_i \frac{\exp(x_t^i \bar{\theta}_t)}{1 + \sum_{j \in S} \exp(x_t^j \bar{\theta}_t)}$$

Challenges in Worst-Case Regret Analysis

- Decomposing worst-case immediate regret:

$$\text{Regret}(t) = \mathbb{E} \left[ R_t(S_t^*, \theta^*) - R_t(S_t, \bar{\theta}_t) \right]$$

- (b) controlled by concentration of $\tilde{\theta}_t$
  - In Bayesian regret, (a) $= 0$ since $\tilde{\theta}_t$ and $\bar{\theta}_t$ are iid
  - Probability each utility is optimistic exponentially small in $K$

Theorem (Bayesian Regret)

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

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

Lemma (Ensuring Optimism)

Let $\alpha = \mathcal{O}(\sqrt{d \log(1+1/d)})$ and take $M = [1 + C \log K]$ samples for some constant $C$. Then $\mathbb{P}(\hat{R}_t(S_t^*) > R_t(S_t, \theta^*), |F_t|) \geq 1 - \frac{1}{4e^2}$

Theorem (Worst-Case Regret)

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

- Matches regret bound for linear TS bands [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., $N < e^d$, $\mathcal{O}(d^{3/2} \sqrt{T} \log \log T)$ 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