

Queueing Matching Bandits with Preference Feedback

Introduction

- Ride-hailing platforms where riders are assigned to drivers.
- •Online labor service markets where tasks are recommended to freelance workers.

Motivation examples include

Problem Statement

- There are N agents (queues) and K arms (servers).
- •At each time, a job for each agent $n \in [N]$ arrives randomly following a Bernoulli distribution with an unknown arrival rate $\lambda_n \in [0, 1]$.
- At each time, a scheduler assigns agents to arms, $\{S_{k,t}\}_{k\in[K]}$.
- Each agent *n* has known *d*-dimensional feature information of $x_n \in \mathbb{R}^d$, and each arm k has latent (unknown) parameter $\theta_k \in \mathbb{R}^d$. Then we adopt the Multi-nomial Logit (MNL) for preference feedback (service rate) as

•Let $Q_n(t)$ be the length of the queue for jobs of agent beginning of time slot t in the system. Queue length evolves as

Objective Function. Here we provide the goal of this problem. For analyzing the stability of the systems, we define the average queue

Jung-hun $Kim¹$ and Min-hwan $Oh¹$ ¹Seoul National University

$$
\mu(n|S_{k,t}, \theta_k) = \frac{\exp(x_n^\top \theta_k)}{1 + \sum_{m \in S_{k,t}} \exp(x_m^\top \theta_k)}
$$

$$
\overline{r_k}
$$
nt
$$
n \in [N]
$$
 at the
$$
r_k
$$
th of the agent
$$
n_k
$$

$$
Q_n(t+1) = (Q_n(t) + A_n(t) - D_n(t|S_{k_{n,t},t}))^+.
$$

where $h_{n,k,t}^{UCB} = x_n^{\top}$ \overline{n} $\hat{\theta}$ $\theta_{k,t} + \beta_t \|x_n\|_V$ −1 k,t . $[Q_n(t)]$.

lengths over horizon time T as

$$
\mathcal{Q}(T) = \frac{1}{T} \sum_{t \in [T]} \sum_{n \in [N]} \mathbb{E}\left[\right]
$$

Definition 1. The systems are denoted to be stable when \bullet $\lim_{T\to\infty}\mathcal{Q}(T)<\infty.$

Assumption 1. (Traffic Slackness) For some traffic slackness $\mathcal{Q}(T) = O$ $0 < \epsilon < 1$, there exists $\{S_k\}_{k\in[K]} \in \mathcal{M}([N])$ such that this set ri satisfies $\lambda_n + \epsilon \leq \mu(n|S_k, \theta_k)$ for all $n \in S_k$ and $k \in [K]$.

Oracle We denote the oracle assignment

 $\{S_k^*$ $\{k,t\}_{{k\in[K]}} = \mathop{\rm argmax}$ $\{S_k\}_{k\in[K]}\in\mathcal{M}(\mathcal{N}_t)$ \sum $k \in [K]$

Proposition 1. Given the prior knowledge of θ_k for all $k \in [K]$, the average queue length of $MaxWeight$ is $\mathcal{Q}(T) = O$ $\left\{\min\{N,K\}\right\}$ ϵ \setminus

 $\mathcal{Q}(T) = O$ rithm achieves stability as

$$
\epsilon[\kappa] = \mathop{\arg\max}_{\{S_k\}_{k\in[K]}\in\mathcal{M}(\mathcal{N}_i)} \sum_{k\in[K]} \sum_{n\in S_k} Q_n(t) \widetilde{\mu}_t^{UCB}(n|S_k, \hat{\theta}_{k,t}).
$$

\nference feedback $y_{n,l} \in \{0, 1\}$ for all $n \in S_{k,l}$ and $k \in [K]$.
\n. The average queue length of UCB-QMB is bounded as $\frac{\inf\{N,K\}}{\epsilon} + \frac{d^2N^2K^2\text{polylog}(T)}{T}$, which implies that the algo-
\nes stability as $\lim_{T\to\infty} \mathcal{Q}(T) = O\left(\frac{\min\{N,K\}}{\epsilon}\right)$.
\nThe policy π of UCB-QMB achieves a regret bound of $= \widetilde{O}\left(\min\left\{\frac{d}{\kappa}\sqrt{KT}Q_{\text{max}}, N\left(\frac{dK}{\kappa^2c^3}\right)^{1/4}T^{3/4}\right\}\right)$.
\n**Algorithm (TS-QMB)** We utilize Thompson Sam-
\nxxWeight as $|\kappa| \leftarrow \underset{\{S_k\}_{k\in[K]}}{\operatorname{argmax}} \sum_{n\in S_k} Q_n(t) \widetilde{\mu}_t^{TS}(n|S_k, \{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M]})$.
\nThe average queue length of TS-QMB is bounded as $\frac{\inf\{N,K\}}{\kappa^4\kappa^8} + \frac{d^2N^2K^2\text{polylog}(T)}{T}$, which implies that the algo-
\nes stability as $\lim_{T\to\infty} \mathcal{Q}(T) = O\left(\frac{\min\{N,K\}}{\epsilon}\right)$.
\nThe policy π of TS-QMB achieves a regret bound of $\tilde{O}\left(\min\left\{\frac{d^{3/2}}{\kappa}\sqrt{KT}Q_{\text{max}}, N\left(\frac{d^2K}{\kappa^2\epsilon^3}\right)^{1/4}T^{3/4}\right\}\right)$.
\n**ments**

ments (MaxWeight) as

\n
$$
\sum_{n \in S_k} Q_n(t) \mu(n|S_k, \theta_k).
$$
\nwledge of θ_k for all $k \in$

\nit is bounded as

.

 $\log \mu(n|S_{k,t},\theta),$

 $\theta_{k,t}$ from minimizing the negative log-likelihood

• At each time t , UCB-QMB offers a set of assortments as

Algorithms & Analyses

UCB-based Algorithm (UCB-QMB)

•We define the negative log-likelihood as

$$
f_{k,t}(\theta) := - \sum_{n \in S_{k,t} \cup \{n_0\}} y_{n,t} \log n
$$

where $y_{n,t} \in \{0,1\}$ is observed preference feedback.

•Construct estimator $\hat{\theta}$ by applying online newton step as

$$
\hat{\theta}_{k,t} = \underset{\theta \in \Theta}{\operatorname{argmin}} \nabla f_{k,t-1}(\hat{\theta}_{k,t-1})^{\top} \theta + \frac{1}{2} ||\theta - \hat{\theta}_{k,t-1}||_{V_{k,t}}^2.
$$

Let UCB index for agent $n \in S_k$ as

$$
\widetilde{\mu}_t^{UCB}(n|S_k, \hat{\theta}_{k,t}) = \frac{\exp(h_{n,k,t}^{UCB})}{1 + \sum_{m \in S_k} \exp(h_{m,k,t}^{UCB})},
$$

 \bullet Construc

$$
= \underset{\theta \in \Theta}{\operatorname{argmin}} \nabla f_{k,t-1}(\hat{\theta}_{k,t-1})^{\top} \theta + \frac{1}{2} ||\theta - \hat{\theta}_{k,t-1}||_{V_{k,t}}^2.
$$

CB index for agent $n \in S_k$ as

$$
\widetilde{\mu}_t^{UCB}(n|S_k, \hat{\theta}_{k,t}) = \frac{\exp(h_{n,k,t}^{UCB})}{1 + \sum_{m \in S_k} \exp(h_{m,k,t}^{UCB})},
$$

rithm achieves stability as

$$
\begin{aligned}\n&\underset{\{S_k\}_{k\in[K]}\in\mathcal{M}(\mathcal{N}_t)\}}{\operatorname{argmax}\sum_{\{S_k\}_{k\in[K]}}\sum_{n\in S_k}Q_n(t)\widetilde{\mu}_t^{UCB}(n|S_k,\hat{\theta}_{k,t}).\\
&\text{e feedback } y_{n,t}\in\{0,1\}\text{ for all }n\in S_{k,t}\text{ and }k\in[K].\\
&\text{average queue length of UCB-QMB is bounded as}\\
&\underset{\mathcal{N}\to\infty}{\text{if $\frac{\lambda^2N^2K^2\text{polylog}(T)}{T}$}}\text{, which implies that the algorithm as}\\
&\underset{\mathcal{N}\to\infty}{\text{lim}}\mathcal{Q}(T)=O\left(\frac{\min\{N,K\}}{\epsilon}\right).\n&\text{policy }\pi\text{ of UCB-QMB achieves a regret bound of}\\
&\left(\min\left\{\frac{d}{\kappa}\sqrt{KT}Q_{\max},N\left(\frac{dK}{\kappa^2\epsilon^3}\right)^{1/4}T^{3/4}\right\}\right).\n&\text{orit him (TS-QMB) We utilize Thompson Sam-ght as}\\
&\underset{\{S_k\}_{k\in[K]}}{\text{argmax}}\sum_{k\in[K]}\sum_{n\in S_k}Q_n(t)\widetilde{\mu}_t^{TS}(n|S_k,\{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M]}).\n&\text{average queue length of TS-QMB is bounded as}\\
&\underset{\mathcal{N}\to\infty}{\text{if $\frac{\lambda^4N^2K^2\text{polylog}(T)}{T}$}}\text{, which implies that the algorithm as}\\
&\underset{\mathcal{N}\to\infty}{\text{using Q}}(T)=O\left(\frac{\min\{N,K\}}{\epsilon}\right).\n&\text{points of TS-QMB achieves a regret bound of}\\
&\underset{\mathcal{N}\to\infty}{\text{min }\left\{\frac{d^{3/2}}{\kappa}\sqrt{KT}Q_{\max},N\left(\frac{d^2K}{\kappa^2\epsilon^3}\right)^{1/4}T^{3/4}\right\}\right).\n&\text{and}\\
&\underset{\mathcal{N}\to\infty}{\text{in this}}\n\end{aligned}
$$

$$
\{S_{k,t}\}_{k\in[K]} = \mathop{\arg\max}_{\{S_k\}_{k\in[K]}} \sum_{\kappa \in M(N)} \sum_{k\in[K]} \sum_{n\in S_k} Q_n(t) \widetilde{\mu}_k^{UCB}(n|S_k, \hat{\theta}_{k,t}).
$$

prove preference feedback
$$
y_{n,t} \in \{0, 1\}
$$
 for all $n \in S_{k,t}$ and $k \in [K]$.
rem 1. The average queue length of UCB-QMB is bounded as

$$
= O\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d^2N^2K^2 \text{polylog}(T)}{T}\right), \text{ which implies that the algo-achieves stability as}
$$

$$
\lim_{T \to \infty} Q(T) = O\left(\frac{\min\{N,K\}}{\epsilon}\right).
$$

rem 2. The policy π of UCB-QMB achieves a regret bound of
 $\mathcal{R}^{\pi}(T) = \widetilde{O}\left(\min\left\{\frac{d}{\kappa}\sqrt{KT}Q_{\text{max}}, N\left(\frac{dK}{\kappa^2\epsilon^3}\right)^{1/4} T^{3/4}\right\}\right).$
based Algorithm (TS-QMB) We utilize Thompson Sam-
with Maxwell as

$$
S_{k,t}\}_{k\in[K]} \leftarrow \mathop{\arg\max}_{\{S_k\}_{k\in[K]}} \sum_{k\in[K]} \sum_{n\in S_k} Q_n(t) \widetilde{\mu}_t^{TS}(n|S_k, \{\widetilde{\theta}_{k,t}^{(i)}\}_{i\in[M]}).
$$

rem 3. The average queue length of TS-QMB is bounded as

$$
= O\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d^4N^2K^2 \text{polylog}(T)}{\tau}\right), \text{ which implies that the algo-archieves stability as}
$$

$$
\lim_{T \to \infty} Q(T) = O\left(\frac{\min\{N,K\}}{\epsilon}\right).
$$

rem 4. The policy π of TS-QMB achieves a regret bound of

$$
\mathcal{R}^{\pi}(T) = \widetilde{O}\left(\min\left\{\frac{d^{3/2}}{\kappa}\sqrt{KT}Q_{\text{max}}, N\left(\frac{d^2K}{\kappa^2\epsilon^3}\right)^{1/4} T^{3/4}\right\}\right).
$$

$$
\{S_{k,t}\}_{k\in[K]} = \underset{\{S_{i}\}_{k\in[N]}}{\operatorname{argmax}} \sum_{\{S_{i}\}_{k\in[N]}} \sum_{h\in[K]} \sum_{n\in S_{k}} Q_{n}(t) \tilde{\mu}_{t}^{UCB}(n|S_{k}, \hat{\theta}_{k,t}).
$$

\nserve preference feedback $y_{n,t} \in \{0, 1\}$ for all $n \in S_{k,t}$ and $k \in [K]$.
\n**orem 1.** The average queue length of UCB-QMB is bounded as
\n
$$
P = O\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d^{CV}_{k}K^{V}\text{polylog}(T)}{T}\right), which implies that the algo-\nn achieves stability as\n
$$
\lim_{T\to\infty} Q(T) = O\left(\frac{\min\{N,K\}}{\epsilon}\right).
$$

\n**orem 2.** The policy π of UCB-QMB achieves a regret bound of
\n
$$
\mathcal{R}^{\pi}(T) = \tilde{O}\left(\min\left\{\frac{d}{\kappa}\sqrt{KT}Q_{\max}, N\left(\frac{dK}{\kappa^{2}\epsilon^{3}}\right)^{1/4}T^{3/4}\right\}\right).
$$

\n**based Algorithm (TS-QMB)** We utilize Thompson Sam-
\n
$$
\text{with Maxwell as}
$$

\n
$$
\{S_{k,t}\}_{k\in[K]} \leftarrow \underset{\{S_{k}\}_{k\in[K]}}{\operatorname{argmax}} \sum_{n\in[K]} \sum_{n\in S_{k}} Q_{n}(t) \tilde{\mu}_{t}^{TS}(n|S_{k}, \{\tilde{\theta}_{k,t}^{(i)}\}_{i\in[M]}).
$$

\n**orem 3.** The average queue length of TS-QMB is bounded as
\n
$$
P = O\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d^{N}N^{2}\epsilon^{2}\text{polylog}(T)}{\epsilon}\right), \text{ which implies that the algo-\nn achieves stability as}\n
$$
\lim_{T\to\infty} Q(T) = O\left(\frac{\min\{N,K\}}{\epsilon}).
$$

\n**orem 4.** The policy π of TS-QMB achieves a regret bound of
\n
$$
\mathcal{R}^{\pi}(T) = \tilde{O}\left(\min\left\{\frac{d^{
$$
$$
$$

$$
\{S_{k,t}\}_{k\in[K]} = \underset{\{S_{k}\}_{k\in[K]}}{\operatorname{argmax}} \sum_{\{S_{k}\}_{k\in[K]}} \sum_{K\in[K]} \sum_{\{G_{l}\}_{l} \in S_{k}} Q_{n}(t) \widetilde{\mu}_{k}^{UCB}(n|S_{k}, \hat{\theta}_{k,t}).
$$
\n
$$
\bullet
$$
 Observe preference feedback $y_{n,t} \in \{0,1\}$ for all $n \in S_{k,t}$ and $k \in [K]$.
\n**Theorem 1.** The average queue length of UCB-QMB is bounded as
\n $Q(T) = O\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d'N'K'\text{polylog}(T)}{T}\right)$, which implies that the algorithm achieves stability as
\n
$$
\lim_{T \to \infty} Q(T) = O\left(\frac{\min\{N,K\}}{\epsilon}\right).
$$
\n**Theorem 2.** The policy π of UCB-QMB achieves a regret bound of
\n $\mathcal{R}^{\pi}(T) = \widetilde{O}\left(\min\left\{\frac{d}{\kappa}\sqrt{KT}Q_{\max}, N\left(\frac{dK}{\kappa^{2}\epsilon^{3}}\right)^{1/4}T^{3/4}\right\}\right).$ \n**TS-based Algorithm (TS-QMB)** We utilize Thompson Sam-
\npling with Maxwell as
\n
$$
\{S_{k,t}\}_{k\in[K]} \leftarrow \underset{\{S_{k}\}_{k\in[K]}}{\operatorname{argmax}} \sum_{K \in [K]} \sum_{n \in S_{k}} Q_{n}(t) \widetilde{\mu}_{t}^{TS}(n|S_{k}, \{\widehat{\theta}_{k,t}^{(i)}\}_{i\in[M]}).
$$
\n**Theorem 3.** The average queue length of TS-QMB is bounded as
\n $Q(T) = O\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d^4N^2K^2\text{polylog}(T)}{\epsilon}\right)$, which implies that the algorithm achieves stability as
\n
$$
\lim_{T \to \infty} Q(T) = O\left(\frac{\min\{N,K\}}{\epsilon}\right).
$$
\n**Theorem 4.** The policy π of TS-QMB achieves a regret bound of
\n $\mathcal{R}^{\pi}(T) = \widetilde{O}\left(\min\left\{\frac{d^{3/2}}{\kappa}\sqrt{KT}Q_{\max}, N\$

$$
\{S_{k,t}\}_{k\in[K]} = \underset{\{S_k\}_{k\in[K]}}{\operatorname{argmax}} \sum_{\{S_k\}_{k\in[K]}} \sum_{\kappa\in[K]} Q_n(t) \widetilde{\mu}_t^{UCB}(n|S_k, \hat{\theta}_{k,t}).
$$

\nserve preference feedback $y_{n,t} \in \{0, 1\}$ for all $n \in S_{k,t}$ and $k \in [K]$.
\norem 1. The average queue length of UCB-QMB is bounded as
\n $= O\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d^2N^2K^2\text{polylog}(T)}{T}\right)$, which implies that the algo-
\na achieves stability as
\n $\lim_{T \to \infty} Q(T) = O\left(\frac{\min\{N, K\}}{\epsilon}\right)$.
\norem 2. The policy π of UCB-QMB achieves a regret bound of
\n $\mathcal{R}^{\pi}(T) = \widetilde{O}\left(\min\left\{\frac{d}{\kappa}\sqrt{KT}Q_{\text{max}}, N\left(\frac{dK}{\kappa^2\epsilon^3}\right)^{1/4} T^{3/4}\right\}\right)$.
\nbased Algorithm (TS-QMB) We utilize Thompson Sam-
\nwith Maxwell as
\n $\{S_{k,t}\}_{k \in [K]} \leftarrow \underset{\{S_k\}_{k \in K}}{\operatorname{argmax}} \sum_{\{K \in K\}} Q_n(t) \widetilde{\mu}_t^{TS}(n|S_k, \{\widetilde{\theta}_{k,t}^{(i)}\}_{i \in [M]})$.
\norem 3. The average queue length of TS-QMB is bounded as
\n $= O\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d^2N^2K^2\text{polylog}(T)}{T}\right)$, which implies that the algo-
\n*i* achieves stability as
\n $\lim_{T \to \infty} Q(T) = O\left(\frac{\min\{N,K\}}{\epsilon}\right)$.
\norem 4. The policy π of TS-QMB achieves a regret bound of
\n $\mathcal{R}^{\pi}(T) = \widetilde{O}\left(\min\left\{\frac{d^3/2}{\kappa}\sqrt{KT}Q_{\text{max}}, N\left(\frac{d^2K}{\kappa^2\epsilon^3}\right)^{1/4}\right\}\right)$.
\n**Definitionents**

$$
S_{k,l} \}_{k \in [K]} = \mathop{\arg\max}_{\{S_k\}_{k \in [K]} \in \mathcal{M}(N_i)} \sum_{k \in [K]} \sum_{n \in S_k} Q_n(t) \tilde{\mu}_t^{UCB}(n|S_k, \hat{\theta}_{k,l}).
$$

we preference feedback $y_{n,t} \in \{0, 1\}$ for all $n \in S_{k,t}$ and $k \in [K]$.
em 1. The average queue length of UCB-QMB is bounded as

$$
= O\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d^2N^2K^2 \text{polylog}(T)}{\epsilon^2 T}\right), \text{ which implies that the algo-schieves stability as}
$$

$$
\lim_{T \to \infty} Q(T) = O\left(\frac{\min\{N, K\}}{\epsilon}\right).
$$

em 2. The policy π of UCB-QMB achieves a regret bound of
 $\mathbb{R}^{\pi}(T) = \widetilde{O}\left(\min\left\{\frac{d}{\kappa}\sqrt{KT}Q_{\max}, N\left(\frac{dK}{\kappa^2\epsilon^3}\right)^{1/4}T^{3/4}\right\}\right).$
ased Algorithm (TS-QMB) We utilize Thompson Sam-
ith Maxwell as
 $\sum_{\{S_k\}_{k \in [K]}} \leftarrow \mathop{\arg\max}_{\{S_k\}_{k \in [K]}} \sum_{n \in S_k} Q_n(t) \tilde{\mu}_t^{TS}(n|S_k, \{\widetilde{\theta}_{k,t}^{(i)}\}_{i \in [M]}).$
em 3. The average queue length of TS-QMB is bounded as

$$
= O\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d^2N^2K^2 \text{polylog}(T)}{\epsilon^2 T}\right), \text{ which implies that the algo-schieves stability as}
$$

$$
\lim_{T \to \infty} Q(T) = O\left(\frac{\min\{N, K\}}{\epsilon}).
$$

em 4. The policy π of TS-QMB achieves a regret bound of
 $\pi(T) = \widetilde{O}\left(\min\left\{\frac{d^{3/2}}{\kappa}\sqrt{KT}Q_{\max}, N\left(\frac{d^2K}{\kappa^2\epsilon^3}\right)^{1/4}T^{3/4}\right\}\right).$

pling with MaxWeight as

Experiments

