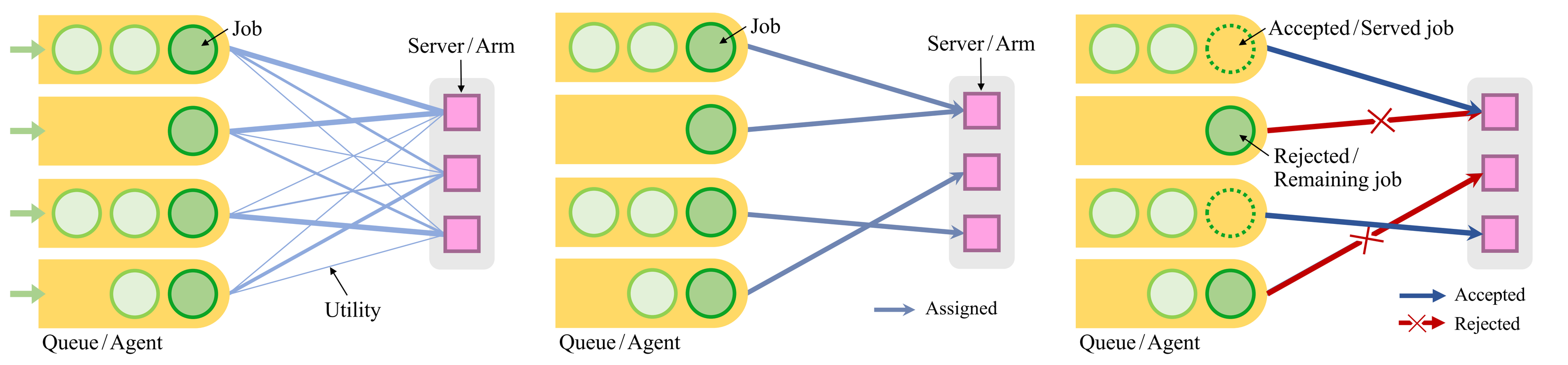
Queueing Matching Bandits with Preference Feedback

Jung-hun Kim[1] and Min-hwan Oh[1]

[1]*Seoul National University*

Introduction



Motivation examples include

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

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

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

- Let $Q_n(t)$ be the length of the queue for jobs of agent $n \in [N]$ at the beginning of time slot t in the system. Queue length of the agent n evolves as

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

Objective Function. Here we provide the goal of this problem. For analyzing the stability of the systems, we define the average queue lengths over horizon time T as

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

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

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

Oracle We denote the oracle assignments (MaxWeight) as

$$\{S^*_{k,t}\}_{k\in[K]} = \underset{\{S_k\}_{k\in[K]}\in\mathcal{M}(\mathcal{N}_t)}{\operatorname{argmax}} \sum_{k\in[K]}\sum_{n\in S_k} Q_n(t)\mu(n|S_k,\theta_k).$$

Proposition 1. *Given the prior knowledge of θ_k for all $k\in[K]$, the average queue length of MaxWeight is bounded as*

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

Algorithms & Analyses

UCB-based Algorithm (`UCB-QMB`)

- We define the negative log-likelihood as

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

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

- Construct estimator $\hat\theta_{k,t}$ from minimizing the negative log-likelihood by applying online newton step as

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

- Construct UCB index for agent $n\in S_k$ as

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

where $h^{UCB}_{n,k,t} = x_n^\top\hat\theta_{k,t} + \beta_t\|x_n\|_{V^{-1}_{k,t}}$.

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

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

- Observe preference feedback $y_{n,t}\in\{0,1\}$ for all $n\in S_{k,t}$ and $k\in[K]$.

Theorem 1. *The average queue length of `UCB-QMB` is bounded as $\mathcal{Q}(T) = O\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d^2N^2K^2\mathrm{polylog}(T)}{\kappa^4\epsilon^6 T}\right)$, which implies that the algorithm achieves stability as*

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

Theorem 2. *The policy π of `UCB-QMB` achieves a regret bound of*

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

TS-based Algorithm (`TS-QMB`) We utilize Thompson Sampling with MaxWeight as

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

Theorem 3. *The average queue length of `TS-QMB` is bounded as $\mathcal{Q}(T) = O\left(\frac{\min\{N,K\}}{\epsilon} + \frac{d^4N^2K^2\mathrm{polylog}(T)}{\kappa^4\epsilon^6 T}\right)$, which implies that the algorithm achieves stability as*

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

Theorem 4. *The policy π of `TS-QMB` achieves a regret bound of*

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

Experiments