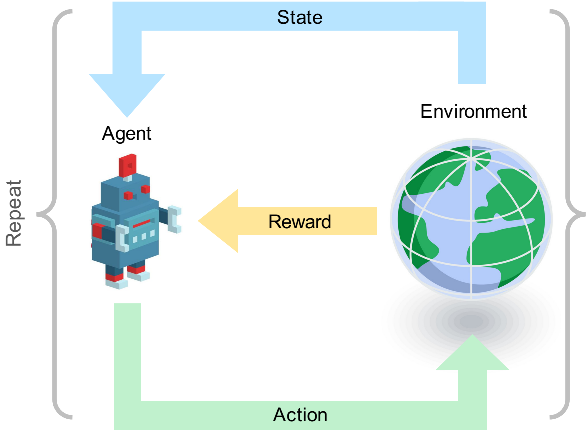
Combinatorial Reinforcement Learning with Preference Feedback

(ICML 2025)

Joongkyu Lee & Min-hwan Oh

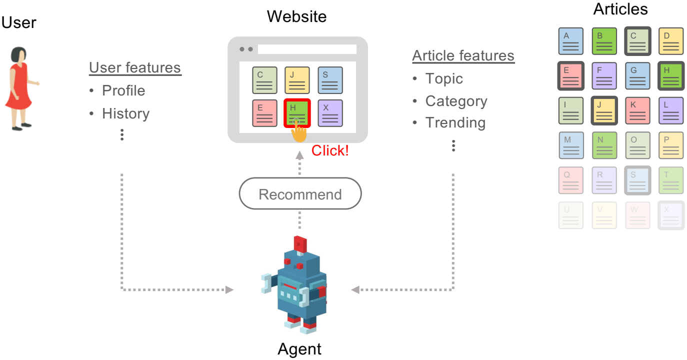
Seoul National University

Standard Reinforcement Learning



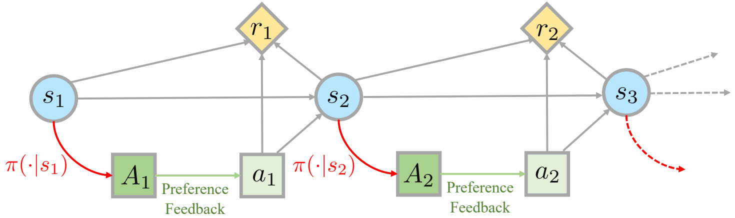
Goal: Maximize sum of future rewards

Sequential Assortment Selection Problem



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

Combinatorial RL with Preference Feedback



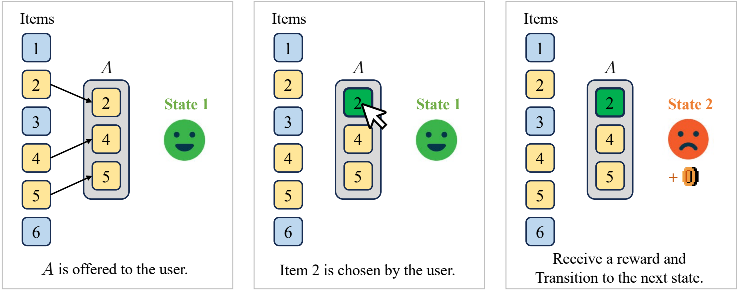
MDPs: $\mathcal{M}(\mathcal{S}, \mathcal{I}, \mathcal{A}, M, \{\mathcal{P}_h\}_{h=1}^{H}, \{P_h\}_{h=1}^{H}, \{r_h\}_{h=1}^{H}, H)$

- $\mathcal{S}$: State space
- $\mathcal{I}$: Set of items (base actions)
- $\mathcal{A}$: Set of (super) actions
- M: Maximum size of item set
- $\mathcal{P}_h$: Preference choice model
- P_h: Transition probability
- r_h: Reward function
- H: Length of each episode

Combinatorial RL with Preference Feedback: Why RL?

Why do we use RL for preference-based feedback systems?

$\Longrightarrow$ To optimize recommendations for <u>long-term user engagement</u>



Combinatorial RL with Preference Feedback: Challenges

1. **Combinatorial action space**: Offering set of items

Combinatorial RL with Preference Feedback: Challenges

1. **Combinatorial action space**: Offering set of items
2. **Preference feedback**: Receive feedback only for chosen item

Combinatorial RL with Preference Feedback: Challenges

1. **Combinatorial action space**: Offering set of items
2. **Preference feedback**: Receive feedback only for chosen item
3. **No theoretical guarantees**:
   - Empirical results: Swaminathan et al. (2017); Ie et al. (2019); McInerney et al. (2020); Vlassis et al. (2021); Chaudhari et al. (2024)

Value Functions & Objective

- **Value function of a policy** π

$$V_h^\pi(s) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, s_{h'+1}) \,|\, s_h = s\right]$$

$$Q_h^\pi(s, A) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, s_{h'+1}) \,|\, s_h = s, A_h = A\right]$$

$$= \sum_{a \in A} \mathcal{P}_h(a|s, A) \underbrace{\left(\sum_{s' \in \mathcal{S}} P_h(s'|s, a)\left(r_h(s, a, s') + V_{h+1}^\pi(s')\right)\right)}_{=:\overline{Q}_h^\pi(s,a)}$$

  - $\mathcal{P}_h$: Preference model, P_h: Transition, r_h: Reward
  - $\overline{Q}_h^\pi(s, a)$: Item-level Q-value function

Value Functions & Objective

- **Value function of a policy π**

$$V_h^\pi(s) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, s_{h'+1}) \,\middle|\, s_h = s\right]$$

$$Q_h^\pi(s, A) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, s_{h'+1}) \,\middle|\, s_h = s, A_h = A\right]$$

$$= \sum_{a \in A} \mathcal{P}_h(a|s, A) \underbrace{\left(\sum_{s' \in \mathcal{S}} P_h(s'|s, a)\left(r_h(s, a, s') + V_{h+1}^\pi(s')\right)\right)}_{=: \overline{Q}_h^\pi(s,a)}$$

  - $\mathcal{P}_h$: Preference model, P_h: Transition, r_h: Reward
  - $\overline{Q}_h^\pi(s, a)$: Item-level Q-value function

- **Optimal value function & policy**

$$V_h^*(s) = \sup_\pi V_h^\pi(s), \quad Q_h^*(s, A) = \sup_\pi Q_h^\pi(s, A), \quad \pi_h^*(s) = \arg\max_{A \in \mathcal{A}} Q_h^*(s, A)$$

Value Functions & Objective

- **Value function of a policy π**

$$V_h^\pi(s) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, s_{h'+1}) \,\middle|\, s_h = s\right]$$

$$Q_h^\pi(s, A) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, s_{h'+1}) \,\middle|\, s_h = s, A_h = A\right]$$

$$= \sum_{a \in A} \mathcal{P}_h(a|s, A) \underbrace{\left(\sum_{s' \in \mathcal{S}} P_h(s'|s, a)\left(r_h(s, a, s') + V_{h+1}^\pi(s')\right)\right)}_{=:\overline{Q}_h^\pi(s,a)}$$

  - $\mathcal{P}_h$: Preference model, P_h: Transition, r_h: Reward
  - $\overline{Q}_h^\pi(s, a)$: Item-level Q-value function

- **Optimal value function & policy**

$$V_h^*(s) = \sup_\pi V_h^\pi(s), \quad Q_h^*(s, A) = \sup_\pi Q_h^\pi(s, A), \quad \pi_h^*(s) = \arg\max_{A \in \mathcal{A}} Q_h^*(s, A)$$

- **Objective**: Minimize $\mathbf{Reg}(K) = \sum_{k=1}^{K} \left(V_1^* - V_1^{\pi_k}\right)(s_{k,1})$

MNL Preference & General Item-Level Q-value

Definition of Q-value

$$Q_h^\pi(s, A) := \sum_{a \in A} \mathcal{P}_h(a|s, A)\overline{Q}_h^\pi(s, a)$$

- Preference model: **Multinomial Logit (MNL) Model** (McFadden, 1977)

  $$\mathcal{P}_h(a|s, A) = \frac{\exp\left(\phi(s, a)^\top \boldsymbol{\theta}_h^\star\right)}{\sum_{a' \in A} \exp\left(\phi(s, a')^\top \boldsymbol{\theta}_h^\star\right)},$$

  where $\phi(s, a) \in \mathbb{R}^d$ is feature vector and $\boldsymbol{\theta}_h^\star \in \mathbb{R}^d$ is unknown parameter

- Item-level Q-value $\overline{Q}_h^\pi(s, a)$: **General function approximation**
  - Can be any function such as neural network, transformer, ...

Main Result

Regret bound

With high probability, the regret of our proposed algorithm is

$$\mathbf{Reg}(K) = \widetilde{\mathcal{O}}\Bigg(\underbrace{d\sqrt{HK}}_{\text{regret from } \mathcal{P}_h} + \underbrace{\sqrt{d_\nu HK \log \mathcal{N}}}_{\text{regret from } \overline{Q}_h^\pi} \Bigg),$$

where d_ν: complexity of function class, $\mathcal{N}$: cardinality of function class.

1. **First theoretical guarantee** in combinatorial RL with preference feedback

Main Result

Regret bound

With high probability, the regret of our proposed algorithm is

$$\mathbf{Reg}(K) = \widetilde{\mathcal{O}}\Bigg(\underbrace{d\sqrt{HK}}_{\text{regret from } \mathcal{P}_h} + \underbrace{\sqrt{d_\nu HK \log \mathcal{N}}}_{\text{regret from } \overline{Q}_h^\pi} \Bigg),$$

where d_ν: complexity of function class, $\mathcal{N}$: cardinality of function class.

1. **First theoretical guarantee** in combinatorial RL with preference feedback
2. **Computationally efficient algorithm**:
   - Avoid <u>combinatorial optimization</u> when computing $\arg\max_{A \in \mathcal{A}} Q_h^k(s, A)$
   - Convert the optimization problem into linear programming (LP) problem

Main Result

Regret bound

With high probability, the regret of our proposed algorithm is

$$\textbf{Reg}(K) = \widetilde{\mathcal{O}}\Bigg(\underbrace{d\sqrt{HK}}_{\text{regret from } \mathcal{P}_h} + \underbrace{\sqrt{d_\nu HK \log \mathcal{N}}}_{\text{regret from } \overline{Q}_h^\pi} \Bigg),$$

where d_ν: complexity of function class, $\mathcal{N}$: cardinality of function class.

1. **First theoretical guarantee** in combinatorial RL with preference feedback
2. **Computationally efficient algorithm**:
   - Avoid combinatorial optimization when computing $\arg\max_{A \in \mathcal{A}} Q_h^k(s, A)$
   - Convert the optimization problem into linear programming (LP) problem
3. When reduced to MNL bandit ($H = 1$), it is **minimax optimal!**
4. In linear MDPs, we obtain $\widetilde{\mathcal{O}}(d\sqrt{HK} + d^{lin}\sqrt{HK})$, which is **minimax optimal**, matching the established lower bound.

Experiment 1: Synthetic Environment

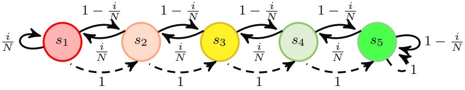


Figure: The “online shopping with budget” environment with $|\mathcal{S}| = 5$.

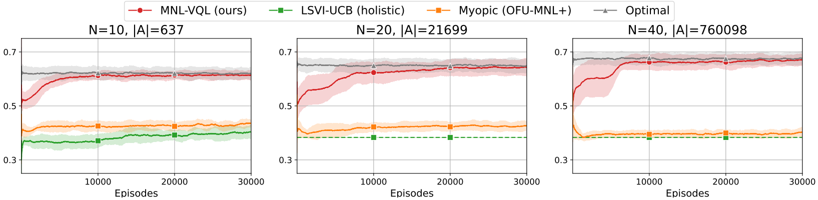


Figure: Episodic returns on “online shopping with budget”.

Experiment 2: Real-World Environment

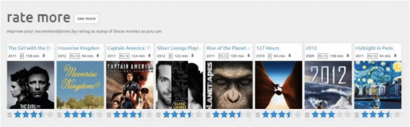


Figure: MovieLens Dataset

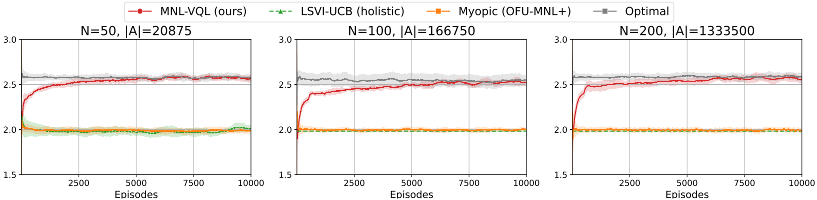


Figure: Episodic returns on “MovieLens experiment”.

References I

Chaudhari, S., Arbour, D., Theocharous, G., and Vlassis, N. (2024). Distributional off-policy evaluation for slate recommendations. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 8265–8273.

Ie, E., Jain, V., Wang, J., Narvekar, S., Agarwal, R., Wu, R., Cheng, H.-T., Chandra, T., and Boutilier, C. (2019). Slateq: A tractable decomposition for reinforcement learning with recommendation sets. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pages 2592–2599.

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

McInerney, J., Brost, B., Chandar, P., Mehrotra, R., and Carterette, B. (2020). Counterfactual evaluation of slate recommendations with sequential reward interactions. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1779–1788.

Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudik, M., Langford, J., Jose, D., and Zitouni, I. (2017). Off-policy evaluation for slate recommendation. Advances in Neural Information Processing Systems, 30.

Vlassis, N., Chandrashekar, A., Amat, F., and Kallus, N. (2021). Control variates for slate off-policy evaluation. Advances in Neural Information Processing Systems, 34:3667–3679.