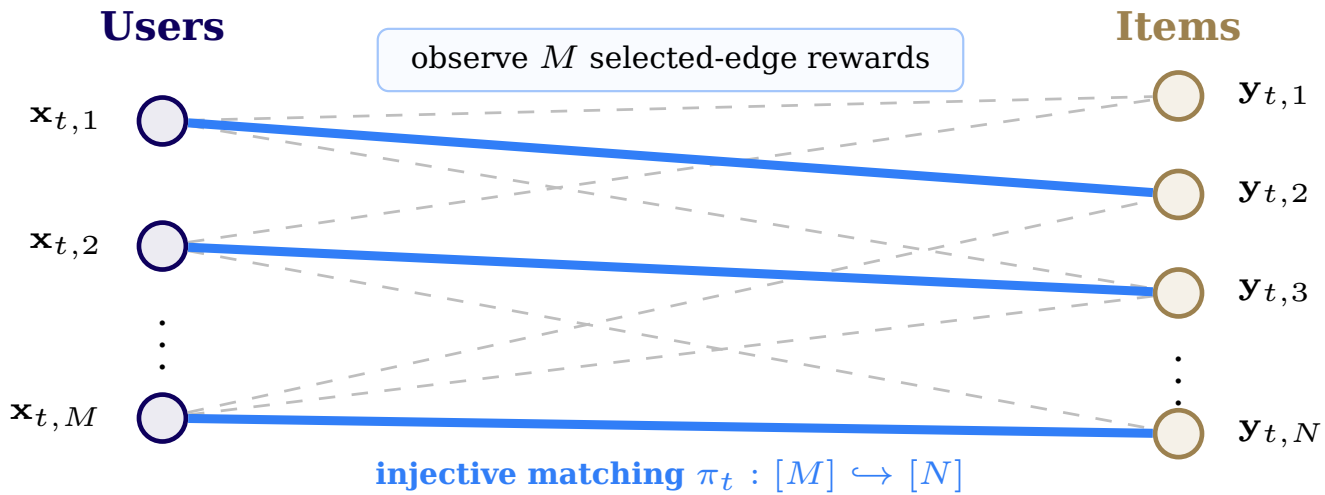
Bilinear Matching Bandits

Wooseong Cho Min-hwan Oh

Seoul National University

Bilinear Matching Bandits



Bilinear Reward Model

$$r_{t,i} = \mathbf{x}_{t,i}^\top \mathbf{\Theta}^\star \mathbf{y}_{t,\pi_t(i)} + \eta_{t,i}, \qquad \mathbf{\Theta}^\star \in \mathbb{R}^{d_1 \times d_2}.$$

Objective: Minimize regret relative to the optimal matching $\pi_t^\star$.

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \sum_{i=1}^{M} \mathbf{x}_{t,i}^\top \mathbf{\Theta}^\star \left(\mathbf{y}_{t,\pi_t^\star(i)} - \mathbf{y}_{t,\pi_t(i)} \right).$$

Kronecker Baseline

- **Direct reduction**

$$\mathbf{z}_{t,i,j} = \mathbf{y}_{t,j} \otimes \mathbf{x}_{t,i} \in \mathbb{R}^p, \qquad p = d_1 d_2, \qquad \mathbf{x}_{t,i}^\top \mathbf{\Theta}^\star \mathbf{y}_{t,j} = \mathbf{z}_{t,i,j}^\top \operatorname{vec}(\mathbf{\Theta}^\star).$$

  The linear contextual combinatorial baseline yields

$$\operatorname{Reg}(T) = \widetilde{\mathcal{O}}\left(p\sqrt{MT}\right) = \widetilde{\mathcal{O}}\left(d_1 d_2 \sqrt{MT}\right).$$

(Takemura et al., 2021)

- **What remains unexploited**
  The baseline handles matching constraints and semi-bandit feedback, but treats each $\mathbf{y}_{t,j} \otimes \mathbf{x}_{t,i}$ as an unstructured vector in $\mathbb{R}^{d_1 d_2}$.

Can we exploit bilinear structure to improve the $d_1 d_2$ dependence?

Algorithm: HybridRRD

For each round $t = 1, \ldots, T$

1. Form the ridge estimate and predict edge rewards.

$$\boldsymbol{\theta}_t = \mathbf{V}_t^{-1}\mathbf{b}_t, \qquad \widehat{r}_{t,i,j} := \mathbf{z}_{t,i,j}^{\top}\boldsymbol{\theta}_t.$$

2. Compute the maximum uncertainty over feasible matchings.

$$\omega_t := \max_{\pi} \sum_i \left\|\mathbf{z}_{t,i,\pi(i)}\right\|^2_{\mathbf{V}_t^{-1}}.$$

3. Choose the exploration rule according to ω_t.

High width $\omega_t > \rho$	Play the maximum-width matching to reduce uncertainty in less explored directions.
Low width $\omega_t \leq \rho$	Use RRD to balance item-span coverage and predicted reward over feasible matchings.

4. Observe the M edge rewards and update the ridge statistics.

Row-Wise Regularized Design

- Build user-wise item covariances using item-span features $\widetilde{\mathbf{y}}_{t,j}$

$$\mathbf{V}_{t,i}(\mathbf{P}) := \sum_{j=1}^{N} (\mathbf{P})_{i,j} \widetilde{\mathbf{y}}_{t,j} \widetilde{\mathbf{y}}_{t,j}^{\top}.$$

- **Balance item coverage and predicted reward**

$$\mathbf{P}_t \in \operatorname*{argmax}_{\mathbf{P} \in \mathcal{M}} \left[\underbrace{\sum_{i=1}^{M} \log\det \mathbf{V}_{t,i}(\mathbf{P})}_{\text{row-wise item coverage}} + \xi \underbrace{\langle \mathbf{P}, \widehat{\mathbf{R}}_t \rangle}_{\text{predicted reward}} \right].$$

- Sample a feasible matching with edge marginals $\mathbf{P}_t$

$$\pi_t \sim \zeta_{\mathbf{P}_t}, \qquad \Pr\{\pi_t(i) = j\} = (\mathbf{P}_t)_{i,j}.$$

Regret Analysis

- **Matching-level uncertainty control.**
  RRD uses row-wise item coverage to control uncertainty in the total rewards of competing matchings.

- **Improved dimension dependence.**
  The ridge estimator operates in $d_1 d_2$ dimensions, while the RRD certificate contributes only $M d_2$.

$$\widetilde{\mathcal{O}}\left(\sqrt{\underbrace{d_1 d_2}_{\text{estimation}} \cdot \underbrace{M d_2}_{\text{RRD coverage}} \cdot T}\right) = \widetilde{\mathcal{O}}\left(d_2\sqrt{d_1 M T}\right)$$

Regret Guarantees

Theorem (High-probability upper bound)

Under boundedness and within-round sub-Gaussian martingale noise,

$$\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(\sigma_\eta d_2 \sqrt{d_1 M T} + B M d_1 d_2\right), \qquad B = S L_x L_y.$$

Theorem (Minimax lower bound)

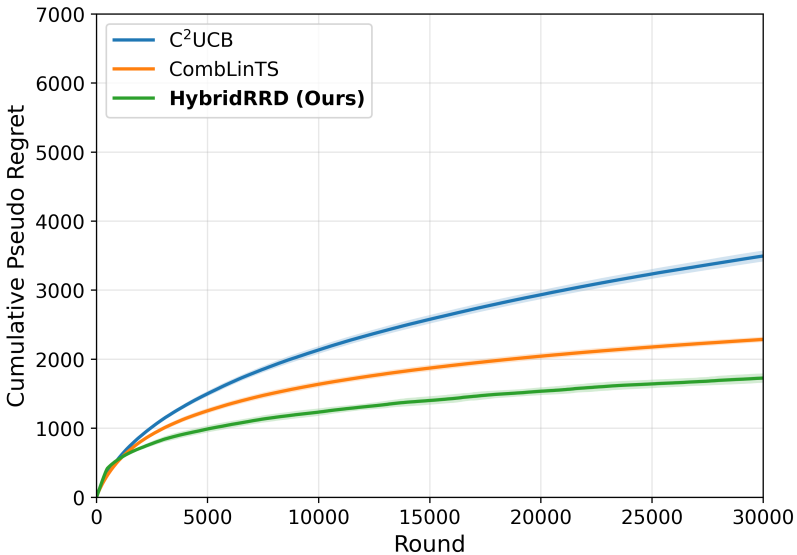
For $T \geq 2d_1$, $N \geq M 2^{d_2}$, and sufficiently large S,

$$\inf_{\mathrm{Alg}} \sup_{I} \mathbb{E}_{I,\mathrm{Alg}}\left[\mathrm{Reg}(T)\right] = \Omega\left(\sigma_\eta d_2 \sqrt{d_1 M T}\right).$$

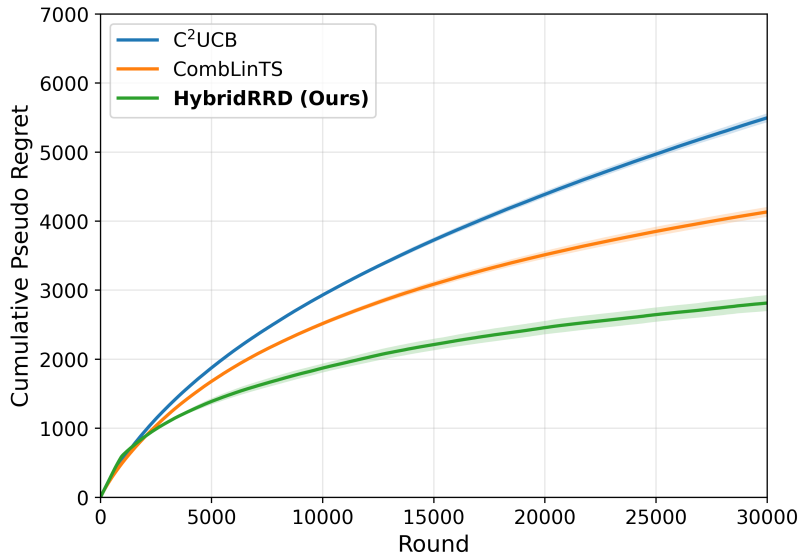
The stochastic term is minimax sharp up to logarithmic factors.

$$\underbrace{d_1 d_2}_{\text{Kronecker baseline}} \longrightarrow \underbrace{d_2\sqrt{d_1}}_{\text{HybridRRD}}.$$

Numerical Experiments



$(d_1, d_2) = (20, 20)$



$(d_1, d_2) = (40, 20)$

- Synthetic experiments: $M = 10$, $N = 30$, $T = 30{,}000$, and $\sigma_\eta = 0.2$; mean $\pm$ one standard deviation over five seeds.
- C^{2}UCB (Qin et al., 2014) and CombLinTS (Wen et al., 2015) use the same Kronecker features, semi-bandit feedback, and matching oracle.
- HybridRRD achieves the lowest cumulative pseudo-regret in both settings, including when d_1 doubles.

References

Qin, L., Chen, S., and Zhu, X. (2014). Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469. SIAM.

Takemura, K., Ito, S., Hatano, D., Sumita, H., Fukunaga, T., Kakimura, N., and Kawarabayashi, K.-i. (2021). Near-optimal regret bounds for contextual combinatorial semi-bandits with linear payoff functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9791–9798.

Wen, Z., Kveton, B., and Ashkan, A. (2015). Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 1113–1122. PMLR.