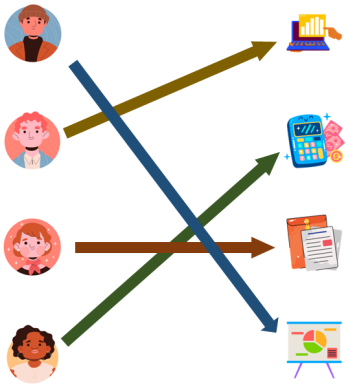
Adaptive Bandit Algorithms for Contextual Matching Markets

Shiyun Lin[1], Simon Mauras[2], Vianney Perchet[3], Nadav Merlis[4]

1. School of Mathematical Sciences, Peking University
2. INRIA, FairPlay Joint Team
3. CREST, ENSAE, Criteo AI Lab, FairPlay Joint Team
4. Technion - Israel Institute of Technology

June 1, 2026

Stable Matching (Matching Markets)



- Two sides of the market: Players & Arms
- Each participant has a preference ordering over the other side.
- Arms over players: $p \succ_a p' \Rightarrow$ arm a strictly prefers player p over p'.
- Players over arms: utility matrix $\boldsymbol{U}$ where $\boldsymbol{U}(p, a) > \boldsymbol{U}(p, a')$ indicates player p prefers arm a over a'.
- **Stable Matching:** No player and arm form a blocking pair, i.e., there is no player-arm pair who would both be better off by matching with each other instead of their current partners.
- **Player-optimal stable matching**: a single stable matching that is utility-maximizing for all players.

Contextual Bandits in Matching Markets

- The preferences of arms over players are fixed, known and strict.
- The preferences of players over arms are dynamic, unknown, and may admit ties, encoded with utility matrix $U(t)$, where $U_{i,j}(t) = \theta_i^\top x_j(t)$.
  - $\theta_i \in \mathbb{R}^d$ is the preference parameter for player p_i and should be learned from interaction.
  - $x_j(t)$ is the contextual information for arm a_j in time t, which is revealed before the matching is implemented each time.
- $y_{i,j}(t) = U_{i,j}(t) + \epsilon_{i,j}(t)$ is the observed reward if player p_i and arm a_j are matched at time t, where $\epsilon_{i,j}(t)$ is a subgaussian noise.
- Define the OSS for every player p_i at time t as $\boldsymbol{U}_i^*(t) = \max_{\mu \in \mathcal{S}_t} \boldsymbol{U}_{i,\mu_i(t)}(t)$.
- Player-optimal stable regret:

$$R_i(T) = \sum_{t=1}^{T} \boldsymbol{U}_i^*(t) - \mathbb{E}\left[\sum_{t=1}^{T} y_i(t)\right], \forall i \in [N].$$

Stochastic Contexts

- $(x_j(t))_{j\in[K]} \underset{\text{i.i.d.}}{\sim} \mathcal{D}_x := (\mathcal{D}_{x_j})_{j\in[K]}$.
- Given the realized contexts, define the minimum difference between any two utilities for player p_i as

$$\delta_{\min}^{(i)}(t) = \min_{j,j'\in[K]} |(\theta_i^*)^\top (x_j(t) - x_{j'}(t))|, \quad \delta_{\min}(t) = \min_{i\in[N]} \delta_{\min}^{(i)}(t).$$

- Previous work assumes that there is a hard threshold for $\delta_{\min}$, i.e., $\delta_{\min}(t) \geq \tilde{\Delta}$ for any $t \in [T]$. And they need to know $\tilde{\Delta}$.
- Relaxed gap definition: Replacing the hard threshold with a soft threshold to quantify the difficulty of the problem:

$$\Delta_{\min} := \max\left\{\Delta : \mathbb{P}(\delta_{\min}(t) \geq \Delta) \geq 1 - \frac{\log T}{T\Delta^2}\right\}.$$

- The gap in previous work could be viewed as $\mathbb{P}\left(\delta_{\min}(t) \geq \tilde{\Delta}\right) = 1 \Longrightarrow \Delta_{\min} \geq \tilde{\Delta}$.

Any regret upper bound established using $\Delta_{\min}$ is valid for $\tilde{\Delta}$.

Batched Adaptive Regret-Balancing (BARB) Algorithm

- Adaptive algorithm: BARB.
  - **Input**: Candidate gap Δ_1 (error tolerance) and η.
  - In each round t_k of batch k:
    1. Observe context $x_j(t_k)$ for arm $a_j, \forall j \in [K]$.
    2. If $\exists i, j$, s.t. $\|x_j(t_k)\|_{\left(\mathbf{V}_i^{(t_k)}\right)^{-1}} > \xi_k := \frac{\Delta_k}{\eta}$ (the current direction induced by the observed contexts is not well estimated), explore by performing a Maximum Cardinality Matching.
    3. Otherwise, exploit by performing a Gale-Shapley matching with the estimated utility matrix $\hat{\boldsymbol{U}}(t_k)$. If there exists overlapping utility CIs, $N_k \leftarrow N_k + 1$.
    4. If $N_k > \frac{3 \log T}{16 \Delta_k^2}$, $\Delta_{k+1} \leftarrow \frac{\Delta_k}{\sqrt{2}}$, enter the next batch.
- Regret upper bound: $O(\frac{\log^2 T}{\Delta_{\min}^2})$.

Matching Asymptotic Regret Bounds

Assumption (Regularity Condition)

Denote the CDF of $\delta_{\min}$ as $F(\Delta)$, for some constant Δ_0 sufficiently close to 0, and for some constant $c > 0$, we have, $\forall \Delta \leq \Delta_0$, $F(\Delta) \leq c\Delta$.

The above assumption ensures that we do not have excessively high probability for small gaps.

Theorem (Asymptotic Regret Upper Bound)

Under the above regularity condition, when T is sufficiently large, the regret of the BARB algorithm satisfies $\mathrm{Reg}_i(T) = \tilde{O}\left(T^{2/3}\right), \forall i \in [N]$.

Theorem (Asymptotic Regret Lower Bound)

For any policy π, there exists an instance satisfying the above regularity condition and at least one player p_i suffers regret $\Omega\left(T^{2/3}\right)$.

- Adversarial contexts: The contexts are chosen by an adversary either independent or with knowledge of the algorithm's previous outcomes.
- We define the α-**approximate** Δ-**optimal stable regret** as follows:

$$R_i^{\alpha,\Delta}(T) = \sum_{t=1}^{T} [\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t) > \Delta\} + \alpha \boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t) \leq \Delta\}] - \mathbb{E}\left[\sum_{t=1}^{T} y_i(t)\right],$$

where $\boldsymbol{U}_i^*(t) = \max_{\mu \in \mathcal{S}_t} \boldsymbol{U}_{i,\mu_t(i)}(t)$ is the OSS for every player p_i at time t, and $\boldsymbol{U}_i^{\varepsilon}(t) = \max_{\mu \in \mathcal{S}_t^{\varepsilon}} \boldsymbol{U}_{i,\mu(i)}(t)$ is the ε-optimal stable share for player p_i at time t.

- We propose Adaptive Explore-Choose Oracle (AdECO) algorithm.
- With suitable parameter setting, the regret upper bound is $Reg_i^{\alpha,\Delta}(T) = \mathcal{O}\left(T^{2/3}\right)$.

Thanks for listening!