Stochastic Matching Bandits with Rare Optimization Updates

Jung-hun Kim
FairPlay joint team, CREST, ENSAE
Institut Polytechnique de Paris
junghun.kim@ensae.fr

Min-hwan Oh
Seoul National University
minoh@snu.ac.kr

Abstract

We introduce a bandit framework for stochastic matching under the multinomial logit (MNL) choice model. In this setting, N agents on one side are assigned to K arms on the other side, where each arm stochastically selects an agent from its assigned pool according to unknown preferences and yields a corresponding reward over a horizon T. The objective is to minimize regret by maximizing the cumulative revenue from successful matches. A naive approach requires solving an NP-hard combinatorial optimization problem at every round, resulting in a prohibitive computational cost. To address this challenge, we propose batched algorithms that strategically limit the number of times matching assignments are updated to $\Theta(\log \log T)$ over the entire horizon. By invoking expensive combinatorial optimization only on a vanishing fraction of rounds, our algorithms substantially reduce overall computational overhead while still achieving a regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$.

1 Introduction

Online matching platforms, including business-to-business procurement and service marketplaces, seek to maximize revenue from repeated transactions. For example, a procurement platform can assign suppliers (agents) to buyers (arms), each of whom chooses at most one supplier from its assigned pool. Suppliers can remain registered across rounds: being unmatched means receiving no transaction in that round, while a successful match represents a repeatable transaction rather than a permanent pairing.

This demand has led to extensive research on online bipartite matching problems [Karp et al., 1990, Mehta et al., 2007, 2013, Gamlath et al., 2019, Fuchs et al., 2005, Kesselheim et al., 2013], where two sets of vertices are considered and one side is revealed sequentially. These studies primarily focus on maximizing the number of matches. However, a significant gap remains between these theoretical models and practical scenarios for maximizing revenue under latent preference functions. Specifically, these models generally assume one-to-one assignments under deterministic matching and focus solely on match count, without incorporating *learning mechanisms* that adapt to observed reward feedback or aim to maximize cumulative revenue.

More recently, stable matching bandits have been proposed to capture online learning dynamics in matching markets [Liu et al., 2020, 2021, Sankararaman et al., 2020, Basu et al., 2021, Zhang et al., 2022, Kong and Li, 2023]. In this line of work, agents are assigned to arms in each round, and the learning objective is to infer the underlying preference/reward structure so as to identify a stable matching [McVitie and Wilson, 1971].

A common modeling abstraction in this literature is that, when multiple agents are tentatively assigned to an arm, the arm resolves the contention according to a fixed known preference order and

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

therefore selects a match in a deterministic manner. This abstraction is well-suited for stability-driven formulations. In many platform settings, however, acceptance decisions can be inherently stochastic and reflect latent preferences. For instance, when a procurement platform offers a buyer several suppliers, the buyer may select one or decline all offers according to preferences that are initially unknown to the platform.

Motivated by such applications, we take a complementary perspective and introduce *stochastic matching bandits* (SMB), which explicitly model stochastic arm-side choice under unknown preferences and target revenue maximization. In SMB, multiple agents may be assigned to the same arm in a round, and the arm then stochastically selects (at most) one agent from its assigned pool. This shift in modeling leads to a different and practically motivated objective: maximizing the platform's cumulative system-wide revenue.

However, realizing this goal comes with substantial computational challenges. In particular, determining the optimal assignment at each round requires solving an NP-hard combinatorial optimization problem. As a result, naive approaches must repeatedly re-solve this expensive optimization to recompute the set of assortments or matchings used in subsequent rounds, leading to a substantial overall computational burden. These considerations naturally raise the following fundamental question:

Can we achieve no-regret learning to maximize revenue in stochastic matching bandits while avoiding combinatorial optimization at every round?

To address this challenge, we propose *batched* algorithms for the SMB framework that strategically limit the number of times matching assignments are updated. By invoking expensive combinatorial optimization only a small number of times over the entire horizon, our algorithms achieve no-regret performance while substantially reducing the overall computational burden. Below, we summarize our main contributions.

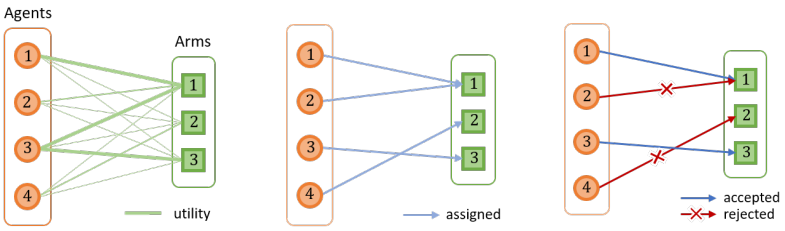
Summary of Our Contributions.

- We introduce a practical framework of *stochastic matching bandits* (SMB), which models stochastic arm-side choices under latent preferences and targets revenue maximization. While this framework captures important features of real-world matching systems, a naive implementation requires solving an NP-hard combinatorial optimization problem at every round, resulting in a substantial overall computational burden.
- Under the SMB framework, we develop a batched algorithm that balances exploration and exploitation with limited matching updates. Assuming knowledge of a non-linearity parameter κ, the algorithm achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret while invoking expensive combinatorial optimization only $\Theta(\log\log T)$ times over the entire horizon, substantially reducing the overall computational burden compared to naive per-round optimization.
- We further propose a second algorithm that removes the requirement of knowing κ in advance. This algorithm retains the same $\widetilde{\mathcal{O}}(\sqrt{T})$ regret guarantee and still requires only $\Theta(\log\log T)$ batch updates, preserving the key property that combinatorial optimization is performed only a small number of times over the horizon.

2 Related Work

Stable Matching Bandits. We review the literature on matching bandits, which studies regret minimization in matching markets. This line of work was initiated by Liu et al. [2020] and extended by Sankararaman et al. [2020], Liu et al. [2021], Basu et al. [2021], Zhang et al. [2022], Kong and Li [2023], focusing on finding optimal stable matchings through stochastic reward feedback, under assumption that the number of agents does not exceed the number of arms ($N \leq K$).

Our proposed *Stochastic Matching Bandits (SMB)* framework departs from this literature in several key ways. First, whereas prior work typically assumes that arms select agents *deterministically* according to known preferences, SMB models arms as making *stochastic* choices driven by unknown latent preferences that must be learned over time. This shift moves the objective from identifying a static stable matching to maximizing cumulative reward through adaptive learning. Second, SMB captures richer preference structures by allowing arm-side choices to depend on the *composition of the assigned set*, so that selection probabilities and expected rewards are determined by the *relative*



(a) Each agent–arm pair is associated with an unknown utility.

(b) Each agent is assigned to an arm by a policy.

(c) Each arm *stochastically* accepts at most one assigned agent.

Figure 1: Illustration of the stochastic matching process with 4 agents ($N = 4$) and 3 arms ($K = 3$).

utilities of agents competing for the same arm. Third, SMB imposes no structural restrictions on market size, naturally accommodating both $N \leq K$ and $N \geq K$ regimes.

Taken together, these modeling choices define a distinct perspective on matching under uncertainty and broaden the applicability of SMB to real-world platforms—such as B2B procurement and service marketplaces—where preferences are stochastic, assignment-dependent, and inherently relative.

MNL-Bandits. In our study, we adopt the Multinomial Logit (MNL) model for arms' choice preferences in matching bandits. As the first MNL bandit method, Agrawal et al. [2017a] proposed an epoch-based algorithm, followed by subsequent contributions from Agrawal et al. [2017b], Chen et al. [2023], Oh and Iyengar [2019, 2021], Lee and Oh [2024]. However, unlike selecting an assortment at each time step, our framework for the stochastic matching market mandates choosing at most K distinct assortments to assign agents to each arm. Consequently, handling K-multiple MNLs simultaneously results in exponential computational complexity. More recently, Kim and Oh [2024] studied MNL-based preferences in matching bandits; however, their focus was on system stability under binary $(0/1)$ rewards, rather than revenue maximization. Additionally, their work did not address the computational intractability arising from per-round combinatorial optimization in this context.

Batch Learning in Bandits. Batch learning in bandit problems has been extensively studied in the context of multi-armed bandits (MAB) [Perchet et al., 2015, Gao et al., 2019], and later extended to (generalized) linear bandit models [Ruan et al., 2021, Hanna et al., 2023, Han et al., 2020, Ren and Zhou, 2024, Sawarni et al., 2024, Ren et al., 2024, Chen et al., 2023]. More recently, Midigeshi et al. [2025] study batched updates for multinomial logistic bandits. However, their setting considers a single-item selection problem, where the learner chooses one arm at each round, and thus does not involve combinatorial action spaces.

Lastly, Kim et al. [2025] study combinatorial semi-bandits with rare update policies. Their analysis assumes that each action is associated with a fixed (unknown) mean reward, independent of other simultaneously selected actions. In contrast, in the MNL bandit setting considered here, the expected reward of an agent–arm match is a nonlinear and set-dependent function of the entire proposed assignment, due to stochastic arm-side choice probabilities. As a result, their framework does not capture the reward coupling induced by MNL-based matching markets.

3 Problem Statement

We study stochastic matching bandits (SMB) with N agents and K arms. The stochastic matching process is illustrated in Figure 1. The detailed formulation is as follows: For each agent $n \in [N]$, feature information is known as $x_n \in \mathbb{R}^d$, and each arm $k \in [K]$ is characterized by latent vector $\theta_k \in \mathbb{R}^d$. We define the set of features as $X = [x_1, \ldots, x_N] \in \mathbb{R}^{d \times N}$ and the rank of X as $rank(X) = r(\leq d)$. At each time $t \in [T]$, each agent n may be assigned to an arm $k_{n,t} \in [K]$. Let assortment $S_{k,t} = \{n \in [N] : k_{n,t} = k\}$, which is the set of agents that are assigned to an arm k at time t. Then given an assortment to each arm k at time t, $S_{k,t}$, each arm k randomly accepts an agent $n \in S_{k,t}$ and receives reward $w_{n,k} \in [0, 1]$ according to the arm's preference specified as

follows. The probability for arm k to accept agent $n \in S_{k,t}$ follows Multinomial Logit (MNL) model [Agrawal et al., 2017a,b, Oh and Iyengar, 2019, 2021, Chen et al., 2023] given by

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

We denote $x_n^\top \theta_k$ as the latent preference utility of arm k for agent n. Following prior work on MNL bandits [Oh and Iyengar, 2019, 2021, Agrawal et al., 2019], we consider that the candidate set size is bounded by $|S_{k,t}| \leq L$ for all arms k and rounds t, and that the reward $w_{n,k}$ is known to the learner.

At each time step t, the agents receive stochastic feedback based on the assortments $\{S_{k,t}\}_{k\in[K]}$. Specifically, for each agent $n \in S_{k,t}$ and arm $k \in [K]$, the feedback is denoted by $y_{n,k,t} \in \{0, 1\}$, where $y_{n,k,t} = 1$ if arm k accepts agent n (i.e., a successful match occurs), and $y_{n,k,t} = 0$ otherwise. Following the standard MNL model, each arm k may also choose an outside option n_0 (i.e., reject all assigned agents) with probability $p(n_0|S_{k,t}, \theta_k) = 1/(1 + \sum_{m \in S_{k,t}} \exp(x_m^\top \theta_k))$, which yields zero reward. Then, given assortments to every arm k, $\{S_k\}_{k\in[K]}$, the expected reward (revenue) for the assortments at time t is defined as

$$\sum_{k\in[K]} R_k(S_k) := \sum_{k\in[K]} \sum_{n\in S_k} w_{n,k} p(n|S_k, \theta_k).$$

The goal of the problem is to maximize the cumulative expected reward over a time horizon T by learning the unknown parameters $\{\theta_k\}_{k\in[K]}$. We define the oracle strategy as the optimal assortment selection when the preference parameters θ_k are known. Let the set of all feasible assignments be: $\mathcal{M} = \{\{S_k\}_{k\in[K]} : S_k \subset [N], |S_k| \leq L \ \ \forall k \in [K], S_k \cap S_l = \emptyset \ \ \forall k \neq l\}$. Then the oracle assortment is given by:

$$\{S_k^*\}_{k\in[K]} = \operatorname*{argmax}_{\{S_k\}_{k\in[K]}\in\mathcal{M}} \sum_{k\in[K]} R_k(S_k).$$

Given $\{S_{k,t}\}_{k\in[K]} \in \mathcal{M}$ for all $t \in [T]$, the expected cumulative regret is defined as

$$\mathcal{R}(T) = \mathbb{E}\Big[\sum_{t\in[T]} \sum_{k\in[K]} R_k(S_k^*) - R_k(S_{k,t})\Big].$$

The objective is to design a policy that minimizes this regret over the time horizon T.

Similar to previous work for logistic and MNL bandit [Oh and Iyengar, 2019, 2021, Lee and Oh, 2024, Goyal and Perivier, 2021, Faury et al., 2020, Abeille et al., 2021], we consider the following regularity condition and non-linearity quantity.

Assumption 3.1. $\|x_n\|_2 \leq 1$ for all $n \in [N]$ and $\|\theta_k\|_2 \leq 1$ for all $k \in [K]$.

Then we define a problem-dependent quantity regarding non-linearity of the MNL structure as follows:

$$\kappa := \inf_{\substack{\theta\in\mathbb{R}^d:\|\theta\|_2\leq 1 \\ n\in S\subseteq[N]:|S|\leq L}} p(n|S, \theta) p(n_0|S, \theta).$$

4 Optimization in Stochastic Matching Bandits: The Curse of Complexity

In this work, we develop algorithms for the Stochastic Matching Bandit (SMB) problem with preference feedback. SMB can be viewed as a generalization of the standard Multinomial Logit (MNL) bandit model with a single assortment [Oh and Iyengar, 2021, Lee and Oh, 2024] to a setting with K simultaneous assortments—one for each arm. Applying existing MNL-based methods to this setting requires dynamically selecting K assortments at each round while simultaneously learning arm preferences in an online fashion. This extension introduces significant computational challenges: the resulting combinatorial optimization problem is NP-hard. In contrast, the standard MNL bandit problem with a single assortment is known to be solvable in polynomial time [Oh and Iyengar, 2021]. Thus, the SMB framework poses a substantially more complex optimization problem, highlighting the need for efficient algorithmic solutions.

Naively extending MNL bandits (e.g. Oh and Iyengar [2021], Lee and Oh [2024]) to SMB requires defining the UCB index for the expected reward of an assortment S_k for all $k \in [K]$ as $R_{k,t}^{UCB}(S_k) =$

$\sum_{n\in S_k} \frac{w_{n,k}\exp(h_{n,k,t})}{1+\sum_{m\in S_k}\exp(h_{m,k,t})}$, where $h_{n,k,t}$ is an UCB index for the utility value between n and k at each time t. Then at each time, the algorithm determines assortments by following the UCB strategy:

$$\{S_{k,t}\}_{k\in[K]} = \underset{\{S_k\}_{k\in[K]}\in\mathcal{M}}{\operatorname{argmax}} \sum_{k\in[K]} R_{k,t}^{UCB}(S_k). \tag{1}$$

While this method can achieve a regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$, it suffers from severe computational limitations. Let C_{opt} denote the worst-case cost of solving (1) once. Since (1) is NP-hard in general, exact maximization can be computationally prohibitive. A trivial worst-case upper bound is obtained by exhaustive search over feasible assignments: when $L \geq N$, the action space has size $\mathcal{O}(K^N)$. Consequently, recomputing the maximizer at every round leads to total optimization cost on the order of $T \cdot C_{\mathrm{opt}} = \mathcal{O}(TK^N)$. Further details of the algorithm and regret analysis are provided in Appendix A.2.

To overcome the overall computational burden, we propose a *batched learning* approach that substantially reduces the total number of combinatorial optimization updates, thereby reducing the overall computational cost.

Remark 4.1 (Approximation oracles and rare updates). Approximation oracles [Kakade et al., 2007, Chen et al., 2013] are a standard way to address the computational difficulty of combinatorial optimization. However, replacing the exact oracle with an approximation oracle changes the benchmark from exact regret to approximation regret. In this work, we instead target exact regret by using exact optimization only at rare batch boundaries. These two ideas are complementary: approximation reduces the cost of each optimization call, whereas batching reduces the number of such calls. Therefore, batch updates can also be combined with approximation oracles to further reduce computation, at the price of targeting the corresponding approximation-regret benchmark. Additional details are provided in Appendix A.8.

5 Batch Learning for Stochastic Matching Bandits

For batch learning to reduce the computational cost, we adopt the elimination-based bandit algorithm [Lattimore and Szepesvári, 2020]. This approach presents several key challenges in the framework of SMB, including efficiently handling the large number of possible matchings between agents and arms for elimination, designing an appropriate estimator for the elimination process, and minimizing the total number of updates to reduce computational overhead. The details of our algorithm (Algorithm 1) are described as follows.

Preprocessing. Before the online rounds, the algorithm computes the SVD $X = U\Sigma V^\top$ and lets $r = \mathrm{rank}(X)$. It then projects each feature vector onto the rank-r subspace, defining $z_n = U_r^\top x_n \in \mathbb{R}^r$, where U_r contains the left singular vectors associated with nonzero singular values. With $\theta_k^* = U_r^\top \theta_k$, the MNL model is equivalently written in terms of (z_n, θ_k^*). This projection does not require prior knowledge of r and also ensures proper regularization in the warm-up phase (Line 9 of Algorithm 1); see Appendix A.3 for details.

Estimation. In what follows, we describe the process for constructing assortments at each time step. The algorithm consists of several epochs. For each $k \in [K]$, from observed feedback $y_{n,k,t} \in \{0,1\}$ for $n \in S_{k,t}$, $t \in \mathcal{T}_{k,\tau-1}$, where $\mathcal{T}_{k,\tau-1}$ is a set of the exploration time steps regarding arm k in the $\tau-1$-th epoch, we first define the negative log-likelihood loss as

$$l_{k,\tau}(\theta) = -\sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}\cup\{n_0\}} y_{n,k,t}\log p(n|S_{k,t},\theta) + \tfrac{1}{2}\|\theta\|_2^2, \tag{2}$$

where, with a slight abuse of notation, $p(n|S_{k,t},\theta) := \exp(z_n^\top\theta)/(1+\sum_{m\in S_{k,t}}\exp(z_m^\top\theta))$. Then at the beginning of each epoch τ, the algorithm estimates $\widehat{\theta}_{k,\tau}$ from the method of Maximum Likelihood Estimation (MLE).

Algorithm 1: Batched Stochastic Matching Bandit (`B-SMB`)

1 **Input:** κ, $M \geq 1$; **Initialize:** $t \leftarrow 1, T_1 \leftarrow \eta_T, \mathcal{M}_0 \leftarrow \mathcal{M}, \mathcal{N}_{k,0} \leftarrow [N], \mathcal{T}_{k,0} \leftarrow \emptyset$

2 Compute the SVD $X = U\Sigma V^\top$, form $U_r = [u_1, \ldots, u_r]$, and set $z_n \leftarrow U_r^\top x_n$ for all $n \in [N]$.

3 **for** $\tau = 1, 2...$ **do**

4 **for** $k \in [K]$ **do**

5 $\widehat{\theta}_{k,\tau} \leftarrow \text{argmin}_{\theta \in \mathbb{R}^r} l_{k,\tau}(\theta)$ with (2) where $\mathcal{T}_{k,\tau-1} = \mathcal{T}^{(1)}_{k,\tau-1} \cup \mathcal{T}^{(2)}_{k,\tau-1}$ and $\mathcal{T}^{(2)}_{k,\tau-1} = \bigcup_{n \in \mathcal{N}_{k,\tau-1}} \mathcal{T}^{(2)}_{n,k,\tau-1}$

`// Assortments Construction & Elimination with` (3)

6 $\{S^{(n,k)}_{l,\tau}\}_{l \in [K]} \leftarrow \underset{\{S_l\}_{l \in [K]} \in \mathcal{M}_{\tau-1} : n \in S_k}{\text{argmax}} \sum_{l \in [K]} R^{UCB}_{l,\tau}(S_l)$ for all $n \in \mathcal{N}_{k,\tau-1}$

7 $\mathcal{N}_{k,\tau} \leftarrow \{n \in \mathcal{N}_{k,\tau-1} : \max_{\{S_l\}_{l \in [K]} \in \mathcal{M}_{\tau-1}} \sum_{l \in [K]} R^{LCB}_{l,\tau}(S_l) \leq \sum_{l \in [K]} R^{UCB}_{l,\tau}(S^{(n,k)}_{l,\tau})\}$

`// G-Optimal Design`

8 $\pi_{k,\tau} \leftarrow \text{argmin}_{\pi \in \mathcal{P}(\mathcal{N}_{k,\tau})} \max_{n \in \mathcal{N}_{k,\tau}} \|z_n\|^2_{(\sum_{n \in \mathcal{N}_{k,\tau}} \pi_{k,\tau}(n) z_n z_n^\top + (1/rT_\tau) I_r)^{-1}}$

`// Exploration`

9 Run `Warm-up` (Algorithm 4) over time steps in $\mathcal{T}^{(1)}_{k,\tau}$ (defined in Algorithm 4)

10 **for** $n \in \mathcal{N}_{k,\tau}$ **do**

11 $t_{n,k} \leftarrow t$, $\mathcal{T}^{(2)}_{n,k,\tau} \leftarrow [t_{n,k}, t_{n,k} + \lceil r \pi_{k,\tau}(n) T_\tau \rceil - 1]$

12 **while** $t \in \mathcal{T}^{(2)}_{n,k,\tau}$ **do**

13 Offer $\{S_{l,t}\}_{l \in [K]} = \{S^{(n,k)}_{l,\tau}\}_{l \in [K]}$ and observe feedback $y_{m,k,t} \in \{0, 1\}$ for all $m \in S_{l,t}$ and $l \in [K]$

14 $t \leftarrow t + 1$

15 $\mathcal{M}_\tau \leftarrow \{\{S_k\}_{k \in [K]} : S_k \subset \mathcal{N}_{k,\tau}, |S_k| \leq L \; \forall k \in [K], S_k \cap S_l = \emptyset \; \forall k \neq l\}$

16 $T_{\tau+1} \leftarrow \eta_T \sqrt{T_\tau}$

Confidence Bound Construction. From the estimator, we define upper and lower confidence bounds for expected reward of assortment S_k as

$$R^{UCB}_{k,\tau}(S_k) := \sum_{n \in S_k} [w_{n,k} p(n|S_k, \widehat{\theta}_{k,\tau})] + 2\beta_T \max_{n \in S_k} \|z_n\|_{V^{-1}_{k,\tau}},$$
$$R^{LCB}_{k,\tau}(S_k) := \sum_{n \in S_k} [w_{n,k} p(n|S_k, \widehat{\theta}_{k,\tau})] - 2\beta_T \max_{n \in S_k} \|z_n\|_{V^{-1}_{k,\tau}}, \tag{3}$$

where confidence term $\beta_T = \frac{C_1}{\kappa}\sqrt{\log(TNK)}$ for some constant $C_1 > 0$ and $V_{k,\tau} = \sum_{t \in \mathcal{T}_{k,\tau-1}} \sum_{n \in S_{k,t}} z_n z_n^\top + I_r$. Unlike prior MNL bandit methods [Oh and Iyengar, 2021, Lee and Oh, 2024], which construct confidence intervals for latent utilities inside the MNL function, we place the confidence term directly at the reward level. This is needed to obtain valid UCB and LCB indices for the expected reward, since lower-bounding latent utilities does not necessarily lower-bound the nonlinear MNL reward.

Efficient Elimination of Suboptimal Matches. For batch updates, we utilize elimination for suboptimal matches. However, exploring all possible matchings naïvely for the elimination is statistically expensive. Therefore, we utilize a statistically efficient exploration strategy by assessing the eligibility of each assignment (n, k) for $n \in \mathcal{N}_{k,\tau-1}$ and $k \in [K]$ as a potential optimal assortment, where $\mathcal{N}_{k,\tau-1}$ is the active set of agents regarding arm k at epoch τ. To evaluate the assignment (n, k), it constructs a representative assortment of $\{S^{(n,k)}_{l,\tau}\}_{l \in [K]}$ from an optimistic view (Line 6 in Algorithm 1). Then based on the representative assortments, it obtains $\mathcal{N}_{k,\tau}$ by eliminating $n \in \mathcal{N}_{k,\tau-1}$ which satisfies an elimination condition (Line 7). From the obtained $\mathcal{N}_{k,\tau}$ for all $k \in [K]$, it constructs an active set of assortments $\mathcal{M}_\tau$ (Line 15), which is likely to contain the optimal assortments as $\{S^*_k\}_{k \in [K]} \in \mathcal{M}_\tau$.

Assortment Assignment. Following the elimination process outlined above, here we describe the policy of assigning assortment $\{S_{k,t}\}_{k\in[K]}$ at each time t corresponding to Lines 8-14 in Algorithm 1. The algorithm initiates the warm-up stage (Line 9 in Algorithm 1) to apply regularization to the estimators, by uniform exploration across all agents $n \in [N]$ for each arm $k \in [K]$. Then for each $k \in [K]$, the algorithm utilizes the G-optimal design problem [Lattimore and Szepesvári, 2020] to obtain proportion $\pi_{k,\tau} \in \mathcal{P}(\mathcal{N}_{k,\tau})$ for learning θ_k^* efficiently by exploring agents in $\mathcal{N}_{k,\tau}$, where $\mathcal{P}(\mathcal{N}_{k,\tau})$ is the probability simplex with vertex set $\mathcal{N}_{k,\tau}$. Notably, the G-optimal design problem can be solved by the Frank-Wolfe algorithm [Damla Ahipasaoglu et al., 2008]. Then, for all $n \in \mathcal{N}_{k,\tau}$, it explores $\{S_{l,\tau}^{(n,k)}\}_{l\in[K]}$ several times using $\pi_{k,\tau}(n)$ which is the corresponding value of n in $\pi_{k,\tau}$.

Scheduling. The algorithm proceeds in epochs τ until the time horizon T is reached. The number of rounds in each epoch is scheduled recursively as $T_\tau = \eta_T \sqrt{T_{\tau-1}}$. This schedule limits the number of updates for assortment assignments, which is crucial for reducing the overall computational cost. We set $\eta_T = \left(\frac{T}{rK}\right)^{\frac{1}{2(1-2^{-M})}}$, where $M \geq 1$ specifies the batch update budget. Let τ_T denote the final epoch before time T, corresponding to the total number of batch updates. As formalized in the following proposition, the parameter M directly bounds τ_T and thus limits the frequency of optimization updates (see Appendix A.5 for the proof).

Proposition 5.1 (Number of Batch Updates)**.** $\tau_T \leq M$.

We establish the following regret bound for our algorithm, with the proof provided in Appendix A.6.

Theorem 5.2. *Algorithm 1 with $M = \mathcal{O}(\log T)$ achieves:*

$$\mathcal{R}(T) = \tilde{\mathcal{O}}\left(\tfrac{1}{\kappa} K^{\frac{3}{2}} \sqrt{rT} \left(\frac{T}{rK}\right)^{\frac{1}{2(2^M-1)}}\right).$$

Corollary 5.3. *For $M = \Theta\big(\log\log(\frac{T}{rK})\big)$, Algorithm 1 achieves:*

$$\mathcal{R}(T) = \tilde{\mathcal{O}}\left(\tfrac{1}{\kappa} K^{3/2} \sqrt{rT}\right).$$

Remark 5.4 (Efficiency via Rare Optimization Updates). As established in Corollary 5.3, Algorithm 1 invokes combinatorial optimization at most $M = \Theta\big(\log\log(\frac{T}{rK})\big)$ times over the entire horizon T, while achieving a regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$. Consequently, the total computational cost associated with combinatorial optimization is $\mathcal{O}(NKC_{\text{opt}} \log\log(\frac{T}{rK}))$. This stands in sharp contrast to the naive approach discussed in Section 4 (e.g., Algorithm 3 in Appendix A.2), which requires solving a combinatorial optimization problem of cost C_{opt} at every round, leading to a total cost of $\mathcal{O}(C_{\text{opt}}T)$. Thus, the dependence on the horizon is reduced from linear in T to doubly logarithmic in T, substantially reducing the overall computational burden.

Discussion on the Tightness of the Regret Bound. We begin by comparing our results to those from previous batch bandit studies under a (generalized) linear structure. Our regret bound, given as $\tilde{\mathcal{O}}(T^{\frac{1}{2}+\frac{1}{2(2^M-1)}}) = \tilde{\mathcal{O}}(T^{\frac{1}{2(1-2^{-M})}})$ for a general $M = \mathcal{O}(\log(T))$, matches the results from Han et al. [2020], Ren and Zhou [2024], Sawarni et al. [2024]. Notably, this bound also aligns with the lower bound for the linear structure, $\Omega(T^{\frac{1}{2(1-2^{-M})}})$ [Han et al., 2020]. For the case of $M = \Theta\big(\log\log(\frac{T}{rK})\big)$, our bound of $\tilde{\mathcal{O}}(\sqrt{T})$ corresponds to the findings for linear bandits in Ruan et al. [2021], Hanna et al. [2023] under $\log\log(T)$ batch updates.

Given that our problem generalizes the single-assortment MNL setting to K-multiple assortments, we can obtain the regret lower bound of $\Omega(K\sqrt{T})$ with respect to K and T for the contextual setting, by simply extending the result of Theorem 3 in Lee and Oh [2024] for single-assortment to K-multiple assortments. In comparison, our analysis indicates a regret dependence of $K^{3/2}$ when $M = \Theta\big(\log\log\big(\frac{T}{rK}\big)\big)$, which is worse by a factor of $\sqrt{K}$ relative to the lower bound. This gap arises from the need to explore all potential matches during the epoch-based elimination procedure in batch updates.

Although Algorithm 1 achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret with amortized computational efficiency due to reduced batch updates, its regret guarantee relies on problem-specific knowledge of κ and, crucially, requires this parameter to be known in advance for setting β_T. The resulting regret bound scales linearly with $1/\kappa$, which can be as large as $\mathcal{O}(L^2)$ in the worst case. In the following section, we propose an algorithm that removes the need for prior knowledge of κ.

6 Without Prior Knowledge κ

Here we provide details of our proposed algorithm (Algorithm 2 in Appendix A.1), focusing on the difference from the algorithm in the previous section.

Local curvature. While we follow the elimination framework of Algorithm 1, to improve the dependence on the unknown parameter κ, we exploit the local curvature of the log-likelihood around the current estimate $\widehat{\theta}_{k,\tau}$. Specifically, we define the regularized empirical Hessian

$$H_{k,\tau}(\widehat{\theta}_{k,\tau}) = \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \left[\sum_{n\in S_{k,t}} p(n|S_{k,t},\widehat{\theta}_{k,\tau}) z_n z_n^\top - \sum_{n,m\in S_{k,t}} p(n|S_{k,t},\widehat{\theta}_{k,\tau}) p(m|S_{k,t},\widehat{\theta}_{k,\tau}) z_n z_m^\top \right], \quad (4)$$

where $\lambda = C_2 r \log(K)$ for some constant $C_2 > 0$ and we denote $H_{k,\tau}(\widehat{\theta}_{k,\tau})$ by $H_{k,\tau}$ when there is no confusion. We define $\tilde{z}_{n,k,\tau}(S_{k,t}) = z_n - \sum_{m\in S_{k,t}} p(m|S_{k,t},\widehat{\theta}_{k,\tau}) z_m$ and we use $\tilde{z}_{n,k,\tau}$ for it, when there is no confusion.

Confidence Bound Construction. For the confidence bound, we define

$$B_\tau(S_{k,t}) := \tfrac{13}{2}\zeta_\tau^2 \max_{n\in S_{k,t}} \|z_n\|^2_{H_{k,\tau}^{-1}} + 2\zeta_\tau^2 \max_{n\in S_{k,t}} \|\tilde{z}_{n,k,\tau}\|^2_{H_{k,\tau}^{-1}} + \zeta_\tau \sum_{n\in S_{k,t}} p(n|S_{k,t},\widehat{\theta}_{k,\tau-1}) \|\tilde{z}_{n,k,\tau}\|_{H_{k,\tau}^{-1}},$$

where $\zeta_\tau = \frac{1}{2}\sqrt{\lambda} + \frac{2r}{\sqrt{\lambda}} \log(4KT(1 + \frac{2(t_\tau - 1)L}{r\lambda}))$ with the start time of τ-th episode t_τ. We note that the first term arises from the second-order term in the Taylor expansion for the error from estimator, while the second and last terms originate from the first-order term. Notably, our confidence bounds for τ-th episode utilize not only the current estimator $\widehat{\theta}_{k,\tau}$ but the previous one $\widehat{\theta}_{k,\tau-1}$ (in the last term) because the historical data in $H_{k,\tau}$ is obtained from the G/D-optimal policy which is optimized under $\widehat{\theta}_{k,\tau-1}$. Then we define upper and lower confidence bounds as

$$\begin{aligned} R_{k,\tau}^{UCB}(S_{k,t}) &:= \sum_{n\in S_{k,t}} w_{n,k} p(n|S_{k,t},\widehat{\theta}_{k,\tau}) + B_\tau(S_{k,t}), \\ R_{k,\tau}^{LCB}(S_{k,t}) &:= \sum_{n\in S_{k,t}} w_{n,k} p(n|S_{k,t},\widehat{\theta}_{k,\tau}) - B_\tau(S_{k,t}). \end{aligned} \quad (5)$$

Exploration Space. For the G/D-optimal design aimed at exploring the space of arms, the algorithm must account for both $V_{k,\tau}$ and $H_{k,\tau}(\widehat{\theta}_{k,\tau})$ to achieve a tight regret bound that avoids dependence on $1/\kappa$. This marks a key distinction from Algorithm 1. From this, the algorithm requires two different types of procedures regarding assortment construction, elimination, and exploration. Let $\mathcal{J}(A)$ be the set of all combinations of subset of A with cardinality bound L as $\mathcal{J}(A) = \{B \subseteq A \mid |B| \le L\}$, and let $\mathcal{K}(A)$ be the set of all combinations of subset A (with cardinality bound L) and its element as $\mathcal{K}(A) = \{(b, B) \mid b \in B \subseteq A, |B| \le L\}$. The G/D-optimal design seeks to minimize the ellipsoidal volume under $V_{k,\tau}$, based on arm selection probabilities within the active set of arms $\mathcal{N}_{k,\tau}$. Additionally, since the action space in $H_{k,\tau}(\widehat{\theta}_{k,\tau})$ depends not only on the selection of actions but also on the selection of assortments, the G/D-optimal design incorporates assortment selection probabilities for $\mathcal{J}(\mathcal{N}_{k,\tau})$ and $\mathcal{K}(\mathcal{N}_{k,\tau})$. Following this policy, the algorithm includes two separate exploration procedures regarding the selection of arms and assortments.

Remark 6.1 (Localized design for MNL assortments). Our localized Gram matrix in (4) differs from prior MNL constructions [Goyal and Perivier, 2021, Lee and Oh, 2024]: it avoids the nonconvex optimism step used by Goyal and Perivier [2021] and, unlike Lee and Oh [2024], uses the current estimator rather than aggregating historical estimators. This makes it naturally compatible with our G/D-optimal design. Moreover, unlike linear or generalized linear bandits [Lattimore and Szepesvári, 2020, Sawarni et al., 2024], our design must account for assortment-dependent MNL feedback, and therefore explores a rescaled feature space over assortments rather than individual actions.

We set $\eta_T = (\frac{T}{rK})^{\frac{1}{2(1-2^{-M})}}$ with a parameter for batch update budget $M \ge 1$. Recall that τ_T is the last epoch over T, which indicates the number of batch updates. Then, by following the same proof of Proposition 5.1, we have the following bound for the number of epochs.

Proposition 6.2 (Number of Batch Updates)**.** $\tau_T \leq M$.

Then, we have the following regret bounds (the proof is provided in Appendix A.1).

Theorem 6.3. *Algorithm 2 with $M = \mathcal{O}(\log(T))$ achieves:* $\mathcal{R}(T) = \tilde{\mathcal{O}}\big(rK^{\frac{3}{2}}\sqrt{T}\big(\frac{T}{rK}\big)^{\frac{1}{2(2^M-1)}}\big)$.

Corollary 6.4. *For $M = \Theta\big(\log\log(\frac{T}{rK})\big)$, Algorithm 2 achieves:* $\mathcal{R}(T) = \tilde{\mathcal{O}}\big(rK^{\frac{3}{2}}\sqrt{T}\big)$.

Remark 6.5 (Improvement on κ). Algorithm 2 does not require prior knowledge of κ and removes the leading $1/\kappa = \mathcal{O}(L^2)$ dependence in Theorem 5.2, at the cost of an additional $\sqrt{r}$ regret factor.

Remark 6.6 (Batch Complexity). Like Algorithm 1, this algorithm requires only $\Theta\big(\log\log(\frac{T}{rK})\big)$ batch updates to achieve a $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound. That is, expensive combinatorial optimization is invoked only a small number of times over the entire horizon, rather than at every round.

7 Conclusion

We introduced a stochastic matching bandit framework in which naive approaches require solving an NP-hard combinatorial optimization problem at every round. To address this, we proposed elimination-based batched algorithms that achieve regret $\widetilde{\mathcal{O}}(\kappa^{-1}K^{3/2}\sqrt{rT})$ with only $\Theta(\log\log(T/(rK)))$ optimization updates when κ is known. We also developed a parameter-free variant that removes prior knowledge of κ and attains regret $\widetilde{\mathcal{O}}(rK^{3/2}\sqrt{T})$ under the same update budget. Thus, no-regret learning in stochastic matching bandits can be achieved with rare calls to expensive optimization. Experimental comparisons are provided in Appendix A.12. **Limitations.** An important future direction is to close the remaining $\sqrt{K}$ gap in the regret dependence, or to determine whether this gap is inherent under rare optimization updates.

Acknowledgments

JK was supported in part by the French National Research Agency (ANR) through grants ANR-20-CE23-0007 and ANR-23-CE23-0002 and the PEPR IA FOUNDRY project (ANR23-PEIA-0003), as well as by Hi! PARIS and the ANR/France 2030 program (ANR23-IACL-0005). This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2026-25536608).

MO was supported by the National Research Foundation of Korea (NRF) grant and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant both funded by the Korea government (MSIT) (No. RS-2022-NR071853, RS-2023- 00222663, RS-2025-25463302, RS-2026-25507282, RS-2026-25613008).

References

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR, 2021.

Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Mnl-bandit: A dynamic learning approach to assortment selection. *arXiv preprint arXiv:1706.03880*, 2017a.

Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Thompson sampling for the mnl-bandit. In *Conference on learning theory*, pages 76–78. PMLR, 2017b.

Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.

Soumya Basu, Karthik Abinav Sankararaman, and Abishek Sankararaman. Beyond $\log^2(t)$ regret for decentralized bandits in matching markets. In *International Conference on Machine Learning*, pages 705–715. PMLR, 2021.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013.

Xi Chen, Akshay Krishnamurthy, and Yining Wang. Robust dynamic assortment optimization in the presence of outlier customers. *Operations Research*, 2023.

S Damla Ahipasaoglu, Peng Sun, and Michael J Todd. Linear convergence of a modified frank–wolfe algorithm for computing minimum-volume enclosing ellipsoids. *Optimisation Methods and Software*, 23(1):5–19, 2008.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

Bernhard Fuchs, Winfried Hochstättler, and Walter Kern. Online matching on a line. *Theoretical Computer Science*, 332(1-3):251–264, 2005.

Buddhima Gamlath, Michael Kapralov, Andreas Maggiori, Ola Svensson, and David Wajc. Online matching with general arrivals. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 26–37. IEEE, 2019.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32, 2019.

Vineet Goyal and Noemie Perivier. Dynamic pricing and assortment under a contextual mnl demand. *arXiv preprint arXiv:2110.10018*, 2021.

Yanjun Han, Zhengqing Zhou, Zhengyuan Zhou, Jose Blanchet, Peter W Glynn, and Yinyu Ye. Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*, 2020.

Osama A Hanna, Lin Yang, and Christina Fragouli. Contexts can be cheap: Solving stochastic contextual bandits with linear bandit algorithms. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1791–1821. PMLR, 2023.

Sham M Kakade, Adam Tauman Kalai, and Katrina Ligett. Playing games with approximation algorithms. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 546–555, 2007.

Richard M Karp, Umesh V Vazirani, and Vijay V Vazirani. An optimal algorithm for on-line bipartite matching. In *Proceedings of the twenty-second annual ACM symposium on Theory of computing*, pages 352–358, 1990.

Thomas Kesselheim, Klaus Radke, Andreas Tönnis, and Berthold Vöcking. An optimal online algorithm for weighted bipartite matching and extensions to combinatorial auctions. In *European symposium on algorithms*, pages 589–600. Springer, 2013.

Jung-hun Kim and Min-hwan Oh. Queueing matching bandits with preference feedback. *arXiv preprint arXiv:2410.10098*, 2024.

Jung-hun Kim, Milan Vojnović, and Min-Hwan Oh. Oracle-efficient combinatorial semi-bandits. *arXiv preprint arXiv:2510.21431*, 2025.

Fang Kong and Shuai Li. Player-optimal stable regret for bandit learning in matching markets. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1512–1522. SIAM, 2023.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Joongkyu Lee and Min-hwan Oh. Nearly minimax optimal regret for multinomial logistic bandit. *arXiv preprint arXiv:2405.09831*, 2024.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

Lydia T Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, pages 1618–1628. PMLR, 2020.

Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *The Journal of Machine Learning Research*, 22(1):9612–9645, 2021.

David G McVitie and Leslie B Wilson. The stable marriage problem. *Communications of the ACM*, 14(7):486–490, 1971.

Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and generalized online matching. *Journal of the ACM (JACM)*, 54(5):22–es, 2007.

Aranyak Mehta et al. Online matching and ad allocation. *Foundations and Trends® in Theoretical Computer Science*, 8(4):265–368, 2013.

Sukruta Prakash Midigeshi, Tanmay Goyal, and Gaurav Sinha. Achieving limited adaptivity for multinomial logistic bandits. *arXiv preprint arXiv:2508.03072*, 2025.

Min-hwan Oh and Garud Iyengar. Thompson sampling for multinomial logit contextual bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Min-hwan Oh and Garud Iyengar. Multinomial logit contextual bandits: Provable optimality and practicality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9205–9213, 2021.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. In *Conference on Learning Theory*, pages 1456–1456. PMLR, 2015.

Xuanfei Ren, Tianyuan Jin, and Pan Xu. Optimal batched linear bandits. *arXiv preprint arXiv:2406.04137*, 2024.

Zhimei Ren and Zhengyuan Zhou. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *Management Science*, 70(2):1315–1342, 2024.

Yufei Ruan, Jiaqi Yang, and Yuan Zhou. Linear bandits with limited adaptivity and learning distributional optimal design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–87, 2021.

Abishek Sankararaman, Soumya Basu, and Karthik Abinav Sankararaman. Dominate or delete: Decentralized competing bandits with uniform valuation. *arXiv preprint arXiv:2006.15166*, 2020.

Ayush Sawarni, Nirjhar Das, Siddharth Barman, and Gaurav Sinha. Generalized linear bandits with limited adaptivity. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Yirui Zhang, Siwei Wang, and Zhixuan Fang. Matching in multi-arm bandit with collision. *Advances in Neural Information Processing Systems*, 35:9552–9563, 2022.

A Appendix

A.1 Algorithm Without Prior Knowledge of κ (Algorithm 2)

A.2 Naive Approach by Extending MNL Bandit

For our framework, we can utilize MNL bandit Lee and Oh [2024] by extending it to K-multiple MNLs (Algorithm 3) as follows. Let the negative log-likelihood $l_{k,t}(\theta) = -\sum_{n\in S_{k,t}\cup\{n_0\}} y_{n,k,t}\log p(n|S_{k,t},\theta)$ where $y_{n,k,t}\in\{0,1\}$ is observed preference feedback (1 denotes a choice, and 0 otherwise). Then we define the gradient of the likelihood as

$$g_{k,t}(\theta) := \nabla_\theta l_{k,t}(\theta) = \sum_{n\in S_{k,t}} (p(n|S_{k,t},\theta) - y_{n,k,t})x_n. \tag{6}$$

We also define gram matrices from $\nabla^2_\theta l_{k,t}(\theta)$ as follows:

$$G_{k,t}(\theta) := \sum_{n\in S_{k,t}} p(n|S_{k,t},\theta)z_n z_n^\top - \sum_{n,m\in S_{k,t}} p(n|S_{k,t},\theta)p(m|S_{k,t},\theta)z_n z_m^\top. \tag{7}$$

We define the UCB index for assortment S_k as

$$R^{UCB}_{k,t}(S_k) = \sum_{n\in S_k} \frac{\exp(h_{n,k,t})}{1+\sum_{m\in S_k}\exp(h_{m,k,t})}, \tag{8}$$

where $h_{n,k,t} = z_n^\top\widehat{\theta}_{k,t} + \gamma_t\|z_n\|_{G^{-1}_{k,t}}$ with $\gamma_t = C_4\log(L)\sqrt{d\log(t)\log(KT)}$ for some $C_4>0$. We set $\lambda = C_5 d\log(K)$ and $\eta = C_6\log(K)$ for some $C_5>0$ and $C_6>0$.

Proposition A.1. *Algorithm 3 achieves a regret bound of $\mathcal{R}(T) = \tilde{\mathcal{O}}(rK\sqrt{T})$ and the computational cost per round is $\mathcal{O}(K^N)$.*

Proof. The proof is provided in Appendix A.10. □

Algorithm 3: Extension of `OFU-MNL+` Lee and Oh [2024]

Compute SVD of $X = U\Sigma V^\top$ and obtain $U_r = [u_1,\dots,u_r]$; Construct $z_n \leftarrow U_r^\top x_n$ for $n\in[N]$
for $t = 1,\dots,T$ **do**
  for $k\in[K]$ **do**
    $\tilde{\mathcal{G}}_{k,t} \leftarrow \lambda I_d + \sum_{s=1}^{t-2} G_{k,s}(\widehat{\theta}_{k,s}) + \eta G_{k,t-1}(\widehat{\theta}_{k,t-1})$ with (7)
    $\mathcal{G}_{k,t} \leftarrow \lambda I_d + \sum_{s=1}^{t-1} G_{k,s}(\widehat{\theta}_{k,s})$ with (7)
    $\widehat{\theta}_{k,t} \leftarrow \mathrm{argmin}_{\theta\in\Theta}\, g_{k,t-1}(\widehat{\theta}_{k,t-1})^\top\theta + \frac{1}{2\eta}\|\theta-\widehat{\theta}_{k,t-1}\|^2_{\tilde{\mathcal{G}}^{-1}_{k,t}}$ with (6)
  $\{S_{k,t}\}_{k\in[K]} \leftarrow \underset{\{S_k\}_{k\in[K]}\in\mathcal{M}}{\mathrm{argmax}} \sum_{k\in[K]} R^{UCB}_{k,t}(S_k)$ with (8)
  Offer $\{S_{k,t}\}_{k\in[K]}$ and observe $y_{n,k,t}$ for all $n\in S_{k,t}$, $k\in[K]$

A.3 Details Regarding Projection in Feature Space

Since x_n for $n\in[N]$ lies in the subspace U_r, we observe that $x_n = U_r b_n$ for some $b_n\in\mathbb{R}^r$. Let $\theta^*_k = U_r^\top\theta_k$. Then we have $x_n^\top\theta_k = z_n^\top\theta^*_k$ by following $x_n^\top\theta_k = b_n^\top U_r^\top\theta_k = b_n^\top(U_r^\top U_r)U_r^\top\theta_k = x_n^\top U_r U_r^\top\theta_k = z_n^\top\theta^*_k$ using $U_r^\top U_r = I_d$. Therefore, we can reformulate the MNL model using r-dimensional feature $z_n\in\mathbb{R}^r$ and latent $\theta^*_k\in\mathbb{R}^r$ in place of d-dimensional $x_n\in\mathbb{R}^d$ and $\theta_k\in\mathbb{R}^d$, respectively, for $n\in[N]$ and $k\in[K]$. We note that this procedure is beneficial not only for reducing feature dimension but also for introducing appropriate regularization for estimators without imposing any assumption about feature distributions considered in Oh and Iyengar [2021] (see Lemma A.5).

A.4 Warm-up Stage for Algorithm 1

Let $\lambda_{\min}(A)$ denote the minimum eigenvalue of matrix A. Then we provide the warm-up stage for Algorithm 1 in Algorithm 4.

Algorithm 2: Batched Stochastic Matching Bandit$^+$ (B-SMB$^+$)

Input: $M \geq 1$; **Init:** $t \leftarrow 1, T_1 \leftarrow C_3 \log(T) \log^2(TKL)$ for some constant $C_3 > 0$

Compute SVD of $X = U\Sigma V^\top$ and obtain $U_r = [u_1, \ldots, u_r]$; Construct $z_n \leftarrow U_r^\top x_n$

for $\tau = 1, 2...$ **do**

  for $k \in [K]$ **do**

    $\widehat{\theta}_{k,\tau} \leftarrow \operatorname{argmin}_{\theta \in \mathbb{R}^r : \|\theta\|_2 \leq 1} l_{k,\tau}(\theta)$ with (2) where $\mathcal{T}_{k,\tau-1} = \bigcup_{n \in \mathcal{N}_{k,\tau-1}} \mathcal{T}_{n,k,\tau-1} \bigcup_{J \in \mathcal{J}(\mathcal{N}_{k,\tau-1})} \mathcal{T}_{J,k,\tau-1}$

    // Assortments Construction with (5)

    $\{S_{l,\tau}^{(n,k)}\}_{l \in [K]} \leftarrow \operatorname{argmax}_{\{S_l\}_{l \in [K]} \in \mathcal{M}_{\tau-1} : n \in S_k} \sum_{l \in [K]} R_{l,\tau}^{UCB}(S_l)$ for all $n \in \mathcal{N}_{k,\tau-1}$

    $\{S_{l,\tau}^{(J,k)}\}_{l \in [K]} \leftarrow \operatorname{argmax}_{\{S_l\}_{l \in [K]} \in \mathcal{M}_{\tau-1} : S_k = J} \sum_{l \in [K]} R_{l,\tau}^{UCB}(S_l)$ for all $J \in \mathcal{J}(\mathcal{N}_{k,\tau-1})$

    // Elimination with (5)

    $\mathcal{N}'_{k,\tau} \leftarrow \{n \in \mathcal{N}_{k,\tau-1} : \max_{\{S_l\}_{l \in [K]} \in \mathcal{M}_{\tau-1}} \sum_{l \in [K]} R_{l,\tau}^{LCB}(S_l) \leq \sum_{l \in [K]} R_{l,\tau}^{UCB}(S_{l,\tau}^{(n,k)})\}$

    $\mathcal{N}_{k,\tau} \leftarrow \{n \in J : J \in \mathcal{J}(\mathcal{N}'_{k,\tau}), \max_{\{S_l\}_{l \in [K]} \in \mathcal{M}_{\tau-1}} \sum_{l \in [K]} R_{l,\tau}^{LCB}(S_l) \leq \sum_{l \in [K]} R_{l,\tau}^{UCB}(S_{l,\tau}^{(J,k)})\}$

    // G-Optimal Design

    $\pi_{k,\tau} \leftarrow \operatorname{argmin}_{\pi \in \mathcal{P}(\mathcal{N}_{k,\tau})} \max_{n \in \mathcal{N}_{k,\tau}} \|z_n\|^2_{(\sum_{n \in \mathcal{N}_{k,\tau}} \pi(n) z_n z_n^\top + (\lambda/rT_\tau) I_r)^{-1}}$

    $\tilde{\pi}_{k,\tau} \leftarrow \operatorname*{argmin}_{\pi \in \mathcal{P}(\mathcal{J}(\mathcal{N}_{k,\tau}))} \max_{J \in \mathcal{J}(\mathcal{N}_{k,\tau})} \left\| \sum_{n \in J} \tilde{z}'_{n,k,\tau}(J) \right\|^2_{(\sum_{J \in \mathcal{J}(\mathcal{N}_{k,\tau})} \pi(J) \sum_{n \in J} \tilde{z}'_{n,k,\tau}(J) \tilde{z}'_{n,k,\tau}(J)^\top + (\lambda/T_\tau r) I_r)^{-1}}$ where $\tilde{z}'_{n,k,\tau}(J) = \sqrt{p(n|J, \widehat{\theta}_{k,\tau})} \tilde{z}_{n,k,\tau}(J)$

    $\bar{\pi}_{k,\tau} \leftarrow \operatorname*{argmin}_{\pi \in \mathcal{P}(\mathcal{K}(\mathcal{N}_{k,\tau}))} \max_{(n,J) \in \mathcal{K}(\mathcal{N}_{k,\tau})} \|\tilde{z}_{n,k,\tau}(J)\|^2_{(\sum_{(n,J) \in \mathcal{K}(\mathcal{N}_{k,\tau})} \pi(n,J) \tilde{z}_{n,k,\tau}(J) \tilde{z}_{n,k,\tau}(J)^\top + (\lambda/T_\tau r) I_r)^{-1}}$

    // Exploration

    for $n \in \mathcal{N}_{k,\tau}$ **do**

      $t_{n,k} \leftarrow t, \mathcal{T}_{n,k,\tau} \leftarrow [t_{n,k}, t_{n,k} + \lceil r \pi_{k,\tau}(n) T_\tau \rceil - 1]$

      while $t \in \mathcal{T}_{n,k,\tau}$ **do**

        Offer $\{S_{l,t}\}_{l \in [K]} = \{S_{l,\tau}^{(n,k)}\}_{l \in [K]}$ and observe feedback $y_{m,k,t} \in \{0, 1\}$ for all $m \in S_{l,t}$ and $l \in [K]$

        $t \leftarrow t + 1$

    for $J \in \mathcal{J}(\mathcal{N}_{k,\tau})$ **do**

      $t_{J,k} \leftarrow t, \mathcal{T}_{J,k,\tau} \leftarrow [t_{J,k}, t_{J,k} + \lceil r \tilde{\pi}_{k,\tau}(J) T_\tau \rceil - 1]$

      while $t \in \mathcal{T}_{J,k,\tau}$ **do**

        Offer $\{S_{l,t}\}_{l \in [K]} = \{S_{l,\tau}^{(J,k)}\}_{l \in [K]}$ and observe feedback $y_{m,k,t} \in \{0, 1\}$ for all $m \in S_{l,t}$ and $l \in [K]$

        $t \leftarrow t + 1$

    for $(n, J) \in \mathcal{K}(\mathcal{N}_{k,\tau})$ **do**

      $t_{n,J,k} \leftarrow t, \mathcal{T}_{n,J,k,\tau} \leftarrow [t_{n,J,k}, t_{n,J,k} + \lceil r \bar{\pi}_{k,\tau}(n, J) T_\tau \rceil - 1]$

      while $t \in \mathcal{T}_{n,J,k,\tau}$ **do**

        Offer $\{S_{l,t}\}_{l \in [K]} = \{S_{l,\tau}^{(J,k)}\}_{l \in [K]}$ and observe feedback $y_{m,k,t} \in \{0, 1\}$ for all $m \in S_{l,t}$ and $l \in [K]$

        $t \leftarrow t + 1$

  $\mathcal{M}_\tau \leftarrow \{\{S_k\}_{k \in [K]} : S_k \subseteq \mathcal{N}_{k,\tau}, |S_k| \leq L \; \forall k \in [K], S_k \cap S_l = \emptyset \; \forall k \neq l\}$

  $T_{\tau+1} \leftarrow \eta_T \sqrt{T_\tau}$

Algorithm 4: `Round-robin Warm-up`

$\lambda_{\min} \leftarrow \lambda_{\min}(\sum_{n\in[N]} z_n z_n^\top)$
$t_k \leftarrow t$, $i \leftarrow \min\{L, N\}$
$T'_k \leftarrow \frac{C_3 N}{i\kappa^2 \lambda_{\min} \log(TKN)}(r + \log(TKN))^2$
$\mathcal{T}^{(1)}_{k,\tau} \leftarrow [t_k, t_k + T'_k - 1]$
for $t \in \mathcal{T}^{(1)}_{k,\tau}$ **do**
  $a \leftarrow ((i(t-1)) \bmod N) + 1$, $b \leftarrow ((it-1) \bmod N) + 1$
  if $a \le b$ **then**
    $S_{k,t} \leftarrow [a, b]$
  else
    $S_{k,t} \leftarrow [1, b] \cup [a, N]$
  Construct any $S_{l,t}$ for $l \in [K]/\{k\}$ satisfying $\{S_{k,t}\}_{k\in[K]} \in \mathcal{M}_0$
  Offer $\{S_{k,t}\}_{k\in[K]}$ and observe feedback $y_{n,k,t} \in \{0, 1\}$ for all $n \in S_{k,t}, k \in [K]$

A.5 Proof of Proposition 5.1

Here we utilize the proof techniques in Sawarni et al. [2024]. Recall that τ_T to be the smallest $\tau \in [T]$ such that

$$\sum_{\tau'\in[\tau]} \sum_{k\in[K]} |\mathcal{T}^{(1)}_{k,\tau'}| + |\mathcal{T}^{(2)}_{k,\tau'}| \ge T.$$

In other words, $\sum_{\tau'\in[\tau_T-1]} \sum_{k\in[K]} |\mathcal{T}^{(1)}_{k,\tau'}| + |\mathcal{T}^{(2)}_{k,\tau'}| < T$. Then we can show that $\tau_T \le M$ by contradiction as follows. Suppose $\tau_T > M$. Then, we have

$$T_{\tau_T-1} \ge (\eta_T)^{\sum_{k=1}^{\tau_T-1}(\frac{1}{2})^{k-1}} \ge (\eta_T)^{2(1-(\frac{1}{2})^{\tau_T-1})} = (T/rK)^{\frac{1-2^{1-\tau_T}}{1-2^{-M}}} \ge T/rK,$$

where the last inequality comes from $M + 1 \le \tau_T$. This implies that $\sum_{\tau'\in[\tau_T-1]} \sum_{k\in[K]} |\mathcal{T}^{(1)}_{k,\tau'}| + |\mathcal{T}^{(2)}_{k,\tau'}| \ge KrT_{\tau_T-1} \ge T$, which is contradiction. Thus, we can conclude that $\tau_T \le M$.

A.6 Proof of Theorem 5.2

In the following proof, with a slight abuse of notation, we use $p(n|S, \theta) = \exp(z_n^\top\theta)/(1 + \sum_{m\in S}\exp(z_m^\top\theta))$ with $z_n \in \mathbb{R}^r$ instead of $x_n \in \mathbb{R}^d$. We provide a lemma for a confidence bound.

Lemma A.2. *For any $\tau \in [T]$, $k \in [K]$, and $n \in [N]$, with probability at least $1 - \delta$, for some constant $C > 0$, we have*

$$|z_n^\top(\widehat{\theta}_{k,\tau} - \theta^*_k)| \le \tfrac{C}{\kappa}\sqrt{\|z_n\|^2_{V^{-1}_{k,\tau}} \log(TKN/\delta)}.$$

Proof. We define the gradient of the likelihood as

$$g_{k,\tau}(\theta) := \sum_{t\in\mathcal{T}_{k,\tau}} \nabla_\theta l_{k,t}(\theta) = \sum_{t\in\mathcal{T}_{k,\tau}} \sum_{n\in S_{k,t}} (p(n|S_{k,t}, \theta) - y_{n,k,t})z_n + \theta.$$

Then we first provide a bound in the following lemma.

Lemma A.3. *For any $n \in [N]$, $k \in [K]$, and $\tau \in [T]$, with probability at least $1 - \delta$, we have*

$$|z_n^\top(\widehat{\theta}_{k,\tau} - \theta^*_k)| \le \frac{3\sqrt{\log(TKN/\delta)}}{\kappa}\|z_n\|_{V^{-1}_{k,\tau}} + \frac{6}{\kappa^2}\|\widehat{\theta}_{k,\tau} - \theta^*_k\|_2 \|g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau}(\theta^*_k)\|_{V^{-1}_{k,\tau}} \|z_n\|_{V^{-1}_{k,\tau}}.$$

Proof. The proof is deferred to Appendix A.9.1 □

Then we define

$$E_1 = \Bigg\{ |z_n^\top (\widehat{\theta}_{k,\tau} - \theta_k^*)| \le \frac{3\sqrt{\log(TKN/\delta)}}{\kappa} \|z_n\|_{V_{k,\tau}^{-1}} + \frac{6}{\kappa^2} \|\widehat{\theta}_{k,\tau} - \theta_k^*\|_2 \|g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau}(\theta_k^*)\|_{V_{k,\tau}^{-1}} \|z_n\|_{V_{k,\tau}^{-1}} \ \forall n \in [N], k \in [K], \tau \in [T] \Bigg\},$$

which holds at least $1-\delta$. Now we provide bounds for $\|\widehat{\theta}_{k,\tau} - \theta_k^*\|_2$ and $\|g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau}(\theta_k^*)\|_{V_{k,\tau}^{-1}}$.

Lemma A.4 (Lemma 7 in Li et al. [2017]). *For all $k \in [K]$, $\tau \in [T]$, with probability at least $1-\delta$ for $\delta > 0$, we have*

$$\|g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau-1}(\theta_k^*)\|_{V_{k,\tau}^{-1}} \le 4\sqrt{2r + \log(KTN/\delta)}.$$

We define $V_{k,\tau}^0 = \sum_{t \in \mathcal{T}_{k,\tau-1}^{(1)}} \sum_{n \in S_{k,t}} z_n z_n^\top$. Then we have the following lemma.

Lemma A.5. *For all $k \in [K]$ and $\tau \ge 2$, we have $\lambda_{\min}(V_{k,\tau}^0) \ge \frac{C_0}{\kappa^2 \log(TKN/\delta)}(r^2 + \log^2(TKN/\delta) + 2r\log(TKN/\delta))$.*

Proof. Let $\lambda' = \frac{C_0}{\kappa^2 \lambda_{\min} \log(TKN/\delta)}(r^2 + \log^2(TKN/\delta) + 2r\log(TKN/\delta))$ and recall $\lambda_{\min} = \lambda_{\min}(\sum_{n\in[N]} z_n z_n^\top)$. From the phase in the warm-up stage (Algorithm 4), we can observe that $V_{k,\tau}^0$ contains $z_n z_n^\top$ for each $n \in [N]$ at least λ'. Since $\sum_{n\in[N]} z_n z_n^\top = \sum_{s\in[r]} \lambda_s u_s u_s{}^\top$, we have $V_{k,\tau}^0 = \sum_{t \in \mathcal{T}_{k,\tau-1}^{(1)}} \sum_{n \in S_{k,t}} z_n z_n^\top = \sum_{s\in[r]} \lambda_s' u_s u_s{}^\top$ where $\lambda_s' \ge \lambda' \lambda_s$. Then from $\lambda_{\min} = \lambda_r$, we can conclude $\lambda_{\min}(V_k^0) \ge \lambda' \lambda_{\min}$. □

Lemma A.6 (Lemma 9 in Kveton et al. [2020]). *Suppose $\lambda_{\min}(V_{k,\tau}^0) \ge \max\{(1/4\kappa^2)(r\log(T/r) + 2\log(KTN/\delta)), 1\}$ for all $k \in [K]$. Then, for all $\tau \in [T]$ and $k \in [K]$, we have*

$$\mathbb{P}(\|\widehat{\theta}_{k,\tau} - \theta_k^*\|_2 \ge 1) \le \delta.$$

We define $E_2 = \{\|\widehat{\theta}_{k,\tau} - \theta_k^*\|_2 \le 1 \ \forall k \in [K], \tau \in [T]\}$. Then from Lemmas A.5, A.6, we have $\mathbb{P}(E_2) \ge 1 - \delta$.

We also denote by E_3 the event of $\{\|g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau-1}(\theta_k^*)\|_{V_{k,\tau}^{-1}} \le 4\sqrt{2r + \log(KTN/\delta)} \ \forall \tau \in [T], k \in [K]\}$, which hold with probability at least $1-\delta$ from Lemma A.4.

Lemma A.7. *Under E_2 and E_3, for any $\tau \in [T]$, $k \in [K]$, we have*

$$\|\widehat{\theta}_{k,\tau} - \theta_k^*\|_2 \le \frac{2}{\kappa}\sqrt{\frac{2r + \log(TNK/\delta)}{\lambda_{\min}(V_k^0)}}.$$

Proof. The proof is deferred to Appendix A.9.2

□

Finally, under $E_1 \cap E_2 \cap E_3$ which holds with probability at least $1 - 3\delta$, we have

$$\begin{aligned}
&|z_n^\top (\widehat{\theta}_{k,\tau} - \theta_k^*)| \\
&\le \frac{2\sqrt{\log(TKN/\delta)}}{\kappa} \|z_n\|_{V_{k,\tau}^{-1}} + (6/\kappa^2) \|z_n\|_{V_{k,\tau}^{-1}} \|\widehat{\theta}_{k,\tau} - \theta_k^*\|_2 \|(g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau}(\theta_k^*))\|_{V_{k,\tau}^{-1}} \\
&\le \frac{2\sqrt{\log(TKN/\delta)}}{\kappa} \|z_n\|_{V_{k,\tau}^{-1}} + \frac{48(2r + \log(KTN/\delta))}{\kappa^2 \sqrt{\lambda_{\min}(V_{k,\tau}^0)}} \|z_n\|_{V_{k,\tau}^{-1}} \\
&\le \frac{3\sqrt{\log(TKN/\delta)}}{\kappa} \|z_n\|_{V_{k,\tau}^{-1}} \\
&= (3/\kappa)\sqrt{\|z_n\|_{V_{\tau,k}^{-1}}^2 \log(TKN/\delta)} := \beta(\delta) \|z_n\|_{V_{\tau,k}^{-1}},
\end{aligned}$$

which concludes the proof. □

Then we define event $E = \{|z_n^\top(\widehat{\theta}_{k,\tau} - \theta_k^*)| \le \beta_T \|z_n\|_{V_{k,\tau}^{-1}} \ \forall \tau \in [T], k \in [K], n \in [N]\}$ for some $c_1 > 0$, which holds at least $1 - 1/T$ with Lemma A.2 and $\delta = 1/T$.

Lemma A.8. *Under E, for all $\tau \in [T]$, $k \in [K]$, and $S \subseteq \mathcal{N}_{k,\tau-1}$, we have*

$$0 \le R_{k,\tau}^{UCB}(S) - R_k(S) \le 4\beta_T \max_{n\in S} \|z_n\|_{V_{k,\tau}^{-1}} \text{ and } -4\beta_T \max_{n\in S} \|z_n\|_{V_{k,\tau}^{-1}} \le R_{k,\tau}^{LCB}(S) - R_k(S) \le 0$$

Proof. Let $u_{n,k} = z_n^\top \theta_k^*$, $\widehat{u}_{n,k} = z_n^\top \widehat{\theta}_{k,\tau}$, and $\widehat{R}_{k,\tau}(S) = \frac{\sum_{n\in S} w_{n,k}\exp(\widehat{u}_{n,k})}{1+\sum_{m\in S}\exp(\widehat{u}_{m,k})}$. Then by the mean value theorem, there exists $\bar{u}_{n,k} = (1-c)\widehat{u}_{n,k} + cu_{n,k}$ for some $c \in (0,1)$ satisfying, for any $S \subset \mathcal{N}_{k,\tau-1}$

$$\begin{aligned}
\left|\widehat{R}_{k,\tau}(S) - R_k(S)\right| &= \left|\frac{\sum_{n\in S} w_{n,k}\exp(\widehat{u}_{n,k})}{1+\sum_{m\in S}\exp(\widehat{u}_{n,k})} - \frac{\sum_{n\in S} w_{n,k}\exp(u_{n,k})}{1+\sum_{m\in S}\exp(u_{m,k})}\right| \\
&= \left|\sum_{n\in S} \nabla_{v_n}\left(\frac{\sum_{m\in S} w_{m,k}\exp(v_m)}{1+\sum_{m\in S}\exp(v_m)}\right)\Big|_{v_n=\bar{u}_{n,k}} (\widehat{u}_{n,k} - u_{n,k})\right| \\
&\le \left|\frac{(1+\sum_{n\in S}\exp(\bar{u}_{n,k}))(\sum_{n\in S} w_{n,k}\exp(\bar{u}_{n,k})(\widehat{u}_{n,k}-u_{n,k}))}{(1+\sum_{n\in S}\exp(\bar{u}_{n,k}))^2}\right| \\
&\quad + \left|\frac{(\sum_{n\in S}\exp(\bar{u}_{n,k}))(\sum_{n\in S} w_{n,k}\exp(\bar{u}_{n,k})(\widehat{u}_{n,k}-u_{n,k}))}{(1+\sum_{n\in S}\exp(\bar{u}_{n,k}))^2}\right| \\
&\le 2\sum_{n\in S}\frac{\exp(\bar{u}_{n,k})}{1+\sum_{m\in S}\exp(\bar{u}_{m,k})}|\widehat{u}_{n,k} - u_{n,k}| \\
&\le 2\max_{n\in S}|\widehat{u}_{n,k} - u_{n,k}| \\
&\le 2\beta_T \max_{n\in S}\|z_n\|_{V_{k,\tau}^{-1}},
\end{aligned}$$

where the last inequality is obtained from, under E, $|z_n^\top\theta_k^* - z_n^\top\widehat{\theta}_{k,\tau}| \le \beta_T\|z_n\|_{V_{k,\tau}^{-1}}$. Then, from the definition of $R_{k,\tau}^{UCB}(S)$ and $R_{k,\tau}^{LCB}(S)$, we can conclude the proof. □

In the following, by adopting the proof technique in Chen et al. [2023], we provide a lemma for showing that $\mathcal{M}_\tau$ is likely to contain the optimal assortment.

Lemma A.9. *Under E, $(S_1^*, \dots, S_K^*) \in \mathcal{M}_{\tau-1}$ for all $\tau \in [T]$.*

Proof. Here we use induction for the proof. Suppose that for fixed τ, we have $(S_1^*, \dots, S_K^*) \in \mathcal{M}_\tau$ for all $k \in [K]$. Recall that $\beta_T = (C_1/\kappa)\sqrt{\log(TKN)}$. From Lemma A.8, we have $R_{k,\tau+1}^{UCB}(S) \ge R_k(S)$ and $R_{k,\tau+1}^{LCB}(S) \le R_k(S)$ for any $S \subset [N]$. Then for $k \in [K]$, $n \in S_k^*$, and any $(S_1, .., S_K) \in \mathcal{M}_\tau$, we have

$$\begin{aligned}
\sum_{l\in[K]} R_{l,\tau+1}^{UCB}(S_{l,\tau+1}^{(n,k)}) &\ge \sum_{l\in[K]} R_{l,\tau+1}^{UCB}(S_l^*) \\
&\ge \sum_{l\in[K]} R_l(S_l^*) \\
&\ge \sum_{l\in[K]} R_l(S_l) \\
&\ge \sum_{l\in[K]} R_{l,\tau+1}^{LCB}(S_l), \qquad (9)
\end{aligned}$$

where the first inequality comes from the elimination condition in the algorithm and $(S_1^*, \dots S_K^*) \in \mathcal{M}_\tau$, and the third inequality comes from the optimality of $(S_1^*, \dots, S_K^*)$. This implies that $n \in \mathcal{N}_{k,\tau+1}$ from the algorithm. Then by following the same statement of (9) for all $n \in S_k^*$ and $k \in [K]$, we have $S_k^* \subset \mathcal{N}_{k,\tau+1}$ for all $k \in [K]$, which implies $(S_1^*, \dots, S_K^*) \in \mathcal{M}_{\tau+1}$. Therefore, with $(S_1^*, \dots, S_K^*) \in \mathcal{M}_1$, we can conclude the proof from the induction. □

From the above Lemmas A.8 and A.9, under E, we have

$$\begin{aligned}\sum_{l\in[K]} R_l(S_l^*) - \sum_{l\in[K]} R_l(S_{l,\tau}^{(n,k)}) &\le \sum_{l\in[K]} R_{l,\tau}^{LCB}(S_l^*) + 4\beta_T \max_{m\in S_l^*} \|z_m\|_{V_{l,\tau}^{-1}} \\ &\quad - \sum_{l\in[K]} R_{l,\tau-1}^{UCB}(S_{l,\tau}^{(n,k)}) + 4\beta_T \max_{m\in S_{l,\tau}^{(n,k)}} \|z_m\|_{V_{l,\tau-1}^{-1}} \\ &\le 4\beta_T \sum_{l\in[K]} \Big(\max_{m\in S_l^*} \|z_m\|_{V_{l,\tau-1}^{-1}} + \max_{m\in S_{l,\tau}^{(n,k)}} \|z_m\|_{V_{l,\tau-1}^{-1}}\Big), \quad (10)\end{aligned}$$

where the last inequality comes from the fact that $(S_1^*, \dots, S_K^*) \in \mathcal{M}_{\tau-1}$ and $\max_{(S_1,\dots,S_K)\in\mathcal{M}_{\tau-1}} \sum_{l\in[K]} R_{l,\tau}^{LCB}(S_l) \le \sum_{l\in[K]} R_{l,\tau}^{UCB}(S_{l,\tau}^{(n,k)})$ from the algorithm.

We define $V(\pi_{k,\tau}) = \sum_{n\in\mathcal{N}_{k,\tau}} \pi_{k,\tau}(n) z_n z_n^\top$ and $\mathrm{supp}(\pi_{k,\tau}) = \{n \in \mathcal{N}_{k,\tau} : \pi_{k,\tau}(n) \neq 0\}$. Then we have the following lemma from the G/D-optimal design problem.

Lemma A.10 (Theorem 21.1 (Kiefer-Wolfowitz) in Lattimore and Szepesvári [2020])**.** *For all $\tau \in [T]$ and $k \in [K]$, we have*

$$\max_{n\in\mathcal{N}_{k,\tau}} \|z_n\|^2_{(V(\pi_{k,\tau})+(1/rT_\tau)I_r)^{-1}} \le r \text{ and } |\mathrm{supp}(\pi_{k,\tau})| \le r(r+1)/2.$$

Proof. For completeness, we provide a proof in Appendix A.11. □

From the definition of $V_{k,\tau}$ and T_τ, we have

$$\begin{aligned}V_{k,\tau} &\succeq \sum_{n\in\mathcal{N}_{k,\tau-1}} r\pi_{k,\tau-1}(n) T_{\tau-1} z_n z_n^\top + I_r \\ &= T_{\tau-1} r (V(\pi_{k,\tau-1}) + (1/T_{\tau-1}r) I_r). \quad (11)\end{aligned}$$

Then from Lemma A.10 and (11), for any $n \in \mathcal{N}_{k,\tau}$ we have

$$\begin{aligned}\beta_T \|z_n\|_{V_{k,\tau}^{-1}} &= (1/\kappa)\sqrt{\|z_n\|^2_{V_{k,\tau}^{-1}} \log(KNT)} \\ &= \tilde{\mathcal{O}}\left((1/\kappa)\sqrt{1/T_{\tau-1}}\sqrt{\|z_n\|^2_{(V(\pi_{k,\tau-1})+(1/T_{\tau-1}r)I_r)^{-1}}/r}\right) \\ &= \tilde{\mathcal{O}}((1/\kappa)\sqrt{1/T_{\tau-1}}). \quad (12)\end{aligned}$$

Therefore under E, from (10) and (12), for $\tau > 1$, we have

$$\sum_{l\in[K]} (R_l(S_l^*) - R_l(S_{l,\tau}^{(n,k)})) = \tilde{\mathcal{O}}((1/\kappa)K\sqrt{1/T_{\tau-1}}).$$

We have

$$\begin{aligned}\mathcal{R}(T) &= \mathbb{E}\left[\sum_{t\in[T]} \sum_{k\in[K]} R_k(S_k^*) - R_k(S_{k,t})\right] \\ &\le \mathbb{E}\left[\sum_{\tau\in[\tau_T]} \sum_{l\in[K]} \sum_{t\in\mathcal{T}_{l,\tau}^{(1)} \bigcap \mathcal{T}_{l,\tau}^{(2)}} \sum_{k\in[K]} R_k(S_k^*) - R_k(S_{k,t})\right], \quad (13)\end{aligned}$$

which consists of regret from the stage of warming up and main. We first analyze the regret from the warming-up as follows:

$$\begin{aligned}\mathbb{E}\left[\sum_{\tau\in[\tau_T]} \sum_{l\in[K]} \sum_{t\in\mathcal{T}_{l,\tau}^{(1)}} \sum_{k\in[K]} R_k(S_k^*) - R_k(S_{k,t})\right] &\le \mathbb{E}\left[\sum_{\tau\in[\tau_T]} \sum_{l\in[K]} K\left|\mathcal{T}_{l,\tau}^{(1)}\right|\right] \\ &= \tilde{\mathcal{O}}(r^2K^2N/(\min\{L,N\}\kappa^2\lambda_{\min})), \quad (14)\end{aligned}$$

where the first equality comes from $\tau_T \leq M = \mathcal{O}(\log(\log(T/rK)))$ from Proposition 5.1. For the regret bound from the main part of the algorithm, with large enough T, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{\tau\in[\tau_T]}\sum_{l\in[K]}\sum_{t\in\mathcal{T}^{(2)}_{l,\tau_T}}\sum_{k\in[K]} R_k(S^*_k) - R_t(S_{k,t})\right] \\
&\leq \mathbb{E}\left[\sum_{\tau\in[\tau_T]}\sum_{l\in[K]}\sum_{t\in\mathcal{T}^{(2)}_{l,\tau}}\sum_{k\in[K]} \left(R_k(S^*_k) - R_k(S_{k,t})\right)\mathbf{1}(E)\right] \\
&\quad + \mathbb{E}\left[\sum_{\tau\in[\tau_T]}\sum_{l\in[K]}\sum_{t\in\mathcal{T}^{(2)}_{l,\tau}}\sum_{k\in[K]} \left(R_k(S^*_k) - R_k(S_{k,t})\right)\mathbf{1}(E^c)\right] \\
&= \tilde{\mathcal{O}}\left((K/\kappa)\sum_{\tau=2}^{\tau_T}\sum_{l\in[K]}\sum_{n\in\mathcal{N}_{l,\tau}} |\mathcal{T}^{(2)}_{n,l,\tau}|\sqrt{1/T_{\tau-1}}\right) + \mathcal{O}(rK\eta_T) + \mathcal{O}(K) \\
&= \tilde{\mathcal{O}}\left((K/\kappa)\sum_{\tau=2}^{\tau_T}\sum_{l\in[K]}\sum_{n\in\mathcal{N}_{l,\tau}} |\mathcal{T}^{(2)}_{n,l,\tau}|\sqrt{1/T_{\tau-1}}\right) + \mathcal{O}(rK\eta_T) \\
&= \tilde{\mathcal{O}}\left((K/\kappa)\sum_{\tau=2}^{\tau_T}\sum_{l\in[K]} (rT_\tau + |\mathrm{supp}(\pi_{l,\tau})|)\sqrt{1/T_{\tau-1}}\right) + \mathcal{O}(rK\eta_T) \\
&= \tilde{\mathcal{O}}\left((K^2/\kappa)\sum_{\tau=2}^{\tau_T}(r\eta_T + r^2\sqrt{1/T_{\tau-1}})\right) \\
&= \tilde{\mathcal{O}}\left((K^2/\kappa)(r\eta_T + r^2)\right) \\
&= \tilde{\mathcal{O}}\left(\tfrac{1}{\kappa}rK^2(T/rK)^{\frac{1}{2(1-2^{-M})}}\right), \qquad (15)
\end{aligned}
$$

where the third last equality comes from Lemma A.10 and the second last equality comes from $\tau_T \leq M = \mathcal{O}(\log(\log(T/rK)))$ from Proposition 5.1. From (13), (14), (15), for $T \geq r^3KN^2/\min\{L,N\}^2\kappa^2\lambda^2_{\min}$, we can conclude the proof.

A.7 Proof of Theorem 6.3

Let $g_{k,\tau}(\theta) = \sum_{t\in\mathcal{T}_{\tau-1}}\sum_{n\in S_{k,t}} p(n|S_{k,t},\theta)z_n + \lambda\theta$ and $\zeta_\tau(\delta) = \frac{1}{2}\sqrt{\lambda} + \frac{2r}{\sqrt{\lambda}}\log\left(\frac{4K}{\delta}\left(1+\frac{2(t_\tau-1)L}{r\lambda}\right)\right)$.

Lemma A.11 (Proposition 2 in Goyal and Perivier [2021])**.** *With probability at least $1-\delta$, for all $\tau \geq 1$ and $k \in [K]$, we have*

$$\|g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau}(\theta^*_k)\|_{H^{-1}_{k,\tau}(\theta^*_k)} \leq \zeta_\tau(\delta).$$

From the above lemma, we define event $E = \{\|g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau}(\theta^*_k)\|_{H^{-1}_{k,\tau}(\theta^*_k)} \leq \zeta_\tau(\delta),\ \forall\tau \geq 1, k \in [K]\}$. Then we have the following lemma.

Lemma A.12. *Under E, for any $\tau \geq 1$ and $k \in [K]$, we have*

$$\|\widehat{\theta}_{k,\tau} - \theta^*_k\|_{H_{k,\tau}(\widehat{\theta}_{k,\tau})} \leq (1+3\sqrt{2})\zeta_\tau(\delta).$$

Proof. Here we utilize the proof techniques in Goyal and Perivier [2021]. Let $G_{k,\tau}(\theta_1, \theta_2) = \int_{v=0}^{1} \nabla g_{k,\tau}(\theta_1 + v(\theta_2 - \theta_1))dv$. By the multivariate mean value theorem, we have

$$g_{k,\tau}(\theta_1) - g_{k,\tau}(\theta_2) = \int_{v=0}^{1} \nabla g_{k,\tau}(\theta_1 + v(\theta_2 - \theta_1))dv(\theta_1 - \theta_2) = G_{k,\tau}(\theta_1, \theta_2)(\theta_1 - \theta_2), \quad (16)$$

which implies

$$\|g_{k,\tau}(\theta_1) - g_{k,\tau}(\theta_2)\|_{G_{k,\tau}^{-1}(\theta_1,\theta_2)} = \|\theta_1 - \theta_2\|_{G_{k,\tau}(\theta_1,\theta_2)}.$$

By following the proof steps of Proposition 3 in Goyal and Perivier [2021] with Proposition C.1 in Lee and Oh [2024], we can show that

$$G_{k,\tau}(\theta_1, \theta_2) \succeq \frac{1}{1+3\sqrt{2}} H_{k,\tau}(\theta_1) \text{ and } G_{k,\tau}(\theta_1, \theta_2) \succeq \frac{1}{1+3\sqrt{2}} H_{k,\tau}(\theta_2).$$

Finally, we have

$$\begin{aligned}
\|\theta_1 - \theta_2\|_{H_{k,\tau}(\theta_1)} &\leq (1+3\sqrt{2})^{1/2} \|\theta_1 - \theta_2\|_{G_{k,\tau}(\theta_1,\theta_2)} \\
&= (1+3\sqrt{2})^{1/2} \|g_{k,\tau}(\theta_1) - g_{k,\tau}(\theta_2)\|_{G_{k,\tau}^{-1}(\theta_1,\theta_2)} \\
&\leq (1+3\sqrt{2}) \|g_{k,\tau}(\theta_1) - g_{k,\tau}(\theta_2)\|_{H_{k,\tau}^{-1}(\theta_2)},
\end{aligned}$$

which concludes the proof with E.

□

From the above lemma and E with $\delta = 1/T$, with probability at least $1 - (1/T)$, for all $\tau \geq 1$ and $k \in [K]$, we have

$$|z_n^\top (\widehat{\theta}_{k,\tau} - \theta_k^*)| \leq \|z_n\|_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})} \|\widehat{\theta}_{k,\tau} - \theta_k^*\|_{H_{k,\tau}(\widehat{\theta}_{k,\tau})} \leq \zeta_\tau \|z_n\|_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})}.$$

In the following proof, with a slight abuse of notation, we define $E = \{|z_n^\top (\widehat{\theta}_{k,\tau} - \theta_k^*)| \leq \zeta_\tau \|z_n\|_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})} \; \forall \tau \geq 1, k \in [K], n \in [N]\}$, which holds at least $1 - (1/T)$. We also use $p(n|S,\theta) = \exp(z_n^\top \theta)/(1 + \sum_{m \in S} \exp(z_m^\top \theta))$ with z_n instead of x_n.

Lemma A.13. *Under E, for all $k \in [K]$ and $\tau \in [T]$, for any $S \subset \mathcal{N}_{k,\tau-1}$, we have*

$$\begin{aligned}
&0 \leq R_{k,\tau}^{UCB}(S) - R_k(S) \\
&\leq 13\zeta_\tau^2 \max_{n \in S} \|z_n\|^2_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})} + 4\zeta_\tau^2 \max_{n \in S} \|\tilde{z}_{n,k,\tau}\|^2_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})} + 2\zeta_\tau \sum_{n \in S} p(n|S, \widehat{\theta}_{k,\tau-1}) \|\tilde{z}_{n,k,\tau}\|_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})},
\end{aligned}$$

$$\begin{aligned}
&0 \leq R_k(S) - R_{k,\tau}^{LCB}(S) \\
&\leq 13\zeta_\tau^2 \max_{n \in S} \|z_n\|^2_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})} + 4\zeta_\tau^2 \max_{n \in S} \|\tilde{z}_{n,k,\tau}\|_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})} + 2\zeta_\tau \sum_{n \in S} p(n|S, \widehat{\theta}_{k,\tau-1}) \|\tilde{z}_{n,k,\tau}\|_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})}.
\end{aligned}$$

Proof. Let $u_{n,k} = z_n^\top \theta_k^*$, $\widehat{u}_{n,k} = z_n^\top \widehat{\theta}_{k,\tau}$, and $\widehat{R}_{k,\tau}(S) = \frac{\sum_{n \in S} w_{n,k} \exp(\widehat{u}_{n,k})}{1 + \sum_{m \in S} \exp(\widehat{u}_{m,k})}$. We also define $u_{n,k} = z_n^\top \theta_k^*$, $\mathbf{u}_k = (u_{n,k} : n \in S)$, $\widehat{\mathbf{u}}_{k,\tau} = (\widehat{u}_{n,k,\tau} : n \in S)$, and $Q(\mathbf{v}) = \sum_{n \in S} \frac{w_{n,k} \exp(v_n)}{1 + \sum_{m \in S} \exp(v_m)}$. Then by a second-order Taylor expansion, we have

$$\begin{aligned}
\left| \widehat{R}_{k,\tau}(S) - R_k(S) \right| &= |Q(\widehat{\mathbf{u}}_{k,\tau}) - Q(\mathbf{u}_k)| \\
&= \left| \nabla Q(\mathbf{u}_k)^\top (\widehat{\mathbf{u}}_{k,\tau} - \mathbf{u}_k) \right| + \left| \frac{1}{2} (\widehat{\mathbf{u}}_{k,\tau} - \mathbf{u}_k)^\top \nabla^2 Q(\bar{\mathbf{u}}_k) (\widehat{\mathbf{u}}_{k,\tau} - \mathbf{u}_k) \right|, \quad (17)
\end{aligned}$$

where $\bar{\mathbf{u}}_k$ is the convex combination of $\widehat{\mathbf{u}}_{k,\tau}$ and $\mathbf{u}_k$. Let $e_{n,k,\tau} = \widehat{u}_{n,k,\tau} - u_{n,k}$, $e_{n_0,k,\tau} = 0$, $\bar{e}_{n,k,\tau} = e_{n,k,\tau} - \sum_{m \in S \cup \{n_0\}} p(m|S, \theta_k^*) e_{m,k,\tau} = e_{n,k,\tau} - \mathbb{E}_{\theta_k^*}[e_{m,k,\tau}]$, and $\tilde{e}_{n,k,\tau} = e_{n,k,\tau} -$

$\sum_{m\in S\cup\{n_0\}} p(m|S,\widehat{\theta}_{k,\tau})e_{m,k,\tau} = e_{n,k,\tau} - \mathbb{E}_{\widehat{\theta}_{k,\tau}}[e_{m,k,\tau}]$. Then the first-order term in the above is bounded as

$$
\begin{aligned}
&\left|\nabla Q(\mathbf{u}_k)^\top(\widehat{\mathbf{u}}_{k,\tau} - \mathbf{u}_k)\right| \\
&= \left|\frac{\sum_{n\in S} w_{n,k}\exp(u_{n,k})(\widehat{u}_{n,k,\tau} - u_{n,k})}{1+\sum_{n\in S}\exp(u_{n,k})} - \frac{(\sum_{n\in S} w_{n,k}\exp(u_{n,k}))(\sum_{n\in S}\exp(u_{n,k})(\widehat{u}_{n,k,\tau} - u_{n,k}))}{(1+\sum_{n\in S}\exp(u_{n,k}))^2}\right| \\
&= \left|\sum_{n\in S} w_{n,k}p(n|S,\theta_k^*)(\widehat{u}_{n,k,\tau} - u_{n,k}) - \sum_{n,m\in S} w_{m,k}p(n|S,\theta_k^*)p(m|S,\theta_k^*)(\widehat{u}_{n,k,\tau} - u_{n,k})\right| \\
&= \left|\sum_{n\in S} w_{n,k}p(n|S,\theta_k^*)\left((\widehat{u}_{n,k,\tau} - u_{n,k}) - \sum_{m\in S} p(m|S,\theta_k^*)(\widehat{u}_{m,k,\tau} - u_{m,k})\right)\right| \\
&\le \sum_{n\in S} w_{n,k}p(n|S,\theta_k^*)\left|e_{n,k,\tau} - \mathbb{E}_{\theta_k^*}[e_{m,k,\tau}]\right| \\
&\le \sum_{n\in S} p(n|S,\theta_k^*)\left|e_{n,k,\tau} - \mathbb{E}_{\theta_k^*}[e_{m,k,\tau}]\right| \\
&= \sum_{n\in S} p(n|S,\theta_k^*)\left|\bar{e}_{n,k,\tau}\right| \\
&\le \sum_{n\in S} p(n|S,\theta_k^*)\left|\bar{e}_{n,k,\tau} - \tilde{e}_{n,k,\tau}\right| + \sum_{n\in S} p(n|S,\theta_k^*)\left|\tilde{e}_{n,k,\tau}\right|
\end{aligned}
$$

For the first term above, we have

$$
\begin{aligned}
&\sum_{n\in S} p(n|S,\theta_k^*)\left|\bar{e}_{n,k,\tau} - \tilde{e}_{n,k,\tau}\right| \\
&= \sum_{n\in S} p(n|S,\theta_k^*)\left|\mathbb{E}_{\theta_k^*}[e_{m,k,\tau}] - \mathbb{E}_{\widehat{\theta}_{k,\tau}}[e_{m,k,\tau}]\right| \\
&= \sum_{n\in S} p(n|S,\theta_k^*)\left|\sum_{m\in S}(p(m|S,\theta_k^*) - p(m|S,\widehat{\theta}_{k,\tau}))e_{m,k,\tau}\right| \\
&\le 2\zeta_\tau^2 \sum_{n\in S} p(n|S,\theta_k^*)\|z_n\|^2_{H_{k,\tau}^{-1}} \\
&\le 2\zeta_\tau^2 \max_{n\in S}\|z_n\|^2_{H_{k,\tau}^{-1}},
\end{aligned}
$$

where the first inequality is obtained by using the mean value theorem. Then for the second term, we have

$$
\begin{aligned}
\sum_{n\in S} p(n|S,\theta_k^*)|\tilde{e}_{n,k,\tau}| &\le \sum_{n\in S}(p(n|S,\theta_k^*) - p(n|S,\widehat{\theta}_{k,\tau-1}))|\tilde{e}_{n,k,\tau}| + \sum_{n\in S} p(n|S,\widehat{\theta}_{k,\tau-1})|\tilde{e}_{n,k,\tau}| \\
&\le 2\zeta_\tau \max_{n\in S}\|z_n\|_{H_{k,\tau}^{-1}}|(\widehat{\theta}_{k,\tau} - \theta_k^*)^\top(z_n - \mathbb{E}_{\widehat{\theta}_{k,\tau}}[z_n])| \\
&\quad + \sum_{n\in S} p(n|S,\widehat{\theta}_{k,\tau-1})|(\widehat{\theta}_{k,\tau} - \theta_k^*)^\top(z_n - \mathbb{E}_{\widehat{\theta}_{k,\tau}}[z_n])| \\
&\le 2\zeta_\tau^2(\max_{n\in S}\|z_n\|^2_{H_{k,\tau}^{-1}} + \max_{n\in S}\|\tilde{z}_{n,k,\tau}\|^2_{H_{k,\tau}^{-1}}) + \zeta_\tau\sum_{n\in S} p(n|S,\widehat{\theta}_{k,\tau-1})\|\tilde{z}_{n,k,\tau}\|_{H_{k,\tau}^{-1}}.
\end{aligned}
$$

From the above inequalities, we have

$$
\left|\nabla Q(\mathbf{u}_k)^\top(\widehat{\mathbf{u}}_{k,\tau} - \mathbf{u}_k)\right| \le 4\zeta_\tau^2 \max_{n\in S}\|z_n\|^2_{H_{k,\tau}^{-1}} + 2\zeta_\tau^2 \max_{n\in S}\|\tilde{z}_{n,k,\tau}\|^2_{H_{k,\tau}^{-1}} + \zeta_\tau\sum_{n\in S} p(n|S,\widehat{\theta}_{k,\tau-1})\|\tilde{z}_{n,k,\tau}\|_{H_{k,\tau}^{-1}}. \tag{18}
$$

Now we focus on the second-order term which is bounded as

$$
\begin{aligned}
&\left|\frac{1}{2}(\widehat{\mathbf{u}}_{k,\tau}-\mathbf{u}_k)^\top\nabla^2 Q(\bar{\mathbf{u}}_k)(\widehat{\mathbf{u}}_{k,\tau}-\mathbf{u}_k)\right| \\
&=\left|\frac{1}{2}\sum_{n,m\in S}(\widehat{u}_{n,k,\tau}-u_{n,k})\frac{\partial^2 Q(\bar{\mathbf{u}}_k)}{\partial_n\partial_m}(\widehat{u}_{m,k,\tau}-u_{m,k})\right| \\
&=\left|\frac{1}{2}\sum_{n,m\in S}(\widehat{u}_{n,k,\tau}-u_{n,k})\frac{\partial^2 Q(\bar{\mathbf{u}}_k)}{\partial_n\partial_m}(\widehat{u}_{m,k,\tau}-u_{m,k})+\frac{1}{2}\sum_{n,m\in S}(\widehat{u}_{n,k,\tau}-u_{n,k})\frac{\partial^2 Q(\bar{\mathbf{u}}_k)}{\partial_n\partial_m}(\widehat{u}_{m,k,\tau}-u_{m,k})\right| \\
&\le\sum_{n,m\in S}|\widehat{u}_{n,k,\tau}-u_{n,k}|\frac{\exp(\bar{u}_{n,k})}{1+\sum_{l\in S}\exp(\bar{u}_{l,k})}\frac{\exp(\bar{u}_{m,k})}{1+\sum_{l\in S}\exp(\bar{u}_{l,k})}|\widehat{u}_{m,k,\tau}-u_{m,k}| \\
&\quad+\frac{3}{2}\sum_{n\in S}(\widehat{u}_{n,k,\tau}-u_{n,k})^2\frac{\exp(\bar{u}_{n,k})}{1+\sum_{l\in S}\exp(\bar{u}_{l,k})} \\
&\le\frac{5}{2}\sum_{n\in S}(\widehat{u}_{n,k,\tau}-u_{n,k})^2\frac{\exp(\bar{u}_{n,k})}{1+\sum_{l\in S}\exp(\bar{u}_{l,k})} \\
&\le\frac{5}{2}\zeta_\tau^2\max_{n\in S}\|z_n\|^2_{H^{-1}_{k,\tau}(\widehat{\theta}_{k,\tau})},
\end{aligned}
\tag{19}
$$

where the first inequality is obtained from Lemma A.22 and the second inequality is obtained from AM-GM inequality. Then from (17), (18), (19), and with the definition of $R^{UCB}_{k,\tau}(S)$ and $R^{LCB}_{k,\tau}(S)$, we can conclude the proof. □

In the following, similar to Lemma A.9, we provide a lemma for showing that $\mathcal{M}_\tau$ is likely to contain the optimal assortment.

Lemma A.14. *Under E, $(S^*_1,\dots,S^*_K)\in\mathcal{M}_{\tau-1}$ for all $\tau\in[T]$.*

Proof. Here we use induction for the proof. Suppose that for fixed τ, we have $(S^*_1,\dots,S^*_K)\in\mathcal{M}_\tau$ for all $k\in[K]$. From E, we have $R^{UCB}_{k,\tau+1}(S)\ge R_k(S)$ and $R^{LCB}_{k,\tau+1}(S)\le R_k(S)$ for any $S\subset[N]$. Then for $k\in[K]$, $n\in S^*_k$, and any $(S_1,..,S_K)\in\mathcal{M}_\tau$, we have

$$
\begin{aligned}
\sum_{l\in[K]}R^{UCB}_{l,\tau+1}(S^{(n,k)}_{l,\tau+1}) &\ge \sum_{l\in[K]}R^{UCB}_{l,\tau+1}(S^*_l) \\
&\ge \sum_{l\in[K]}R_l(S^*_l) \\
&\ge \sum_{l\in[K]}R_l(S_l) \\
&\ge \sum_{l\in[K]}R^{LCB}_{l,\tau+1}(S_l),
\end{aligned}
\tag{20}
$$

where the first inequality comes from the elimination condition in the algorithm and $(S^*_1,\dots S^*_K)\in\mathcal{M}_\tau$, and the third inequality comes from the optimality of $(S^*_1,\dots,S^*_K)$. This implies that $n\in\mathcal{N}'_{k,\tau+1}$ from the algorithm. Then by following the same statement of (20) for all $n\in S^*_k$ and $k\in[K]$, we have $S^*_k\subseteq\mathcal{N}'_{k,\tau+1}$ for all $k\in[K]$.

Then for $k \in [K]$, $J = S_k^*$, and any $(S_1, .., S_K) \in \mathcal{M}_\tau$, we have

$$\begin{aligned}\sum_{l\in[K]} R_{l,\tau+1}^{UCB}(S_{l,\tau+1}^J) &\geq \sum_{l\in[K]} R_{l,\tau+1}^{UCB}(S_l^*) \\ &\geq \sum_{l\in[K]} R_l(S_l^*) \\ &\geq \sum_{l\in[K]} R_l(S_l) \\ &\geq \sum_{l\in[K]} R_{l,\tau+1}^{LCB}(S_l), \end{aligned} \tag{21}$$

where the first inequality comes from the elimination condition in the algorithm and $(S_1^*, \dots S_K^*) \in \mathcal{M}_\tau$, and the third inequality comes from the optimality of $(S_1^*, \dots, S_K^*)$. This implies that $J(= S_k^*) \in \mathcal{J}(\mathcal{N}'_{k,\tau+1})$ from the algorithm. Then by following the same statement of (21) for all $k \in [K]$, we have $S_k^* \subseteq \mathcal{N}_{k,\tau+1}$ for all $k \in [K]$, which implies $(S_1^*, \dots, S_K^*) \in \mathcal{M}_{\tau+1}$. Therefore, with $(S_1^*, \dots, S_K^*) \in \mathcal{M}_1$, we can conclude the proof from the induction.

□

We define $\bar{V}(\bar{\pi}_{k,\tau}) = \sum_{n\in J\in\mathcal{J}_{k,\tau}} \bar{\pi}_{k,\tau}(n,J)\tilde{z}_{n,k,\tau}(J)\tilde{z}_{n,k,\tau}(J)^\top$ and $\tilde{V}(\tilde{\pi}_{k,\tau}) = \sum_{J\in\mathcal{J}_{k,\tau}} \tilde{\pi}_{k,\tau}(J) \sum_{n\in J} p(n|J, \widehat{\theta}_{k,\tau})\tilde{z}_{n,k,\tau}(J)\tilde{z}_{n,k,\tau}(J)^\top$. Then we have the following lemma from the G/D-optimal design problem.

Lemma A.15 (Kiefer-Wolfowitz)**.** *For all $\tau \in [T]$ and $k \in [K]$, we have*

$$\max_{n\in J\in\mathcal{J}(\mathcal{N}_{k,\tau})} \|\tilde{z}_{n,k,\tau}(J)\|^2_{(\bar{V}(\bar{\pi}_{k,\tau})+(\lambda/T_\tau r)I_r)^{-1}} \leq r \text{ and } |\mathrm{supp}(\bar{\pi}_{k,\tau})| \leq r(r+1)/2,$$

$$\max_{J\in\mathcal{J}(\mathcal{N}_{k,\tau})} \sum_{n\in J} p(n|J, \widehat{\theta}_{k,\tau})\|\tilde{z}_{n,k,\tau}(J)\|^2_{(\tilde{V}(\tilde{\pi}_{k,\tau})+(\lambda/T_\tau r)I_r)^{-1}} \leq r \text{ and } |\mathrm{supp}(\tilde{\pi}_{k,\tau})| \leq r(r+1)/2.$$

Proof. Fix $\tau \in [T]$ and $k \in [K]$, and let

$$\alpha := \frac{\lambda}{T_\tau r}.$$

(1) The bound for $\bar{\pi}_{k,\tau}$. Recall the design space

$$\mathcal{K}(\mathcal{N}_{k,\tau}) := \{(n, J) : J \in \mathcal{J}(\mathcal{N}_{k,\tau}),\ n \in J\}.$$

For a distribution $\bar{\pi}$ over $\mathcal{K}(\mathcal{N}_{k,\tau})$, define

$$\bar{V}(\bar{\pi}) := \sum_{(n,J)\in\mathcal{K}(\mathcal{N}_{k,\tau})} \bar{\pi}(n,J)\, \tilde{z}_{n,k,\tau}(J)\tilde{z}_{n,k,\tau}(J)^\top, \qquad \bar{W}(\bar{\pi}) := \bar{V}(\bar{\pi}) + \alpha I_r,$$

and the criterion

$$\bar{g}(\bar{\pi}) := \max_{(n,J)\in\mathcal{K}(\mathcal{N}_{k,\tau})} \tilde{z}_{n,k,\tau}(J)^\top \bar{W}(\bar{\pi})^{-1} \tilde{z}_{n,k,\tau}(J) = \max_{n\in J\in\mathcal{J}(\mathcal{N}_{k,\tau})} \|\tilde{z}_{n,k,\tau}(J)\|^2_{\bar{W}(\bar{\pi})^{-1}}.$$

Let $\bar{\pi}_{k,\tau}$ be a minimizer of $\bar{g}(\cdot)$.

Step 1 (equalization on the support). As in Lemma A.10, we may assume that for every $(n, J) \in \mathrm{supp}(\bar{\pi}_{k,\tau})$,

$$\tilde{z}_{n,k,\tau}(J)^\top \bar{W}(\bar{\pi}_{k,\tau})^{-1} \tilde{z}_{n,k,\tau}(J) = \bar{g}(\bar{\pi}_{k,\tau}). \tag{22}$$

(Otherwise, if some $(n_0, J_0) \in \mathrm{supp}(\bar{\pi}_{k,\tau})$ is strictly below the maximum, one can redistribute an arbitrarily small amount of its probability mass to points attaining the maximum and (by continuity of $\bar{g}$) obtain another feasible design with strictly smaller objective, contradicting optimality. This is the same argument used in Lemma A.10.)

Step 2 (trace bound, giving $\bar{g}(\bar{\pi}_{k,\tau}) \le r$). Using (22) and $\sum_{(n,J)} \bar{\pi}_{k,\tau}(n,J) = 1$,

$$\begin{aligned}
\bar{g}(\bar{\pi}_{k,\tau}) &= \sum_{(n,J)} \bar{\pi}_{k,\tau}(n,J)\, \tilde{z}_{n,k,\tau}(J)^\top \bar{W}(\bar{\pi}_{k,\tau})^{-1} \tilde{z}_{n,k,\tau}(J) \\
&= \operatorname{tr}\left(\sum_{(n,J)} \bar{\pi}_{k,\tau}(n,J)\, \tilde{z}_{n,k,\tau}(J) \tilde{z}_{n,k,\tau}(J)^\top \bar{W}(\bar{\pi}_{k,\tau})^{-1} \right) \\
&= \operatorname{tr}\big(\bar{V}(\bar{\pi}_{k,\tau})\, \bar{W}(\bar{\pi}_{k,\tau})^{-1} \big) \\
&= \operatorname{tr}\big((\bar{W}(\bar{\pi}_{k,\tau}) - \alpha I_r)\, \bar{W}(\bar{\pi}_{k,\tau})^{-1} \big) \\
&= \operatorname{tr}(I_r) - \alpha \operatorname{tr}(\bar{W}(\bar{\pi}_{k,\tau})^{-1}) \;\le\; r.
\end{aligned}$$

Hence

$$\max_{n \in J \in \mathcal{J}(\mathcal{N}_{k,\tau})} \|\tilde{z}_{n,k,\tau}(J)\|^2_{(\bar{V}(\bar{\pi}_{k,\tau}) + \alpha I_r)^{-1}} = \bar{g}(\bar{\pi}_{k,\tau}) \le r.$$

Step 3 (support size bound). Let $S := \operatorname{supp}(\bar{\pi}_{k,\tau})$ and define $A_{(n,J)} := \tilde{z}_{n,k,\tau}(J) \tilde{z}_{n,k,\tau}(J)^\top \in \mathbb{S}^r$. If $|S| > \dim(\mathbb{S}^r) = r(r+1)/2$, then the matrices $\{A_{(n,J)} : (n,J) \in S\}$ are linearly dependent, so there exists a nonzero $v : S \to \mathbb{R}$ such that

$$\sum_{(n,J) \in S} v(n,J)\, A_{(n,J)} = 0.$$

Taking the trace after multiplying by $\bar{W}(\bar{\pi}_{k,\tau})^{-1}$ gives

$$\begin{aligned}
0 &= \operatorname{tr}\left(\bar{W}(\bar{\pi}_{k,\tau})^{-1} \sum_{(n,J) \in S} v(n,J) A_{(n,J)} \right) \\
&= \sum_{(n,J) \in S} v(n,J)\, \tilde{z}_{n,k,\tau}(J)^\top \bar{W}(\bar{\pi}_{k,\tau})^{-1} \tilde{z}_{n,k,\tau}(J) \\
&= \bar{g}(\bar{\pi}_{k,\tau}) \sum_{(n,J) \in S} v(n,J),
\end{aligned}$$

where the last equality uses (22). Thus $\sum_{(n,J) \in S} v(n,J) = 0$. Define $\bar{\pi}(t) = \bar{\pi}_{k,\tau} + tv$. Then $\sum_{(n,J) \in S} \bar{\pi}(t)(n,J) = 1$ for all t and

$$\bar{V}(\bar{\pi}(t)) = \bar{V}(\bar{\pi}_{k,\tau}) + t \sum_{(n,J) \in S} v(n,J) A_{(n,J)} = \bar{V}(\bar{\pi}_{k,\tau}),$$

so $\bar{W}(\bar{\pi}(t)) = \bar{W}(\bar{\pi}_{k,\tau})$ and hence $\bar{g}(\bar{\pi}(t)) = \bar{g}(\bar{\pi}_{k,\tau})$. Let

$$t^* := \sup\{ t > 0 :\ \bar{\pi}_{k,\tau}(n,J) + t\, v(n,J) \ge 0 \ \ \forall (n,J) \in S \}.$$

At $t = t^*$ at least one coordinate becomes 0, giving an equally optimal design with strictly smaller support. Iterating yields an optimal design with $|\operatorname{supp}(\bar{\pi}_{k,\tau})| \le r(r+1)/2$.

This completes the first claim.

(2) The bound for $\tilde{\pi}_{k,\tau}$. Now the design space is $\mathcal{J}(\mathcal{N}_{k,\tau})$. For $\tilde{\pi}$ on this set define

$$\tilde{V}(\tilde{\pi}) := \sum_{J \in \mathcal{J}(\mathcal{N}_{k,\tau})} \tilde{\pi}(J) \sum_{n \in J} p(n \mid J, \hat{\theta}_{k,\tau})\, \tilde{z}_{n,k,\tau}(J) \tilde{z}_{n,k,\tau}(J)^\top, \qquad \tilde{W}(\tilde{\pi}) := \tilde{V}(\tilde{\pi}) + \alpha I_r,$$

and

$$\tilde{g}(\tilde{\pi}) := \max_{J \in \mathcal{J}(\mathcal{N}_{k,\tau})} \sum_{n \in J} p(n \mid J, \hat{\theta}_{k,\tau})\, \tilde{z}_{n,k,\tau}(J)^\top \tilde{W}(\tilde{\pi})^{-1} \tilde{z}_{n,k,\tau}(J).$$

Let $\tilde{\pi}_{k,\tau}$ be a minimizer of $\tilde{g}(\cdot)$.

Step 1 (equalization on the support). As before, for every $J \in \operatorname{supp}(\tilde{\pi}_{k,\tau})$ we may assume

$$\sum_{n \in J} p(n \mid J, \hat{\theta}_{k,\tau})\, \tilde{z}_{n,k,\tau}(J)^\top \tilde{W}(\tilde{\pi}_{k,\tau})^{-1} \tilde{z}_{n,k,\tau}(J) = \tilde{g}(\tilde{\pi}_{k,\tau}). \tag{23}$$

Step 2 (trace bound). Using (23),

$$\begin{aligned}\tilde{g}(\tilde{\pi}_{k,\tau}) &= \sum_{J} \tilde{\pi}_{k,\tau}(J) \sum_{n \in J} p(n \mid J, \hat{\theta}_{k,\tau})\, \tilde{z}_{n,k,\tau}(J)^{\top} \tilde{W}(\tilde{\pi}_{k,\tau})^{-1} \tilde{z}_{n,k,\tau}(J) \\ &= \operatorname{tr}\Big(\tilde{V}(\tilde{\pi}_{k,\tau})\, \tilde{W}(\tilde{\pi}_{k,\tau})^{-1}\Big) \\ &= \operatorname{tr}\Big((\tilde{W}(\tilde{\pi}_{k,\tau}) - \alpha I_r)\, \tilde{W}(\tilde{\pi}_{k,\tau})^{-1}\Big) \\ &= \operatorname{tr}(I_r) - \alpha \operatorname{tr}(\tilde{W}(\tilde{\pi}_{k,\tau})^{-1}) \;\le\; r.\end{aligned}$$

Therefore,

$$\max_{J \in \mathcal{J}(\mathcal{N}_{k,\tau})} \sum_{n \in J} p(n \mid J, \hat{\theta}_{k,\tau})\, \|\tilde{z}_{n,k,\tau}(J)\|^2_{(\tilde{V}(\tilde{\pi}_{k,\tau}) + \alpha I_r)^{-1}} = \tilde{g}(\tilde{\pi}_{k,\tau}) \le r.$$

Step 3 (support size bound). Let $S := \operatorname{supp}(\tilde{\pi}_{k,\tau})$ and define the symmetric matrices

$$B_J := \sum_{n \in J} p(n \mid J, \hat{\theta}_{k,\tau})\, \tilde{z}_{n,k,\tau}(J) \tilde{z}_{n,k,\tau}(J)^{\top} \in \mathbb{S}^r.$$

If $|S| > r(r+1)/2$, then $\{B_J : J \in S\}$ are linearly dependent, so $\exists\, v : S \to \mathbb{R} \setminus \{0\}$ with $\sum_{J \in S} v(J) B_J = 0$. Multiplying by $\tilde{W}(\tilde{\pi}_{k,\tau})^{-1}$ and taking traces yields

$$0 = \sum_{J \in S} v(J) \operatorname{tr}(\tilde{W}(\tilde{\pi}_{k,\tau})^{-1} B_J) = \sum_{J \in S} v(J)\, \tilde{g}(\tilde{\pi}_{k,\tau}),$$

where we used (23). Hence $\sum_{J \in S} v(J) = 0$. Define $\tilde{\pi}(t) = \tilde{\pi}_{k,\tau} + tv$. Then $\sum_J \tilde{\pi}(t)(J) = 1$ and $\tilde{V}(\tilde{\pi}(t)) = \tilde{V}(\tilde{\pi}_{k,\tau})$, so $\tilde{g}(\tilde{\pi}(t)) = \tilde{g}(\tilde{\pi}_{k,\tau})$. Choosing t^* as before to hit the boundary removes at least one support element without changing optimality. Iterating yields $|\operatorname{supp}(\tilde{\pi}_{k,\tau})| \le r(r+1)/2$. This completes the second claim and the proof. □

From the above Lemmas A.14 and A.8, under E, we have

$$
\begin{aligned}
&\sum_{l\in[K]} R_l(S_l^*) - \sum_{l\in[K]} R_l(S_{l,\tau}^{(n,k)}) \\
&\le \sum_{l\in[K]} \Bigg[R_{l,\tau}^{LCB}(S_l^*) + 13\zeta_\tau^2 \max_{m\in S_l^*} \|z_m\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} + 4\zeta_\tau^2 \max_{m\in S_l^*} \|\tilde{z}_{m,l,\tau}\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} \\
&\qquad +2\zeta_\tau \sum_{m\in S_l^*} p(m|S_l^*, \widehat{\theta}_{l,\tau-1}) \|\tilde{z}_{m,l,\tau}\|_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} \Bigg] \\
&\quad - \sum_{l\in[K]} \Bigg[R_{l,\tau}^{UCB}(S_{l,\tau}^{(n,k)}) - 13\zeta_\tau^2 \max_{m\in S_{l,\tau}^{(n,k)}} \|z_m\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} - 4\zeta_\tau^2 \max_{m\in S_{l,\tau}^{(n,k)}} \|z_m\|^2_{H_{l,\tau-1}^{-1}(\widehat{\theta}_{l,\tau-1})} \\
&\qquad -2\zeta_\tau \sum_{m\in S_{l,\tau}^{(n,k)}} p(m|S_{l,\tau}^{(n,k)}, \widehat{\theta}_{l,\tau-1}) \|\tilde{z}_{m,l,\tau}\|_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} \Bigg] \\
&\lesssim \sum_{l\in[K]} \Bigg[\zeta_\tau^2 \max_{m\in S_l^*} \|z_m\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} + \zeta_\tau^2 \max_{m\in S_l^*} \|\tilde{z}_{m,l,\tau}\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} + \zeta_\tau \sum_{m\in S_l^*} p(m|S_l^*, \widehat{\theta}_{l,\tau-1}) \|\tilde{z}_{m,l,\tau}\|_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} \\
&\qquad +\zeta_\tau^2 \max_{m\in S_{l,\tau}^{(n,k)}} \|z_m\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} + \zeta_\tau^2 \max_{m\in S_{l,\tau}^{(n,k)}} \|\tilde{z}_{m,l,\tau}\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} \\
&\qquad\quad +\zeta_\tau \sum_{m\in S_{l,\tau}^{(n,k)}} p(m|S_{l,\tau}^{(n,k)}, \widehat{\theta}_{l,\tau-1}) \|\tilde{z}_{m,l,\tau}\|_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} \Bigg] \\
&\le \sum_{l\in[K]} \Bigg[\zeta_\tau^2 \max_{m\in S_l^*} \|z_m\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} + \zeta_\tau^2 \max_{m\in S_l^*} \|\tilde{z}_{m,l,\tau}\|_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} + \zeta_\tau^2 \max_{m\in S_{l,\tau}^{(n,k)}} \|z_m\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} \\
&\qquad +\zeta_\tau^2 \max_{m\in S_{l,\tau}^{(n,k)}} \|\tilde{z}_{m,l,\tau}\|_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} + \zeta_\tau \sqrt{\sum_{m\in S_l^*} p(m|S_l^*, \widehat{\theta}_{l,\tau-1})} \sqrt{\sum_{m\in S_l^*} p(m|S_l^*, \widehat{\theta}_{l,\tau-1}) \|\tilde{z}_{m,l,\tau}\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})}} \\
&\qquad +\zeta_\tau \sqrt{\sum_{m\in S_{l,\tau}^{(n,k)}} p(m|S_{l,\tau}^{(n,k)}, \widehat{\theta}_{l,\tau-1})} \sqrt{\sum_{m\in S_{l,\tau}^{(n,k)}} p(m|S_{l,\tau}^{(n,k)}, \widehat{\theta}_{l,\tau-1}) \|\tilde{z}_{m,l,\tau}\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})}} \Bigg],
\end{aligned}
\tag{24}
$$

where the second inequality comes from the fact that $(S_1^*, \ldots, S_K^*) \in \mathcal{M}_{\tau-1}$ and $\max_{(S_1,\ldots,S_K)\in\mathcal{M}_{\tau-1}} \sum_{l\in[K]} R_{l,\tau}^{LCB}(S_l) \le \sum_{l\in[K]} R_{l,\tau}^{UCB}(S_{l,\tau}^{(n,k)})$ from the algorithm.

Likewise, we also have

$$
\begin{aligned}
&\sum_{l\in[K]} R_l(S_l^*) - \sum_{l\in[K]} R_l(S_{l,\tau}^{(J,k)}) \\
&\lesssim \sum_{l\in[K]} \Bigg[\zeta_\tau^2 \max_{m\in S_l^*} \|z_m\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} + \zeta_\tau^2 \max_{m\in S_l^*} \|\tilde{z}_{m,l,\tau}\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} + \zeta_\tau^2 \max_{m\in S_{l,\tau}^{(J,k)}} \|z_m\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} \\
&\quad + \zeta_\tau^2 \max_{m\in S_{l,\tau}^{(J,k)}} \|\tilde{z}_{m,l,\tau}\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})} + \zeta_\tau \sqrt{\sum_{m\in S_l^*} p(m|S_l^*, \widehat{\theta}_{l,\tau-1})} \sqrt{\sum_{m\in S_l^*} p(m|S_l^*, \widehat{\theta}_{l,\tau-1}) \|\tilde{z}_{m,l,\tau}\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})}} \\
&\quad + \zeta_\tau \sqrt{\sum_{m\in S_{l,\tau}^{(n,k)}} p(m|S_{l,\tau}^{(J,k)}, \widehat{\theta}_{l,\tau-1})} \sqrt{\sum_{m\in S_{l,\tau}^{(J,k)}} p(m|S_{l,\tau}^{(J,k)}, \widehat{\theta}_{l,\tau-1}) \|\tilde{z}_{m,l,\tau}\|^2_{H_{l,\tau}^{-1}(\widehat{\theta}_{l,\tau})}} \Bigg].
\end{aligned} \tag{25}
$$

We can show that

$$
\begin{aligned}
&H_{k,\tau}(\widehat{\theta}_{k,\tau}) \\
&= \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} p(n|S_{k,t}, \widehat{\theta}_{k,\tau}) z_n z_n^\top - \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} \sum_{m\in S_{k,t}} p(n|S_{k,t}, \widehat{\theta}_{k,\tau}) p(m|S_{k,t}, \widehat{\theta}_{k,\tau}) z_n z_m^\top \\
&= \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} p(n|S_{k,t}, \widehat{\theta}_{k,\tau}) z_n z_n^\top - \frac{1}{2} \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} \sum_{m\in S_{k,t}} p(n|S_{k,t}, \widehat{\theta}_{k,\tau}) p(m|S_{k,t}, \widehat{\theta}_{k,\tau}) (z_n z_m^\top + z_n z_m^\top) \\
&\succeq \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} p(n|S_{k,t}, \widehat{\theta}_{k,\tau}) z_n z_n^\top - \frac{1}{2} \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} \sum_{m\in S_{k,t}} p(n|S_{k,t}, \widehat{\theta}_{k,\tau}) p(m|S_{k,t}, \widehat{\theta}_{k,\tau}) (z_n z_n^\top + z_m z_m^\top) \\
&= \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} p(n|S_{k,t}, \widehat{\theta}_{k,\tau}) z_n z_n^\top - \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} \sum_{m\in S_{k,t}} p(n|S_{k,t}, \widehat{\theta}_{k,\tau}) p(m|S_{k,t}, \widehat{\theta}_{k,\tau}) z_n z_n^\top \\
&= \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} p(n|S_{k,t}, \widehat{\theta}_{k,\tau}) \left(1 - \sum_{m\in S_{k,t}} p(m|S_{k,t}, \widehat{\theta}_{k,\tau})\right) z_n z_n^\top \\
&= \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} p(n|S_{k,s}, \widehat{\theta}_{k,s}) p(n_0|S_{k,s}, \widehat{\theta}_{k,}) z_n z_n^\top \succeq \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} \kappa z_n z_n^\top \\
&\succeq \lambda I_r + \sum_{n\in\mathcal{N}_{k,\tau-1}} \kappa r \pi_{k,\tau-1}(n) T_{\tau-1} z_n z_n^\top = \kappa T_{\tau-1} r (V(\pi_{k,\tau-1}) + (\lambda/\kappa r T_{\tau-1}) I_r) \\
&\succeq \kappa T_{\tau-1} r (V(\pi_{k,\tau-1}) + (\lambda/r T_{\tau-1}) I_r).
\end{aligned} \tag{26}
$$

From Lemma A.10 and (26), we also have, for any $n \in \mathcal{N}_{k,\tau}$

$$
\begin{aligned}
\|z_n\|^2_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})} &= \mathcal{O}\left(\frac{\|z_n\|^2_{(V(\pi_{k,\tau-1}) + (\lambda/rT_{\tau-1})I_r)^{-1}}}{\kappa r T_{\tau-1}} \right) \\
&= \mathcal{O}\left(\frac{1}{\kappa T_{\tau-1}} \right).
\end{aligned} \tag{27}
$$

We have

$$
\begin{aligned}
&H_{k,\tau}(\widehat{\theta}_{k,\tau}) \\
&= \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} p(n|S_{k,t},\widehat{\theta}_{k,\tau}) z_n z_n^\top - \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} \sum_{m\in S_{k,t}} p(n|S_{k,t},\widehat{\theta}_{k,\tau}) p(m|S_{k,t},\widehat{\theta}_{k,\tau}) z_n z_m^\top \\
&= \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \mathbb{E}_{\widehat{\theta}_{k,\tau}}[z_n z_n^\top] - \mathbb{E}_{\widehat{\theta}_{k,\tau}}[z_n]\mathbb{E}_{\widehat{\theta}_{k,\tau}}[z_n]^\top \\
&= \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \mathbb{E}_{\widehat{\theta}_{k,\tau}}[\tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^\top] \\
&= \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} p(n|S_{k,t},\widehat{\theta}_{k,\tau}) \tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^\top \\
&\succeq \lambda I_r + \sum_{J\in\mathcal{J}(\mathcal{N}_{k,\tau-1})} \sum_{t\in\mathcal{T}_{J,k,\tau-1}} \sum_{n\in J} p(n|J,\widehat{\theta}_{k,\tau}) \tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^\top \\
&\succeq \lambda I_r + \sum_{J\in\mathcal{J}(\mathcal{N}_{k,\tau-1})} r\bar{\pi}_{k,\tau-1}(J) T_{\tau-1} \sum_{n\in J} \kappa \tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^\top \\
&\succeq \kappa T_{\tau-1} r \left(\bar{V}(\bar{\pi}_{k,\tau-1}) + (\lambda/T_{\tau-1}r) I_r\right).
\end{aligned} \tag{28}
$$

From Lemma A.15 and (28) with $\mathcal{N}_{k,\tau} \subseteq \mathcal{N}_{k,\tau-1}$, we also have, for any $n \in J \in \mathcal{J}(\mathcal{N}_{k,\tau})$

$$
\begin{aligned}
\|\tilde{z}_{n,k,\tau}(J)\|^2_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})} &= \mathcal{O}\left(\frac{\|\tilde{z}_{n,k,\tau}(J))\|^2_{(\bar{V}(\bar{\pi}_{k,\tau-1})+(\lambda/rT_{\tau-1})I_r)^{-1}}}{\kappa r T_{\tau-1}}\right) \\
&= \mathcal{O}\left(\frac{1}{\kappa T_{\tau-1}}\right).
\end{aligned} \tag{29}
$$

We have

$$
\begin{aligned}
&H_{k,\tau}(\widehat{\theta}_{k,\tau}) \\
&= \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} p(n|S_{k,t},\widehat{\theta}_{k,\tau}) z_n z_n^\top - \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} \sum_{m\in S_{k,t}} p(n|S_{k,t},\widehat{\theta}_{k,\tau}) p(m|S_{k,t},\widehat{\theta}_{k,\tau}) z_n z_m^\top \\
&= \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \mathbb{E}_{\widehat{\theta}_{k,\tau}}[z_n z_n^\top] - \mathbb{E}_{\widehat{\theta}_{k,\tau}}[z_n]\mathbb{E}_{\widehat{\theta}_{k,\tau}}[z_n]^\top \\
&= \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \mathbb{E}_{\widehat{\theta}_{k,\tau}}[\tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^\top] \\
&= \lambda I_r + \sum_{t\in\mathcal{T}_{k,\tau-1}} \sum_{n\in S_{k,t}} p(n|S_{k,t},\widehat{\theta}_{k,\tau}) \tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^\top \\
&\succeq \lambda I_r + \sum_{J\in\mathcal{J}(\mathcal{N}_{k,\tau-1})} \sum_{t\in\mathcal{T}_{J,k,\tau-1}} \sum_{n\in J} p(n|J,\widehat{\theta}_{k,\tau}) \tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^\top \\
&\succeq \lambda I_r + \sum_{J\in\mathcal{J}(\mathcal{N}_{k,\tau-1})} r\tilde{\pi}_{k,\tau-1}(J) T_{\tau-1} \sum_{n\in J} p(n|J,\widehat{\theta}_{k,\tau}) \tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^\top \\
&\succeq \lambda I_r + \sum_{J\in\mathcal{J}(\mathcal{N}_{k,\tau-1})} r\tilde{\pi}_{k,\tau-1}(J) T_{\tau-1} \sum_{n\in J} p(n|J,\widehat{\theta}_{k,\tau-1}) \tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^\top \\
&\qquad - 2\zeta_\tau \sum_{J\in\mathcal{J}(\mathcal{N}_{k,\tau-1})} r\tilde{\pi}_{k,\tau-1}(J) T_{\tau-1} \max_{n\in J}(\|z_n\|_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})} + \|z_n\|_{H_{k,\tau-1}^{-1}(\widehat{\theta}_{k,\tau-1})}) \tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^\top \\
&= T_{\tau-1} r \Big(\tilde{V}(\tilde{\pi}_{k,\tau-1}) + (\lambda/T_{\tau-1}r) I_r \\
&\qquad -2\zeta_\tau \sum_{J\in\mathcal{J}(\mathcal{N}_{k,\tau-1})} \tilde{\pi}_{k,\tau-1}(J) \max_{n\in J}(\|z_n\|_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})} + \|z_n\|_{H_{k,\tau-1}^{-1}(\widehat{\theta}_{k,\tau-1})}) \tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^\top \Big),
\end{aligned} \tag{30}
$$

where the last inequality is obtained from, using the mean value theorem,

$$\begin{aligned}
&\sum_{n\in J}(p(n|J,\widehat{\theta}_{k,\tau})-p(n|J,\widehat{\theta}_{k,\tau-1})\tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^{\top}\\
&=\sum_{n\in J}(p(n|J,\widehat{\theta}_{k,\tau})-p(n|J,\theta_k^*)+p(n|J,\theta_k^*)-p(n|J,\widehat{\theta}_{k,\tau-1}))\tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^{\top}\\
&\succeq -2\zeta_\tau(\max_{n\in J}\|z_n\|_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})}+\max_{n\in J}\|z_n\|_{H_{k,\tau-1}^{-1}(\widehat{\theta}_{k,\tau-1})})\tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^{\top}. \qquad (31)
\end{aligned}$$

Let $B=2\zeta_\tau\sum_{J\in\mathcal{J}(\mathcal{N}_{k,\tau-1})}\tilde{\pi}_{k,\tau-1}(J)\max_{n\in J}(\|z_n\|_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})}+\|z_n\|_{H_{k,\tau-1}^{-1}(\widehat{\theta}_{k,\tau-1})})\tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^{\top}$ and we have $B\preceq 4\zeta_\tau\sqrt{\frac{1}{\kappa T_{\tau-2}}}\sum_{J\in\mathcal{J}(\mathcal{N}_{k,\tau-1})}\tilde{\pi}_{k,\tau-1}(J)\max_{n\in J}\tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^{\top}$ from (27). Then for $\tau\geq 3$, we have

$$\begin{aligned}
&\tilde{V}(\tilde{\pi}_{k,\tau-1})-B\\
&\succeq \frac{1}{2}\tilde{V}(\tilde{\pi}_{k,\tau-1})+\frac{1}{2}\tilde{V}(\tilde{\pi}_{k,\tau-1})-B\\
&\succeq \frac{1}{2}\tilde{V}(\tilde{\pi}_{k,\tau-1})+\frac{1}{2}\sum_{J\in\mathcal{J}(\mathcal{N}_{k,\tau})}\tilde{\pi}_{k,\tau}(J)\sum_{n\in J}\kappa\tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^{\top}-4\zeta_\tau\sqrt{\frac{1}{\kappa T_{\tau-2}}}\sum_{J\in\mathcal{J}(\mathcal{N}_{k,\tau-1})}\tilde{\pi}_{k,\tau-1}(J)\max_{n\in J}\tilde{z}_{n,k,\tau}\tilde{z}_{n,k,\tau}^{\top}\\
&\succeq \frac{1}{2}\tilde{V}(\tilde{\pi}_{k,\tau-1}), \qquad (32)
\end{aligned}$$

where the last inequality is obtained from $\frac{1}{2}\kappa\geq 4\zeta_\tau\sqrt{\frac{1}{\kappa T_{\tau-2}}}$ because $T_{\tau-2}\geq\min\{T_1,\eta_T\}$ with large enough T such that $T\geq\max\{\frac{r^3K}{\kappa^6}\log^4(KTL),\exp(\frac{r}{\kappa^3})\}$.
Then, we have

$$\begin{aligned}
\|\tilde{z}_{n,k,\tau}\|^2_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})}&\leq rT_{\tau-1}\|\tilde{z}_{n,k,\tau}\|^2_{(\tilde{V}(\tilde{\pi}_{k,\tau-1})+(\lambda/T_{\tau-1}r)I_r-B)^{-1}}\\
&\leq rT_{\tau-1}\|\tilde{z}_{n,k,\tau}\|^2_{(\frac{1}{2}\tilde{V}(\tilde{\pi}_{k,\tau-1})+\frac{1}{2}(\lambda/T_{\tau-1}r)I_r)^{-1}}\\
&\leq 2rT_{\tau-1}\|\tilde{z}_{n,k,\tau}\|^2_{(\tilde{V}(\tilde{\pi}_{k,\tau-1})+(\lambda/T_{\tau-1}r)I_r)^{-1}}.
\end{aligned}$$

Then from the above, Lemma A.15, and (30) with $\mathcal{N}_{k,\tau}\subseteq\mathcal{N}_{k,\tau-1}$, we have, for any $J\in\mathcal{J}(\mathcal{N}_{k,\tau})$

$$\begin{aligned}
&\sum_{n\in J}p(n|J,\widehat{\theta}_{k,\tau-1})\|\tilde{z}_{n,k,\tau}\|^2_{H_{k,\tau}^{-1}(\widehat{\theta}_{k,\tau})}\\
&=\mathcal{O}\left(\frac{\sum_{n\in J}p(n|J,\widehat{\theta}_{k,\tau-1})\|\tilde{z}_{n,k,\tau}\|^2_{(\tilde{V}(\tilde{\pi}_{k,\tau-1})+(\lambda/T_{\tau-1}r)I_r)^{-1}}}{rT_{\tau-1}}\right)\\
&=\mathcal{O}\left(\frac{1}{T_{\tau-1}}\right). \qquad (33)
\end{aligned}$$

Therefore under E, from (24), (25), (27),(29), and (33), we have the following.
For $t\in\bigcup_{n\in\mathcal{N}_{k,\tau},k\in[K]}\mathcal{T}_{n,k,\tau}\bigcup_{J\in\mathcal{J}(\mathcal{N}_{k,\tau}),k\in[K]}\mathcal{T}_{J,k,\tau}\bigcup_{n\in J\in\mathcal{J}(\mathcal{N}_{k,\tau}),k\in[K]}\mathcal{T}_{n,J,k,\tau}$,

$$\sum_{k\in[K]}(R_k(S_k^*)-R_k(S_{k,t}))=\mathcal{O}\left(K\left(\sqrt{\frac{r}{T_{\tau-1}}}+\frac{r}{T_{\tau-1}\kappa}\right)\right).$$

For the regret bound, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t\in[T]}\sum_{k\in[K]} R_k(S_k^*) - R_t(S_{k,t})\right] \\
&\le \mathbb{E}\left[\sum_{t\in[T]}\sum_{k\in[K]} \left(R_k(S_k^*) - R_k(S_{k,t})\right)\mathbf{1}(E)\right] + \mathbb{E}\left[\sum_{t\in[T]}\sum_{k\in[K]} \left(R_k(S_k^*) - R_k(S_{k,t})\right)\mathbf{1}(E^c)\right] \\
&= \tilde{\mathcal{O}}\left(K\sum_{\tau=3}^{\tau_T}\sum_{k\in[K]}\left(\sum_{J\in\mathcal{J}(\mathcal{N}_{k,\tau})}|\mathcal{T}_{J,k,\tau}| + \sum_{n\in\mathcal{N}_{k,\tau}}|\mathcal{T}_{n,k,\tau}|\right)\left(\sqrt{\frac{r}{T_{\tau-1}}} + \frac{r}{T_{\tau-1}\kappa}\right)\right) + \tilde{\mathcal{O}}(rK\eta_T) + \mathcal{O}(K) \\
&= \tilde{\mathcal{O}}\left(K\sum_{\tau=3}^{\tau_T}\sum_{k\in[K]}\left(\sum_{J\in\mathcal{J}(\mathcal{N}_{k,\tau})}|\mathcal{T}_{J,k,\tau}| + \sum_{n\in\mathcal{N}_{k,\tau}}|\mathcal{T}_{n,k,\tau}|\right)\left(\sqrt{\frac{r}{T_{\tau-1}}} + \frac{r}{T_{\tau-1}\kappa}\right)\right) + \tilde{\mathcal{O}}(rK\eta_T) \\
&= \tilde{\mathcal{O}}\left(K\sum_{\tau=3}^{\tau_T}\sum_{k\in[K]}\left(rT_\tau + |Supp(\pi_{k,\tau})| + |Supp(\tilde{\pi}_{k,\tau})|\right)\left(\sqrt{\frac{r}{T_{\tau-1}}} + \frac{r}{T_{\tau-1}\kappa}\right)\right) + \tilde{\mathcal{O}}(rK\eta_T) \\
&= \tilde{\mathcal{O}}\left(K^2\sum_{\tau=3}^{\tau_T}\left(r^{3/2}\eta_T + r^2\frac{1}{\kappa\sqrt{T_{\tau-1}}}\eta_T\right)\right) \\
&= \tilde{\mathcal{O}}\left(K^2 r^{3/2}\eta_T\right) \\
&= \tilde{\mathcal{O}}\left(r^{3/2}K^2(T/rK)^{\frac{1}{2(1-2^{-M})}}\right), \qquad (34)
\end{aligned}
$$

where the third last equality comes from Lemma A.10 and the second last equality comes from $\tau_T \le M = \tilde{\mathcal{O}}(1)$ and $T_{\tau-1} \ge \eta_T$ for $\tau \ge 3$.

A.8 Approximation Oracle

Here we discuss the combinatorial optimization in our algorithm. We can utilize an α-approximation oracle with $0 \le \alpha \le 1$, first introduced in Kakade et al. [2007]. Instead of obtaining the exact optimal assortment solution, the α-approximation oracle, denoted by $\mathbb{O}^\alpha$, outputs $\{S_k^\alpha\}_{k\in[K]}$ satisfying $\sum_{k\in[K]} f_k(S_k^\alpha) \ge \max_{\{S_k\}_{k\in[K]}\in\mathcal{M}} \sum_{k\in[K]} \alpha f_k(S_k)$.

We introduce an algorithm (Algorithm 5 in Appendix A.8) by modifying Algorithm 1 to incorporate α-approximation oracles for the optimization. Due to the redundancy, we explain only the distinct parts of the algorithm here. (Approximation oracles can also be applied to Algorithm 2 similarly, but we omit it in this discussion.) For testing the assignment (n,k), the algorithm constructs assortment $\{S_{l,\tau}^{\alpha,(n,k)}\}_{l\in[K]}$ (where $n \in S_{k,\tau}^{\alpha,(n,k)}$) in an optimistic view with an α-approximation oracle to resolve computation issue as follows. We define an approximation oracle $\mathbb{O}_{UCB}^{\alpha,(n,k)}$ which outputs $\{S_{l,\tau}^{\alpha,(n,k)}\}_{l\in[K]}$ satisfying

$$
\max_{\{S_l\}_{l\in[K]}\in\mathcal{M}_{\tau-1}:n\in S_k} \sum_{l\in[K]} \alpha R_{l,\tau}^{UCB}(S_l) \le \sum_{l\in[K]} R_{l,\tau}^{UCB}(S_{l,\tau}^{\alpha,(n,k)}), \qquad (35)
$$

which replaces Line 6 in Algorithm 1. For the elimination procedure, we define another β-approximation oracle, denoted by $\mathbb{O}_{LCB}^{\beta}$, which outputs $\{S_{l,\tau}^{\beta}\}_{l\in[K]}$ satisfying

$$
\max_{\{S_l\}_{l\in[K]}\in\mathcal{M}_{\tau-1}} \sum_{l\in[K]} \beta R_{l,\tau}^{LCB}(S_l) \le \sum_{l\in[K]} R_{l,\tau}^{LCB}(S_{l,\tau}^{\beta}). \qquad (36)
$$

Then it updates $\mathcal{N}_{k,\tau}$ by eliminating $n \in \mathcal{N}_{k,\tau-1}$ which satisfies the elimination condition of

$$
\sum_{l\in[K]} \alpha R_{l,\tau}^{LCB}(S_{l,\tau}^{\beta}) > \sum_{l\in[K]} R_{l,\tau}^{UCB}(S_{l,\tau}^{\alpha,(n,k)}), \qquad (37)
$$

which replaces Line 7 in Algorithm 1. We note that the algorithm utilizes the two different types of approximation oracles, $\mathbb{O}_{UCB}^{\alpha,(n,k)}$ and $\mathbb{O}_{LCB}^{\beta}$. Then the algorithm achieves a regret bound for γ-regret defined as $\mathcal{R}^{\gamma}(T) = \mathbb{E}[\sum_{t\in[T]} \sum_{k\in[K]} \gamma R_k(S_k^*) - R_k(S_{k,t})]$ in the following theorem.

Theorem A.16. *Algorithm 5 with $M = \mathcal{O}(\log(T))$ achieves a regret bound with $\gamma = \alpha\beta$ as*

$$\mathcal{R}^{\gamma}(T) = \tilde{\mathcal{O}}\left(\tfrac{1}{\kappa} K^{\frac{3}{2}} \sqrt{rT} \left(\frac{T}{rK}\right)^{\frac{1}{2(2^M-1)}}\right).$$

Proof. The proof is provided in Appendix A.8.2. □

A.8.1 α-Approximated Algorithm (Algorithm 5)

Algorithm 5: Batched Stochastic Matching Bandit with β-Approximation Oracle

Input: β, κ, $M \geq 1$; **Init:** $t \leftarrow 1, T_1 \leftarrow \eta_T$

Compute SVD of $X = U\Sigma V^\top$ and obtain $U_r = [u_1, \ldots, u_r]$; Construct $z_n \leftarrow U_r^\top x_n$ for $n \in [N]$

for $\tau = 1, 2...$ **do**

  for $k \in [K]$ **do**

    `// Estimation`

    $\widehat{\theta}_{k,\tau} \leftarrow \mathrm{argmin}_{\theta\in\mathbb{R}^r} l_{k,\tau}(\theta)$ with (2) where $\mathcal{T}_{k,\tau-1} = \mathcal{T}_{k,\tau-1}^{(1)} \cup \mathcal{T}_{k,\tau-1}^{(2)}$ and $\mathcal{T}_{k,\tau-1}^{(2)} = \bigcup_{n\in\mathcal{N}_{k,\tau-1}} \mathcal{T}_{n,k,\tau-1}^{(2)}$

    `// Assortments Construction`

    $\{S_{l,\tau}^{\alpha,(n,k)}\}_{l\in[K]} \leftarrow \mathbb{O}_{UCB}^{\alpha,(n,k)}$ from (35) for all $n \in \mathcal{N}_{k,\tau-1}$ with (3)

    `// Elimination`

    $\{S_{l,\tau}^{\beta}\}_{l\in[K]} \leftarrow \mathbb{O}_{LCB}^{\beta}$ from (36)

    $\mathcal{N}_{k,\tau} \leftarrow \{n \in \mathcal{N}_{k,\tau} : \sum_{l\in[K]} \alpha R_{l,\tau}^{LCB}(S_{l,\tau}^{\beta}) \leq \sum_{l\in[K]} R_{l,\tau}^{UCB}(S_{l,\tau}^{\alpha,(n,k)})\}$ for $k \in [K]$

    `// G/D-optimal design`

    $\pi_{k,\tau} \leftarrow \mathrm{argmax}_{\pi\in\mathcal{P}(\mathcal{N}_{k,\tau})} \log\det(\sum_{n\in\mathcal{N}_{k,\tau}} \pi_{k,\tau}(n) z_n z_n^\top + (1/rT_\tau) I_r)$

    `// Exploration`

    Run `Warm-up` (Algorithm 4) over time steps in $\mathcal{T}_{k,\tau}^{(1)}$ (defined in Algorithm 4)

    for $n \in \mathcal{N}_{k,\tau}$ **do**

      $t_{n,k} \leftarrow t$, $\mathcal{T}_{n,k,\tau}^{(2)} \leftarrow [t_{n,k}, t_{n,k} + \lceil r\pi_{k,\tau}(n) T_\tau \rceil - 1]$

      while $t \in \mathcal{T}_{n,k,\tau}^{(2)}$ **do**

        Offer $\{S_{l,t}\}_{l\in[K]} = \{S_{l,\tau}^{(n,k)}\}_{l\in[K]}$ and observe feedback $y_{m,k,t} \in \{0,1\}$ for all $m \in S_{l,t}$ and $l \in [K]$

        $t \leftarrow t + 1$

  $\mathcal{M}_\tau \leftarrow \{\{S_k\}_{k\in[K]} : S_k \subset \mathcal{N}_{k,\tau}, |S_k| \leq L\ \forall k \in [K], S_k \cap S_l = \emptyset\ \forall k \neq l\}$;

  $T_{\tau+1} \leftarrow \eta_T \sqrt{T_\tau}$

A.8.2 Proof of Theorem A.16

In this proof, we provide only the parts that are different from the proof of Theorem 5.2.

Lemma A.17. *Under E, $(S_1^*, \ldots, S_K^*) \in \mathcal{M}_{\tau-1}$ for all $\tau \in [T]$.*

Proof. Here we use induction for the proof. Suppose that for fixed τ, we have $(S_1^*, \ldots, S_K^*) \in \mathcal{M}_\tau$ for all $k \in [K]$. Recall that $\beta_T = (C_1/\kappa)\sqrt{\log(TKN)}$. From Lemma A.8, we have $R_{k,\tau+1}^{UCB}(S) \geq$

$R_k(S)$ and $R^{LCB}_{k,\tau+1}(S) \leq R_k(S)$ for any $S \subset [N]$. Then for $k \in [K]$, $n \in S^*_k$, and any $(S_1, .., S_K) \in \mathcal{M}_\tau$, we have

$$\begin{aligned}
\sum_{l\in[K]} R^{UCB}_{l,\tau+1}(S^{\alpha,(n,k)}_{l,\tau+1}) &\geq \max_{\{S_k\}_{k\in[K]}\in\mathcal{M}_\tau : n\in S_k} \sum_{l\in[K]} \alpha R^{UCB}_{l,\tau+1}(S_l) \\
&\geq \sum_{l\in[K]} \alpha R^{UCB}_{l,\tau+1}(S^*_l) \\
&\geq \sum_{l\in[K]} \alpha R_l(S^*_l) \\
&\geq \sum_{l\in[K]} \alpha R_l(S^{\beta}_{l,\tau+1}) \\
&\geq \sum_{l\in[K]} \alpha R^{LCB}_{l,\tau+1}(S^{\beta}_{l,\tau+1}),
\end{aligned} \tag{38}$$

where the first inequality comes from (35), the second inequality comes from $(S^*_1, \dots S^*_K) \in \mathcal{M}_\tau$, and the firth inequality comes from the optimality of $(S^*_1, \dots, S^*_K)$. This implies that $n \in \mathcal{N}_{k,\tau+1}$ from the algorithm. Then by following the same statement of (38) for all $n \in S^*_k$ and $k \in [K]$, we have $S^*_k \subset \mathcal{N}_{k,\tau+1}$ for all $k \in [K]$, which implies $(S^*_1, \dots, S^*_K) \in \mathcal{M}_{\tau+1}$. Therefore, with $(S^*_1, \dots, S^*_K) \in \mathcal{M}_1$, we can conclude the proof from the induction. □

From Lemmas A.17 and A.8, under E, we have

$$\begin{aligned}
\sum_{l\in[K]} \alpha\beta R_l(S^*_l) - \sum_{l\in[K]} R_l(S^{\alpha,(n,k)}_{l,\tau}) &\leq \sum_{l\in[K]} \alpha\beta R^{LCB}_{l,\tau}(S^*_l) + 4\beta_T \max_{m\in S^*_l} \|z_m\|_{V^{-1}_{l,\tau}} \\
&\quad - \sum_{l\in[K]} R^{UCB}_{l,\tau}(S^{\alpha,(n,k)}_{l,\tau}) + 4\beta_T \max_{m\in S^{(n,k)}_{l,\tau}} \|z_m\|_{V^{-1}_{l,\tau}} \\
&\leq \sum_{l\in[K]} \alpha R^{LCB}_{l,\tau}(S^{\beta}_{l,\tau}) + 4\beta_T \max_{m\in S^*_l} \|z_m\|_{V^{-1}_{l,\tau}} \\
&\quad - \sum_{l\in[K]} R^{UCB}_{l,\tau}(S^{\alpha,(n,k)}_{l,\tau}) + 4\beta_T \max_{m\in S^{(n,k)}_{l,\tau}} \|z_m\|_{V^{-1}_{l,\tau}} \\
&\leq 4\beta_T \sum_{l\in[K]} (\max_{m\in S^*_l} \|z_m\|_{V^{-1}_{l,\tau}} + \max_{m\in S^{(n,k)}_{l,\tau}} \|z_m\|_{V^{-1}_{l,\tau-1}}),
\end{aligned} \tag{39}$$

where the second inequality comes from (36) and last inequality comes from the fact that $(S^*_1, \dots, S^*_K) \in \mathcal{M}_{\tau-1}$ and $\sum_{l\in[K]} \alpha R^{LCB}_{l,\tau}(S^{\beta}_{l,\tau}) \leq \sum_{l\in[K]} R^{UCB}_{l,\tau}(S^{\alpha,(n,k)}_{l,\tau})$ from the algorithm. Then, by following the proof in Theorem 1, we can conclude the proof.

A.9 Proof of Lemmas

A.9.1 Proof of Lemma A.3

For the poof, we follow the proof steps in (Bounding the Prediction Error) Oh and Iyengar [2021]. We define

$$H_{k,\tau}(\theta) = \sum_{t\in\mathcal{T}_{k,\tau}} \left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\theta) z_n z_n^\top - \sum_{n\in S_{k,t}} \sum_{m\in S_{k,t}} p(n|S_{k,t},\theta) p(m|S_{k,t},\theta) z_n z_m^\top \right) + I_r.$$

We note that $g_{k,\tau}(\theta_1)-g_{k,\tau}(\theta_2)=\sum_{t\in\mathcal{T}_{k,\tau}}\sum_{n\in S_{k,t}}(p(n,|S_{k,t},\theta_1)-p(n,|S_{k,t},\theta_2))z_n+(\theta_1-\theta_2)$. Then from the mean value theorem, there exists $\bar{\theta}=c\theta_1+(1-c)\theta_2$ with some $c\in(0,1)$ such that

$$\begin{aligned}
&g_{k,\tau}(\theta_1)-g_{k,\tau}(\theta_2)\\
&=\nabla_\theta g_{k,\tau}(\theta)\big|_{\theta=\bar{\theta}}(\theta_1-\theta_2)\\
&=\left(\sum_{t\in\mathcal{T}_{k,\tau}}\left(\sum_{n\in S_{k,t}}p(n|S_{k,t},\bar{\theta})z_nz_n^\top-\sum_{n\in S_{k,t}}\sum_{m\in S_{k,t}}p(n|S_{k,t},\bar{\theta})p(m|S_{k,t},\bar{\theta})z_nz_m^\top\right)+I_r\right)(\theta_1-\theta_2)\\
&=H_{k,\tau}(\bar{\theta})(\theta_1-\theta_2) \qquad (40)
\end{aligned}$$

We define $L_{k,\tau}=H_{k,\tau}(\theta_k^*)$ and $E_{k,\tau}=H_{k,\tau}(\bar{\theta}_k)-H_{k,\tau}(\theta_k^*)$ where $\bar{\theta}_k=c\theta_k^*+(1-c)\widehat{\theta}_{k,\tau}$ for some constant $c\in(0,1)$.

From (40), we have $g_{k,\tau}(\widehat{\theta}_{k,\tau})-g_{k,\tau}(\theta_k^*)=(L_{k,\tau}+E_{k,\tau})(\widehat{\theta}_{k,\tau}-\theta_k^*)$. Then, for any $z\in\mathbb{R}^r$, we have

$$\begin{aligned}
z^\top(\widehat{\theta}_{k,\tau}-\theta_k^*)&=z^\top(L_{k,\tau}+E_{k,\tau})^{-1}(g_{k,\tau}(\widehat{\theta}_{k,\tau})-g_{k,\tau}(\theta_k^*))\\
&=z^\top L_{k,\tau}^{-1}(g_{k,\tau}(\widehat{\theta}_{k,\tau})-g_{k,\tau}(\theta_k^*))-z^\top L_{k,\tau}^{-1}E_{k,\tau}(L_{k,\tau}+E_{k,\tau})^{-1}(g_{k,\tau}(\widehat{\theta}_{k,\tau})-g_{k,\tau}(\theta_k^*)).
\end{aligned} \qquad (41)$$

For obtaining a bound for $|z^\top(\widehat{\theta}_{k,\tau}-\theta_k^*)|$, we analyze the two terms in (41). We first provide a bound for $|z^\top L_{k,\tau}^{-1}(g_{k,\tau}(\widehat{\theta}_{k,\tau})-g_{k,\tau}(\theta_k^*))|$. Let $\epsilon_{n,t}=y_{n,k,t}-p(n|S_{k,t},\theta_k^*)$ for $n\in S_{k,t}$. Since $\widehat{\theta}_{k,\tau}$ is the solution from MLE such that $\sum_{t\in\mathcal{T}_{k,\tau}}\sum_{n\in S_{k,t}}(p(n|S_{k,t},\widehat{\theta}_{k,\tau})-y_{n,k,\tau})z_n=0$, we have

$$\begin{aligned}
&g_{k,\tau}(\widehat{\theta}_{k,\tau})-g_{k,\tau}(\theta_k^*)\\
&=\sum_{t\in\mathcal{T}_{k,\tau}}\sum_{n\in S_{k,t}}\left(p(n|S_{k,t},\widehat{\theta}_{k,\tau})-p(n|S_{k,t},\theta_k^*)\right)z_n+(\widehat{\theta}_{k,\tau}-\theta_k^*)\\
&=\sum_{t\in\mathcal{T}_{k,\tau}}\sum_{n\in S_{k,t}}\left(p(n|S_{k,t},\widehat{\theta}_{k,\tau})-y_{n,k,t}\right)z_n+\widehat{\theta}_{k,\tau}+\sum_{t\in\mathcal{T}_{k,\tau}}\sum_{n\in S_{k,t}}\left(y_{n,k,\tau}-p(n|S_{k,t},\theta_k^*)\right)z_n-\theta_k^*\\
&=0+\sum_{t\in\mathcal{T}_{k,\tau}}\sum_{n\in S_{k,t}}\epsilon_{n,t}z_n-\theta_k^* \qquad (42)
\end{aligned}$$

We define

$$\begin{aligned}
Z_{k,t}&=[z_n:n\in S_{k,t}]^\top\in\mathbb{R}^{|S_{k,t}|\times r}\text{ for }t\in\mathcal{T}_{k,\tau},\\
D_{k,\tau}&=[Z_{k,t}:t\in\mathcal{T}_{k,\tau}]^\top\in\mathbb{R}^{(\sum_{t\in\mathcal{T}_{k,\tau}}|S_{k,t}|)\times r},\\
\mathcal{E}_{k,t}&=[\epsilon_{n,t}:n\in S_{k,t}]^\top\in\mathbb{R}^{|S_{k,t}|}.
\end{aligned}$$

Then using Hoeffding inequality, we have

$$\begin{aligned}
\mathbb{P}(|z^\top L_{k,\tau}^{-1}(g_{k,\tau}(\widehat{\theta}_{k,\tau})-g_{k,\tau}(\theta_k^*))|\geq\nu)&\leq\mathbb{P}\left(\left|\sum_{t\in\mathcal{T}_{k,\tau}}z^\top L_{k,\tau}^{-1}Z_{k,t}^\top\mathcal{E}_{k,t}\right|\geq\nu-|z^\top L_{k,\tau}^{-1}\theta_k^*|\right)\\
&\leq\mathbb{P}\left(\left|\sum_{t\in\mathcal{T}_{k,\tau}}z^\top L_{k,\tau}^{-1}Z_{k,t}^\top\mathcal{E}_{k,t}\right|\geq\nu-1\right)\\
&\leq 2\exp\left(-\frac{2(\nu-1)^2}{\sum_{t\in\mathcal{T}_{k,\tau}}(2\sqrt{2}\|z^\top L_{k,\tau}^{-1}Z_{k,t}^\top\|_2)^2}\right)\\
&=2\exp\left(-\frac{(\nu-1)^2}{4\|z^\top L_{k,\tau}^{-1}D_{k,\tau}^\top\|_2^2}\right)\\
&\leq 2\exp\left(-\frac{\kappa^2(\nu-1)^2}{4\|z\|_{V_{k,\tau}^{-1}}^2}\right),
\end{aligned} \qquad (43)$$

where the last inequality is obtained from the fact that

$$
\begin{aligned}
L_{k,\tau} &= \sum_{t\in\mathcal{T}_{k,\tau}} \left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\theta_k^*) z_n z_n^\top - \sum_{n\in S_{k,t}} \sum_{m\in S_{k,t}} p(n|S_{k,t},\theta_k^*) p(m|S_{k,t},\theta_k^*) z_n z_m^\top \right) \\
&= \sum_{t\in\mathcal{T}_{k,\tau}} \left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\theta_k^*) z_n z_n^\top - \frac{1}{2} \sum_{n\in S_{k,t}} \sum_{m\in S_{k,t}} p(n|S_{k,t},\theta_k^*) p(m|S_{k,t},\theta_k^*) (z_n z_m^\top + z_m z_n^\top) \right) \\
&\succeq \sum_{t\in\mathcal{T}_{k,\tau}} \left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\theta_k^*) z_n z_n^\top - \frac{1}{2} \sum_{n\in S_{k,t}} \sum_{m\in S_{k,t}} p(n|S_{k,t},\theta_k^*) p(m|S_{k,t},\theta_k^*) (z_n z_n^\top + z_m z_m^\top) \right) \\
&= \sum_{t\in\mathcal{T}_{k,\tau}} \left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\theta_k^*) z_n z_n^\top - \sum_{n\in S_{k,t}} \sum_{m\in S_{k,t}} p(n|S_{k,t},\theta_k^*) p(m|S_{k,t},\theta_k^*) z_n z_n^\top \right) \\
&= \sum_{t\in\mathcal{T}_{k,\tau}} \left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\theta_k^*) p(n_0|S_{k,t},\theta_k^*) z_n z_n^\top \right) \\
&\succeq \kappa D_\tau^\top D_\tau (= \kappa V_{k,\tau}),
\end{aligned}
$$

where the first inequality is obtained from $(z_n - z_m)(z_n - z_m)^\top = z_n z_n^\top + z_m z_m^\top - z_n z_m^\top - z_m z_n^\top \succeq 0$. Then from (43) using $\nu = (2/\kappa)\sqrt{\log(2TKN/\delta)}\|z\|_{V_{k,\tau}^{-1}} + 1$ and the union bound, with probability at least $1-\delta$, for all $\tau \in [T], k \in [K]$, we have

$$
|z^\top L_{k,\tau}^{-1}(g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau}(\theta_k^*))| \le \frac{3\sqrt{\log(TKN/\delta)}}{\kappa} \|z\|_{V_{k,\tau}^{-1}}. \tag{44}
$$

Now we provide a bound for the second term in (41) of $|z^\top L_{k,\tau}^{-1} E_{k,\tau}(L_{k,\tau} + E_{k,\tau})^{-1}(g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau}(\theta_k^*))|$. We have

$$
\begin{aligned}
&|z^\top L_{k,\tau}^{-1} E_{k,\tau}(L_{k,\tau} + E_{k,\tau})^{-1}(g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau}(\theta_k^*))| \\
&\le \|z\|_{L_{k,\tau}^{-1}} \|L_{k,\tau}^{-1/2} E_{k,\tau}(L_{k,\tau} + E_{k,\tau})^{-1} L^{1/2}\|_2 \|g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau}(\theta_k^*)\|_{L_{k,\tau}^{-1}} \\
&\le (1/\kappa)\|z\|_{V_{k,\tau}^{-1}} \|L_{k,\tau}^{-1/2} E_{k,\tau}(L_{k,\tau} + E_{k,\tau})^{-1} L^{1/2}\|_2 \|g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau}(\theta_k^*)\|_{V_{k,\tau}^{-1}}.
\end{aligned} \tag{45}
$$

Then it follows that

$$
\begin{aligned}
&\|L_{k,\tau}^{-1/2} E_{k,\tau}(L_{k,\tau} + E_{k,\tau})^{-1} L^{1/2}\|_2 \\
&= \|L_{k,\tau}^{-1/2} E_{k,\tau}(L_{k,\tau}^{-1} - L_{k,\tau}^{-1} E_{k,\tau}(L_{k,\tau} + E_{k,\tau})^{-1} L^{1/2}\|_2 \\
&\le \|L_{k,\tau}^{-1/2} E_{k,\tau} L_{k,\tau}^{-1/2}\|_2 + \|L_{k,\tau}^{-1/2} E_{k,\tau} L_{k,\tau}^{-1/2}\|_2 \|L_{k,\tau}^{-1/2} E_{k,\tau}(L_{k,\tau} + E_{k,\tau})^{-1} L_{k,\tau}^{1/2}\|_2,
\end{aligned}
$$

which implies

$$
\begin{aligned}
\|L_{k,\tau}^{-1/2} E_{k,\tau}(L_{k,\tau} + E_{k,\tau})^{-1} L_{k,\tau}^{1/2}\|_2 &\le \frac{\|L_{k,\tau}^{-1/2} E_{k,\tau} L_{k,\tau}^{-1/2}\|_2}{1 - \|L_{k,\tau}^{-1/2} E_{k,\tau} L_{k,\tau}^{-1/2}\|_2} \\
&\le 2\|L_{k,\tau}^{-1/2} E_{k,\tau} L_{k,\tau}^{-1/2}\|_2 \\
&\le \frac{6}{\kappa}\|\widehat{\theta}_{k,\tau} - \theta_k^*\|_2,
\end{aligned} \tag{46}
$$

where the last inequality is obtained from (17) and (18) in Oh and Iyengar [2021]. Then from (45), (46), we have

$$
\begin{aligned}
&|z^\top L_{k,\tau}^{-1} E_{k,\tau}(L_{k,\tau} + E_{k,\tau})^{-1}(g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau}(\theta_k^*))| \\
&\le \frac{6}{\kappa^2}\|\widehat{\theta}_{k,\tau} - \theta_k^*\|_2 \|g_{k,\tau}(\widehat{\theta}_{k,\tau}) - g_{k,\tau}(\theta_k^*)\|_{V_{k,\tau}^{-1}} \|z\|_{V_{k,\tau}^{-1}}.
\end{aligned} \tag{47}
$$

We can conclude the proof from (44) and (47).

A.9.2 Proof of Lemma A.7

We note that $g_{k,\tau}(\theta_1)-g_{k,\tau}(\theta_2) = \sum_{t\in\mathcal{T}_{k,\tau}}\sum_{n\in S_{k,t}}(p(n,|S_{k,t},\theta_1)-p(n,|S_{k,t},\theta_2))z_n+(\theta_1-\theta_2)$.
Define $H_{k,\tau}(\theta) = \sum_{t\in\mathcal{T}_{k,\tau}}\left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\theta)z_n z_n^\top - \sum_{n\in S_{k,t}}\sum_{m\in S_{k,t}} p(n|S_{k,t},\theta)p(m|S_{k,t},\theta)z_n z_m^\top\right)+ I_r$. Then we can show that there exists $\bar\theta = c\theta_1+(1-c)\theta_2$ with some $c\in(0,1)$ such that

$$
\begin{aligned}
&g_{k,\tau}(\theta_1) - g_{k,\tau}(\theta_2)\\
&= \nabla_\theta g_{k,\tau}(\theta)\big|_{\theta=\bar\theta}(\theta_1-\theta_2)\\
&= \left(\sum_{t\in\mathcal{T}_{k,\tau}}\left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\bar\theta)z_n z_n^\top - \sum_{n\in S_{k,t}}\sum_{m\in S_{k,t}} p(n|S_{k,t},\bar\theta)p(m|S_{k,t},\bar\theta)z_n z_m^\top\right) + I_r\right)(\theta_1-\theta_2)\\
&= H_{k,\tau}(\bar\theta)(\theta_1-\theta_2). \qquad (48)
\end{aligned}
$$

Define $\bar H_{k,\tau}(\bar\theta) = \sum_{t\in\mathcal{T}_{k,\tau}}\sum_{n\in S_{k,t}} p(n|S_{k,t},\bar\theta)p(n_0|S_{k,t},\bar\theta)z_n z_n^\top + I_r$. Then we have $H_{k,\tau}(\bar\theta)\succeq \bar H_{k,\tau}(\bar\theta)$ from the following.

$$
\begin{aligned}
&\sum_{t\in\mathcal{T}_{k,\tau}}\left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\bar\theta)z_n z_n^\top - \sum_{n\in S_{k,t}}\sum_{m\in S_{k,t}} p(n|S_{k,t},\bar\theta)p(m|S_{k,t},\bar\theta)z_n z_m^\top\right)\\
&= \sum_{t\in\mathcal{T}_{k,\tau}}\left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\bar\theta)z_n z_n^\top - \sum_{n\in S_{k,t}}\sum_{m\in S_{k,t}} p(n|S_{k,t},\bar\theta)p(m|S_{k,t},\bar\theta)z_n z_m^\top\right)\\
&= \sum_{t\in\mathcal{T}_{k,\tau}}\left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\bar\theta)z_n z_n^\top - \frac{1}{2}\sum_{n\in S_{k,t}}\sum_{m\in S_{k,t}} p(n|S_{k,t},\bar\theta)p(m|S_{k,t},\bar\theta)(z_n z_m^\top + z_m z_n^\top)\right)\\
&\succeq \sum_{t\in\mathcal{T}_{k,\tau}}\left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\bar\theta)z_n z_n^\top - \frac{1}{2}\sum_{n\in S_{k,t}}\sum_{m\in S_{k,t}} p(n|S_{k,t},\bar\theta)p(m|S_{k,t},\bar\theta)(z_n z_n^\top + z_m z_m^\top)\right)\\
&= \sum_{t\in\mathcal{T}_{k,\tau}}\left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\bar\theta)z_n z_n^\top - \sum_{n\in S_{k,t}}\sum_{m\in S_{k,t}} p(n|S_{k,t},\bar\theta)p(m|S_{k,t},\bar\theta)z_n z_n^\top\right)\\
&= \sum_{t\in\mathcal{T}_{k,\tau}}\left(\sum_{n\in S_{k,t}} p(n|S_{k,t},\bar\theta)p(n_0|S_{k,t},\bar\theta)z_n z_n^\top\right), \qquad (49)
\end{aligned}
$$

where the inequality is obtained from $(z_n - z_m)(z_n - z_m)^\top \succeq 0$. Under E_1, we have $\|\widehat\theta_{k,\tau}\|_2 - \|\theta_k^*\|_2 \le 1$ implying $\|\widehat\theta_{k,\tau}\|_2 \le 1 + \|\theta_k^*\|_2 = 1 + \|U_r^\top\theta_k\|_2 \le 2$. Then for $\bar\theta = c\widehat\theta_{k,\tau} + (1-c)\theta_k^*$ for some $c\in(0,1)$, we have $\|U_r\bar\theta\|_2\le 2$. Then from $p(n|S_{k,t},\bar\theta) = \exp(z_n^\top\bar\theta)/(1+\sum_{m\in S_{k,t}}\exp(z_m^\top\bar\theta)) = \exp(x_n^\top(U_r\bar\theta))/(1+\sum_{m\in S_{k,t}}\exp(x_m^\top(U_r\bar\theta)))$, we can show that $\bar H_{k,\tau}(\bar\theta)\succeq\kappa V_{k,\tau}$, which implies $H_{k,\tau}(\bar\theta)\succeq\bar H_{k,\tau}(\bar\theta)\succeq\kappa V_{k,\tau}$.

Then we have

$$
\begin{aligned}
\|\widehat\theta_{k,\tau}-\theta_k^*\|_2^2 &\le (1/\lambda_{\min}(V_{k,\tau}))(\widehat\theta_{k,\tau}-\theta_k^*)^\top V_{k,\tau}(\widehat\theta_{k,\tau}-\theta_k^*)\\
&\le (1/\kappa\lambda_{\min}(V_{k,\tau}^0))(\widehat\theta_{k,\tau}-\theta_k^*)^\top H_{k,\tau}(\bar\theta)(\widehat\theta_{k,\tau}-\theta_k^*)\\
&\le (1/\kappa\lambda_{\min}(V_{k,\tau}^0))(\widehat\theta_{k,\tau}-\theta_k^*)^\top H_{k,\tau}(\bar\theta)H_{k,\tau}(\bar\theta)^{-1}H_{k,\tau}(\bar\theta)(\widehat\theta_{k,\tau}-\theta_k^*)\\
&\le (1/\kappa^2\lambda_{\min}(V_{k,\tau}^0))(g_{k,\tau}(\widehat\theta_{k,\tau})-g_{k,\tau}(\theta_k^*))^\top V_{k,\tau}^{-1}(g_{k,\tau}(\widehat\theta_{k,\tau})-g_{k,\tau}(\theta_k^*))\\
&\le (1/\kappa^2\lambda_{\min}(V_{k,\tau}^0))\|g_{k,\tau}(\widehat\theta_{k,\tau})-g_{k,\tau}(\theta_k^*))\|_{V_{k,\tau}^{-1}}^2. \qquad (50)
\end{aligned}
$$

Then from E_2, we can conclude that

$$
\|\widehat\theta_{k,\tau}-\theta_k^*\|_2 \le \frac{4}{\kappa}\sqrt{\frac{2r+\log(KTN/\delta)}{\lambda_{\min}(V_{k,\tau}^0)}}.
$$

A.10 Proof of Proposition A.1

We first provide a lemma for a confidence bound. Let $\gamma_t(\delta) = c_1\sqrt{d}\log(L)\left(\log(t) + \sqrt{\log(t)\log(K/\delta)}\right)$ for some $c_1 > 0$.

Lemma A.18 (Lemma 1 in Lee and Oh [2024])**.** *With probability at least $1-\delta$, for all $t \geq 1$ and $k \in [K]$ we have*

$$\|\widehat{\theta}_{k,t} - \theta_k^*\|_{\mathcal{G}_{k,t}} \leq \gamma_t(\delta).$$

Let $\delta = 1/T$. From the above lemma, we define event $E = \{\|\widehat{\theta}_{k,t} - \theta_k^*\|_{\mathcal{G}_{k,t}} \leq \gamma_t \, \forall k \in [K] \text{ and } t \geq 1\}$, which holds with probability at least $1 - 1/T$. Then we provide a lemma for the optimism.

Lemma A.19. *Under E, for all $t \geq 1$, we have*

$$\sum_{k\in[K]} R_k(S_k^*) \leq \sum_{k\in[K]} R_{k,t}^{UCB}(S_{k,t}).$$

Proof. Under E, we have

$$|z_n^\top \widehat{\theta}_{k,t} - z_n^\top \theta_k^*| \leq \|z_n\|_{\mathcal{G}_{k,t}^{-1}} \|\widehat{\theta}_{k,t} - \theta_k^*\|_{\mathcal{G}_{k,t}} \leq \gamma_t \|z_n\|_{\mathcal{G}_{k,t}^{-1}},$$

which implies $z_n^\top \theta_k^* \leq z_n^\top \widehat{\theta}_{k,t} + \gamma_t \|z_n\|_{\mathcal{G}_{k,t}^{-1}} = h_{n,k,t}$. Therefore, from Lemma A.3 in Agrawal et al. [2017a], we have $R_k(S_k^*) \leq R_{k,t}^{UCB}(S_k^*)$. Then using definition of $S_{k,t}$ in the algorithm, we can conclude that

$$\sum_{k\in[K]} R_k(S_k^*) \leq \sum_{k\in[K]} R_{k,t}^{UCB}(S_k^*) \leq \sum_{k\in[K]} R_{k,t}^{UCB}(S_{k,t}).$$

□

Now we provide a lemma which is critical to bound regret under optimism.

Lemma A.20. *Under E, for all $k \in [K]$, we have*

$$\sum_{t=1}^{T} R_{k,t}^{UCB}(S_{k,t}) - R_k(S_{k,t}) = O\left(r\sqrt{T} + \frac{1}{\kappa}r^2\right)$$

Proof. By following the proof steps in Theorem 4 in Lee and Oh [2024], we can show this lemma. □

Then from Lemmas A.18 and A.20, we can conclude the proof for the regret as follows.

$$\begin{aligned}
\mathcal{R}(T) &= \mathbb{E}\left[\sum_{t\in[T]}\sum_{k\in[K]} R_k(S_{k,t}^*) - R_k(S_{k,t})\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in[K]} \left(R_k(S_{k,t}^*) - R_k(S_{k,t})\right)\mathbf{1}(E)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in[K]} \left(R_k(S_{k,t}^*) - R_k(S_{k,t})\right)\mathbf{1}(E^c)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k\in[K]} \left(R_{k,t}^{UCB}(S_{k,t}) - R_k(S_{k,t})\right)\mathbf{1}(E)\right] + \sum_{t=1}^{T}\sum_{k\in[K]} \mathbb{P}(E^c) \\
&= \tilde{\mathcal{O}}\left(rK\sqrt{T} + \frac{1}{\kappa}r^2K\right) = \tilde{\mathcal{O}}\left(rK\sqrt{T}\right).
\end{aligned}$$

Now we discuss the computational cost. Since there exists $\mathcal{O}(K^N)$ number of assortment candidate in $\mathcal{M}$, especially for $L \geq N$, the cost per round is $\mathcal{O}(K^N)$ from Line 7.

A.11 Proof of Lemma A.10

Fix an epoch τ and an arm k. Recall that the projected features satisfy $z_n \in \mathbb{R}^r$ and define, for any $\pi \in \mathcal{P}(\mathcal{N}_{k,\tau})$,

$$V(\pi) := \sum_{n \in \mathcal{N}_{k,\tau}} \pi(n)\, z_n z_n^\top, \qquad W(\pi) := V(\pi) + \alpha I_r, \quad \alpha := \frac{1}{rT_\tau},$$

and

$$g(\pi) := \max_{n \in \mathcal{N}_{k,\tau}} z_n^\top W(\pi)^{-1} z_n = \max_{n \in \mathcal{N}_{k,\tau}} \|z_n\|^2_{(V(\pi)+\alpha I_r)^{-1}}.$$

Let $\pi_{k,\tau}$ be a (regularized) G-optimal design, i.e., $\pi_{k,\tau} \in \arg\min_{\pi \in \mathcal{P}(\mathcal{N}_{k,\tau})} g(\pi)$, and denote $W_* := W(\pi_{k,\tau})$ and $g_* := g(\pi_{k,\tau})$.

Step 1: Active constraints on the support. Let $S := \mathrm{supp}(\pi_{k,\tau})$. We claim that for every $n \in S$,

$$z_n^\top W_*^{-1} z_n = g_*. \tag{51}$$

This is a standard optimality (equivalence/KKT) condition. Indeed, consider the equivalent regularized D-optimal problem

$$\max_{\pi \in \mathcal{P}(\mathcal{N}_{k,\tau})} \log\det(W(\pi)).$$

Since $\log\det(\cdot)$ is concave and $W(\pi)$ is affine in π, KKT conditions are necessary and sufficient. The gradient is

$$\frac{\partial}{\partial \pi(n)} \log\det(W(\pi)) = \mathrm{trace}\big(W(\pi)^{-1} z_n z_n^\top\big) = z_n^\top W(\pi)^{-1} z_n.$$

Thus there exists $\lambda \in \mathbb{R}$ such that

$$z_n^\top W_*^{-1} z_n \le \lambda \quad \forall n \in \mathcal{N}_{k,\tau}, \qquad \text{and} \qquad z_n^\top W_*^{-1} z_n = \lambda \quad \forall n \in S.$$

By definition, $\lambda = \max_n z_n^\top W_*^{-1} z_n = g_*$, proving (51).

Step 2: Bound $g(\pi_{k,\tau}) \le r$. Using (51) and $\sum_{n \in S} \pi_{k,\tau}(n) = 1$, we obtain

$$\sum_{n \in \mathcal{N}_{k,\tau}} \pi_{k,\tau}(n)\, z_n^\top W_*^{-1} z_n = \sum_{n \in S} \pi_{k,\tau}(n)\, g_* = g_*.$$

On the other hand,

$$\begin{aligned}
\sum_{n \in \mathcal{N}_{k,\tau}} \pi_{k,\tau}(n)\, z_n^\top W_*^{-1} z_n &= \mathrm{trace}\left(\sum_{n \in \mathcal{N}_{k,\tau}} \pi_{k,\tau}(n)\, z_n z_n^\top\, W_*^{-1} \right) \\
&= \mathrm{trace}\big(V(\pi_{k,\tau})\, W_*^{-1}\big) \\
&= \mathrm{trace}\big((W_* - \alpha I_r) W_*^{-1}\big) = \mathrm{trace}(I_r) - \alpha\, \mathrm{trace}(W_*^{-1}) \\
&= r - \alpha\, \mathrm{trace}(W_*^{-1}) \le r.
\end{aligned}$$

Therefore,

$$g(\pi_{k,\tau}) = g_* \le r.$$

Step 3: Support-size bound $|S| \le r(r+1)/2$. Each matrix $z_n z_n^\top$ is symmetric $r \times r$, hence lies in the vector space $\mathbb{S}^r$ of symmetric matrices with dimension $\dim(\mathbb{S}^r) = r(r+1)/2$. If $|S| > r(r+1)/2$, then the set $\{z_n z_n^\top : n \in S\}$ is linearly dependent, so there exists a nonzero $v : S \to \mathbb{R}$ such that

$$\sum_{n \in S} v(n)\, z_n z_n^\top = 0.$$

Multiplying by W_*^{-1} and taking trace yields

$$0 = \mathrm{trace}\left(W_*^{-1} \sum_{n \in S} v(n)\, z_n z_n^\top \right) = \sum_{n \in S} v(n)\, \mathrm{trace}\big(W_*^{-1} z_n z_n^\top\big) = \sum_{n \in S} v(n)\, z_n^\top W_*^{-1} z_n.$$

Using (51), $z_n^\top W_*^{-1} z_n = g_*$ for all $n \in S$, hence

$$0 = g_* \sum_{n \in S} v(n).$$

Since $W_* \succ 0$ and $g_* = \max_n z_n^\top W_*^{-1} z_n \geq 0$ (and is positive whenever some $z_n \neq 0$), we obtain

$$\sum_{n \in S} v(n) = 0.$$

Define, for $t \in \mathbb{R}$, a perturbed design $\pi(t)$ by

$$\pi(t)(n) = \begin{cases} \pi_{k,\tau}(n) + t\, v(n), & n \in S, \\ 0, & n \notin S. \end{cases}$$

Because $\sum_{n \in S} v(n) = 0$, we have $\sum_n \pi(t)(n) = 1$ for all t. Moreover,

$$V(\pi(t)) = \sum_{n \in S} (\pi_{k,\tau}(n) + tv(n)) z_n z_n^\top = V(\pi_{k,\tau}) + t \sum_{n \in S} v(n) z_n z_n^\top = V(\pi_{k,\tau}),$$

so $W(\pi(t)) = W(\pi_{k,\tau}) = W_*$ for all t such that $\pi(t) \geq 0$. Hence $g(\pi(t)) = \max_n z_n^\top W(\pi(t))^{-1} z_n = \max_n z_n^\top W_*^{-1} z_n = g_*$.

Let

$$t' = \sup\Big\{ t > 0 : \ \pi_{k,\tau}(n) + tv(n) \geq 0 \ \ \forall n \in S \Big\}.$$

Then $t' > 0$ and at $t = t'$ at least one coordinate in S becomes zero; otherwise we could increase t further and stay nonnegative, contradicting the definition of t'. Thus $\pi(t')$ is feasible, achieves the same value g_*, and satisfies $|\mathrm{supp}(\pi(t'))| \leq |S| - 1$.

Iterating this support-reduction argument yields an optimal design π such that

$$|\mathrm{supp}(\pi)| \leq \frac{r(r+1)}{2}.$$

This completes the proof.

A.12 Experiments

We compare B-SMB and B-SMB$^+$ with three per-round baselines: the adapted MNL bandit algorithm OFU-MNL$^+$ [Lee and Oh, 2024], and the stable-MNL matching bandit algorithms UCB-QMB and TS-QMB [Kim and Oh, 2024]. Details on the adaptation of OFU-MNL$^+$ are provided in Appendix A.2. We sample $x_n, \theta_k \sim \mathrm{Unif}([-1,1]^d)$ and normalize them, and draw $w_{n,k} \sim \mathrm{Unif}[0,1]$. We use $N = 3$, $K = 2$, $r = 2$, and $T = 5000$ in Figure 2, and $N = 7$, $K = 4$ in Figure 3.

Figures 2 and 3 report the cumulative number of optimization updates, runtime, and cumulative regret. As discussed in Section 4, the per-round baselines repeatedly solve a combinatorial optimization problem, so their optimization updates grow linearly with time. In contrast, our batched algorithms require only a few updates, consistent with Propositions 5.1 and 6.2, and therefore achieve substantial runtime savings, while maintaining regret comparable to OFU-MNL$^+$. Specifically, the runtime is only about 25% of the benchmark runtime for $N = 3, K = 2$, and only about 2% for $N = 7, K = 4$. Equivalently, this corresponds to roughly 75% and 98% reductions in runtime, respectively, even in these relatively small instances. This efficiency gap becomes more pronounced as the problem size increases. Additional results are reported in Appendix A.12.1.

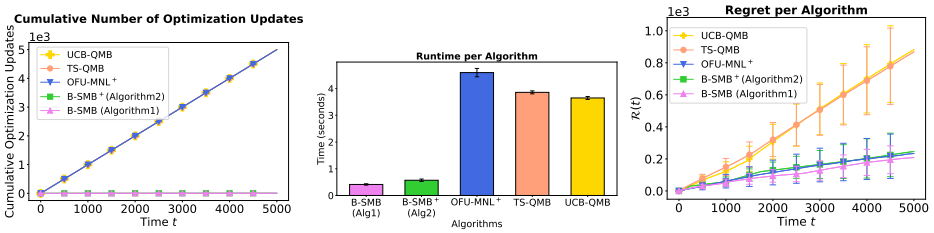


Figure 2: Results for $N = 3$, $K = 2$: optimization updates, runtime, and cumulative regret.

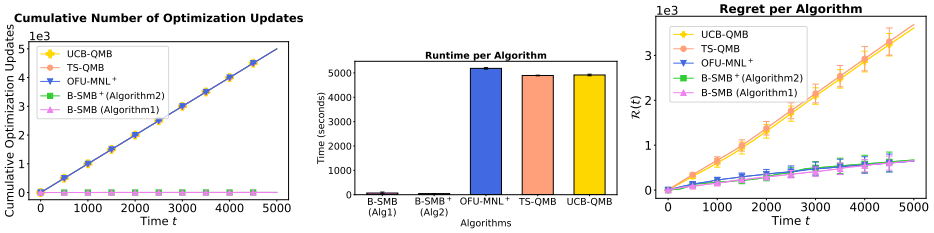


Figure 3: Results for $N = 7$, $K = 4$: cumulative optimization updates, runtime, cumulative regret.

A.12.1 Additional Scalability Experiment

We also conducted a larger experiment with $N = 18$ and $K = 5$. This experiment is intended as a computational stress test rather than a full regret comparison, since the per-round exact baseline did not complete the $5{,}000$-round run within 24 hours on the same setup. In this instance, `B-SMB` and `B-SMB`$^+$ completed all $5{,}000$ rounds in 1419 and 3961 seconds, respectively, while requiring only 4 and 3 optimization updates. By contrast, within the same 3961-second time budget, the exact baseline `OFU-MNL`$^+$ completed only 44 rounds, which is just 0.88% of the rounds completed by our methods.

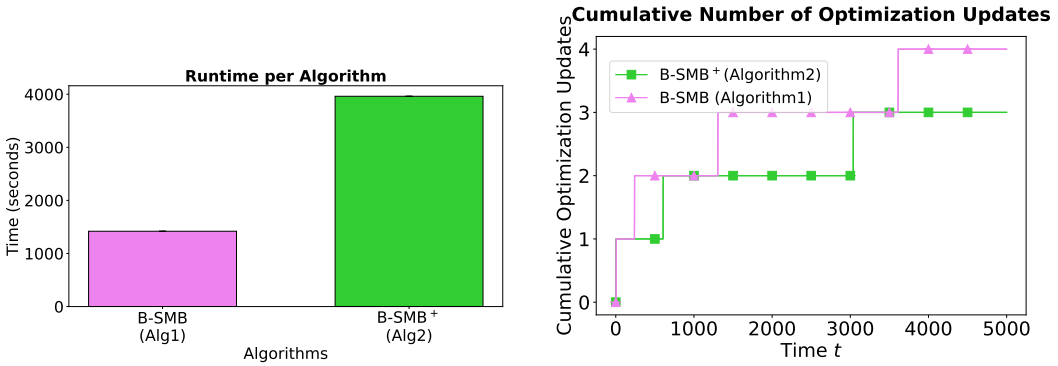


Figure 4: Additional scalability experiment with $N = 18$, $K = 5$. Left: runtime comparison. Right: number of optimization updates.

A.13 Auxiliary Lemmas

Lemma A.21 (Lemma E.2 in Lee and Oh [2024])**.** *For all $t \geq 1$ and $k \in [K]$, we have*

$$(i) \quad \sum_{s=1}^{t} \sum_{n \in S_{k,s}} p(n|S_{k,s}, \widehat{\theta}_{k,s}) p(n_0|S_{k,s}, \widehat{\theta}_{k,s}) \|z_n\|^2_{H_{k,s}^{-1}} \leq 2r \log\left(1 + \tfrac{t}{r\lambda}\right),$$

$$(ii) \quad \sum_{s=1}^{t} \max_{n \in S_{k,s}} \|z_n\|^2_{H_{k,s}^{-1}} \leq \tfrac{1}{\kappa} 2r \log\left(1 + \tfrac{t}{r\lambda}\right).$$

Lemma A.22 (Lemma E.3 in Lee and Oh [2024])**.** *Define $\tilde{Q} : \mathbb{R}^{|S|} \to \mathbb{R}$ for $S \in [N]$, such that for any $\boldsymbol{u} = (u_1, \ldots, u_{|S|}) \in \mathbb{R}^{|S|}$, $\tilde{Q}(\boldsymbol{u}) = \sum_{n \in S} \frac{\exp(u_n)}{1+\sum_{m \in S} \exp(u_m)}$. Let $p_n(\boldsymbol{u}) = \frac{\exp(u_n)}{1+\sum_{m \in S} \exp(u_m)}$. Then for all $n \in S$, we have*

$$\left| \frac{\partial^2 \tilde{Q}}{\partial u_n \partial u_m} \right| \leq \begin{cases} 3p_n(\boldsymbol{u}), & \text{if } n = m \\ 2p_n(\boldsymbol{u}) p_m(\boldsymbol{u}), & \text{if } n \neq m \end{cases}$$

A.14 Broader Impacts

The work may positively impact online matching and recommendation systems by reducing computational burden. Since the paper is primarily theoretical and uses only synthetic experiments, we do not identify direct negative societal impacts.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the main contributions: the stochastic matching bandit framework, rare optimization updates, and $\widetilde{\mathcal{O}}(\sqrt{T})$-type regret guarantees.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper explicitly discusses the remaining $\sqrt{K}$ gap between the regret upper bound and the known lower bound, and identifies closing this gap—or proving it inherent under rare optimization updates—as an important future direction.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: The assumptions are stated in the problem formulation and theorem statements. Full proofs of the main regret and batch-update guarantees, as well as supporting lemmas, are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: The experiments use synthetic data, and the paper specifies the data-generation process, problem dimensions, horizon, baselines, and algorithmic procedures. Baseline adaptation details and the additional scalability experiment are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is provided in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Appendix A.12 specifies the synthetic data distribution, values of N, K, r, T, and the compared baselines.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: The figures report variability across independent synthetic runs using error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The work is primarily theoretical and the experiments are simple synthetic simulations that do not require GPUs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work is theoretical and uses only synthetic experiments. It does not involve private data, human subjects, or high-risk model deployment, and it preserves submission anonymity.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The work may positively impact online matching and recommendation systems by reducing computational burden. Since the paper is primarily theoretical and uses only synthetic experiments, we do not identify direct negative societal impacts (see Appendix A.14).

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release high-risk models, scraped datasets, or generative systems. The experiments use synthetic data.

12.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

13. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A].

    Justification: The paper does not use external datasets, pretrained models, or existing code assets. Prior algorithms and related works are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

14. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: The paper does not introduce or release a new dataset, benchmark, or model asset. The experiments are based on synthetic data generated according to the procedure described in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

15. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

16. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve human subjects or crowdsourcing experiments, so IRB approval is not applicable.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

17. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The core methodology, theory, algorithms, and experiments do not use LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.