Interactive Learning with Pricing for
Optimal and Stable Allocations in Markets

Yigit Efe Erginbas
UC Berkeley

Soham Phade
Salesforce Research

Kannan Ramchandran
UC Berkeley

Abstract

Large-scale online recommendation systems must facilitate the allocation of a limited number of items among competing users while learning their preferences from user feedback. As a principled way of incorporating market constraints and user incentives in the design, we consider our objectives to be two-fold: maximal social welfare with minimal instability. To maximize social welfare, our proposed framework enhances the quality of recommendations by exploring allocations that optimistically maximize the rewards. To minimize instability, a measure of users' incentives to deviate from recommended allocations, the algorithm prices the items based on a scheme derived from the Walrasian equilibria. Though it is known that these equilibria yield stable prices for markets with known user preferences, our approach accounts for the inherent uncertainty in the preferences and further ensures that the users accept their recommendations under offered prices. To the best of our knowledge, our approach is the first to integrate techniques from combinatorial bandits, optimal resource allocation, and collaborative filtering to obtain an algorithm that achieves sub-linear social welfare regret as well as sub-linear instability. Empirical studies on synthetic and real-world data also demonstrate the efficacy of our strategy compared to approaches that do not fully incorporate all these aspects.

1 INTRODUCTION

Online recommendation systems have become an integral part of our socioeconomic life with the rapid increases in online services that help users discover options match-
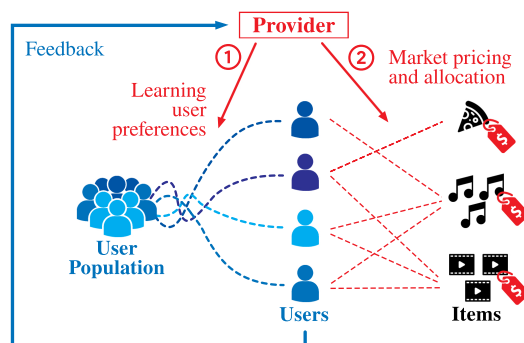
Figure 1: The provider interactively learns the user preferences to achieve socially optimal and stable allocations.

ing their preferences. Despite providing efficient ways to discover information about the preferences of users, they have played a largely complementary role in searching and browsing with little consideration of the accompanying *markets* within which items are offered to users at certain prices. Indeed, in many real-world scenarios, the recommendation of items that have associated notions of limited capacities naturally give rise to a market setting where the users compete for the *allocation* of these items.

Allocation constraints and underlying marketplaces are common in numerous recommendation contexts such as point-of-interest recommendation (Yelp), accommodation (Airbnb, Booking.com), e-commerce (Amazon, eBay), ride-share (Uber, Lyft), or job search (TaskRabbit, LinkedIn). As similar systems become more ubiquitous and impactful in the broader aspects of daily life, there is a huge application drive and potential for delivering recommendations that respect the requirements of the market. Therefore, it is crucial to consider market-aware recommendation systems to maximize the user experience.

Main Challenges: We model user preferences as rewards that users obtain by consuming different items, while social welfare is the aggregate reward over the entire system comprising multiple users with heterogeneous preferences. The *provider*, who continually recommends items to the users, receives interactive feedback from them. The provider's goal is to maximize social welfare at all times while respecting the (time-varying) allocation constraints: indeed

we consider system dynamics in terms of user demands and item capacities that change with time to be an important aspect of our problem. In the process of identifying the best allocation of items to target users, the provider encounters challenges regarding two separate aspects of the problem. The first challenge is to learn user preferences from interactive feedback, and the second is to find allocations and pricing that will result in maximal social welfare within the constraints of the market.

The first challenge of *learning* arises from the fact that the provider does not have exact knowledge of the user preferences ahead of time, and hence has to learn them while continually taking action. In order to learn user preferences efficiently, previous works have established interactive recommender systems that query the users with well-chosen recommendations (Kawale et al., 2015; Zhao et al., 2013). Typically, these works consider a setting where a single user arrives at each round, and the system makes a recommendation that will match the user's preferences. However, this approach is no longer applicable in settings having an associated market structure, as the simultaneous recommendations made to different users must also respect the constraints of the market.

This brings us to the second challenge that relates to the *allocation* aspect of the problem induced by market constraints. Even if providing the users with their most preferred items would result in high rewards, such an allocation may not respect the constraints of the market in general. Therefore, to achieve better outcomes, the provider needs to take the constraints into account and only make allocations within the limits of these constraints. However, since the users care only to maximize their own gains, these allocations may be at odds with the incentives of the users.

One common solution to this problem is to employ *pricing* mechanisms. As the prices provide users with incentives to change their behavior, they can be used to induce *stable* outcomes in which no user has an incentive to deviate from the recommended allocation. The celebrated theory of competitive equilibrium introduced by Arrow (1951) shows that stable allocation and prices can be computed when the user preferences are *fully known*. However, when the preferences are not known, it is not clear whether stability is even achievable. In this work, we address this question in detail and show that it is indeed possible to be stable by learning preferences efficiently. We accomplish this by learning the relevant user preferences efficiently and showing that instability grows only sublinearly over time so as to guarantee exact stability asymptotically.

Besides achieving stability, it is important to give users the choice of determining whether to accept or reject their recommendation based on the prices. To this end, we assume that a recommendation is accepted by a user only if the user expects to obtain more reward from this recommended item

than its price. Therefore, while deciding on the prices of items, the provider also needs to be aware of user choices to achieve higher social welfare. To the best of our knowledge, we are not aware of any prior work that introduces price-based choice in online recommendation systems.

Interactive Learning for Allocation and Pricing (ILAP): As depicted in Figure 1, the provider has two tasks: (1) to learn user preferences by making recommendations that will give rise to informative responses, (2) to find allocations and pricing that will result in maximal social welfare with minimal instability. With these goals, we develop a market-aware mechanism for the provider. By recommending items, the provider helps the users to narrow down their options so that users can evaluate their preferences among a smaller set of items. In addition, being aware of the market structure, the provider determines the item prices that play the role of an intermediary for stabilizing the market.

Since there exists a trade-off between maximizing social welfare using historical data and gathering new information to improve performance in the future, the provider encounters the well-known *exploration-exploitation* dilemma. In the literature of interactive recommendations, this dilemma is typically formulated as a multi-armed bandit problem where each arm corresponds to the recommendation of an item to the target user (Zhao et al., 2013; Barraza-Urbina, 2017; Wang et al., 2019). After the provider makes recommendations to the user, the user feeds their reward information back to the provider to improve future recommendations. However, in contrast to these prior works, our setting further requires that a collection of actions taken for different users does not cause instability and respects the constraints of the market.

Based on the standard OFU (Optimism in Face of Uncertainty) principle (Dani et al., 2008; Abbasi-Yadkori, 2011), we devise a procedure that estimates the mean reward values optimistically so as to solve the system problem of allocation and pricing. As is standard with OFU-based methods, our algorithm maintains a confidence set of the mean rewards for all user-item pairs. If it has less information about some user-item allocation pair, the confidence set becomes larger in the corresponding direction. Then, due to optimism, the algorithm becomes more inclined to attempt the corresponding allocation pairs to explore and collect more information. To reduce instability, it chooses the prices based on the equilibrium prices resulting from the problem with the optimistic estimation of the mean rewards. However, instead of directly offering these equilibrium prices, it reduces the prices by an amount that depends on the current confidence of the algorithm, to ensure that recommendations are likelier to be accepted.

To capture the correlation between the preferences of different users for different items, we employ latent factor (also known as matrix factorization) models that have been

widely applied in recommender systems (Resnick and Varian, 1997; Sarwar et al., 2001; Schafer et al., 2007; Bennett and Lanning, 2007; Koren et al., 2009). We associate each user and item with an unknown feature vector and assume that expected rewards can be represented through a linear interaction of these vectors. We first consider the scenario where the features of the items are available as contexts to the provider and the provider only needs to learn the hidden information regarding the users. Then, relaxing this assumption, we consider the case where all of the feature vectors are unknown, which is equivalent to assuming that the expected reward matrix is low rank.

Our Contributions

- We formulate the joint pricing and recommendation problem for achieving socially optimal and stable allocation of items (with time-varying capacities) to users (with time-varying demands). To our knowledge, we are the first to consider the direct effects of pricing *actions* on the performance metrics in an online learning setup.

- We model the structure of the preferences using linear latent factors. Considering settings with and without contextual information, we propose algorithms that achieve sublinear social welfare regret and instability bounds.

- We model users' ability to accept or reject offers based on their valuation of the items and the offered prices.

- For the setting where features of the items are given as context, we obtain an algorithm that achieves $\widetilde{\mathcal{O}}((NR)^{1/4}(nMT)^{3/4})$ social welfare regret and instability for a problem with T rounds, N users (at most n of which are simultaneously active), M items and length-R feature vectors. Additionally, if the users are not allowed to reject the items, we show that the algorithm can achieve $\widetilde{\mathcal{O}}(\sqrt{NMnRT})$ regret and instability.

- For the more challenging setting where the provider has no contextual information about the items, we obtain $\widetilde{\mathcal{O}}(((N + M)R)^{1/4}(NMT)^{3/4})$ regret and instability.

Experiments: We run experiments both on synthetic and real-world datasets to show the efficacy of the proposed algorithms. We demonstrate the significance of various aspects of our framework by contrasting our strategy with algorithms that fall short of capturing all of the components. Results show that the proposed algorithm can obtain significant improvements over these naive approaches.

1.1 Related Work

Combinatorial Multi-Armed Bandits (CMAB): The semi-bandits framework of Audibert et al. (2011) and the CMAB frameworks of Chen et al. (2013) and Kveton et al. (2015) model problems where a player chooses a subset of arms in each round and observes random rewards from the played arms. Since the allocations in our scenario correspond to selecting a subset of user-item pairs, it shows parallelism with these frameworks. However, the actions of the provider in our problem are not limited to the allocation, but also include the choice of accompanying pricing. Since the selection of prices affects social welfare and instability, it adds an additional degree of complexity to the problem of the provider, as we address in this work.

Structured Linear Bandits: The structured linear bandit frameworks of Johnson et al. (2016) and Combes et al. (2017) as well as low-rank linear bandit framework of Lu et al. (2021) offer methods to deal with the intrinsic structures between different actions in bandit problems. However, these works only concentrate on linear reward mechanisms (which would correspond to only observing the total reward of all allocations) while our approach has the additional ability to leverage the combinatorial nature of the reward feedback in our problem setting.

Bandits with Knapsacks: Wu et al. (2015), Agrawal and Devanur (2016), and Yang et al. (2020) address the problem of online allocations with constraints using primal-dual type approaches. However, in these frameworks, the dual variables appear only as latent variables that are introduced to balance the trade-off between future and current allocations. On the other hand, in our problem formulation, the prices are a part of the *action* taken by the platform and play a direct role in market instability and social welfare because the acceptance/rejection decisions are based on prices.

Bandits in Economics: Approaches based on bandits have also been applied to other economic contexts. One recent line of literature studies algorithms for learning socially-optimal matching in two-sided markets without transfers or incentives (Liu et al., 2020; Johari et al., 2021). Jagadeesan et al. (2021) considers learning stable one-to-one matching with transfers, but their model does not capture the effects of transfers on whether the users accept or reject the offered match. In the context of economics, bandit methods have also been applied to dynamic pricing (Kleinberg and Leighton, 2003), incentivizing exploration (Frazier et al., 2014) learning under competition (Aridor et al., 2019).

Pricing Mechanisms for Stable Outcomes: In the literature of microeconomics, one of the standard approaches to discovering optimal allocations with stable prices is through a process called *tâtonnement* (groping, trial-and-error) (Walras, 2014). During this process, each user repeatedly submits its demand at prices chosen by the provider while the provider successively adjusts the prices in response to the user's demand, so that capacity constraints are satisfied in equilibrium (Arrow, 1951). Since this equilibrium ensures that the items are allocated to users that will obtain the largest reward, it has found applications in various contexts such as auction design (Smith, 1965) or bandwidth allocation over networks (Kelly, 1997). How-

ever, these mechanisms assume that the users know their preferences for all possible items and require the users to repeatedly report their potential demands at many different prices. Since it might take many iterations until convergence, asking the users to repeatedly respond would create an unacceptable level of burden for them. Our approach circumvents both of these obstacles by learning the user preferences and making relevant recommendations.

Recommendation with Capacity Constraints: Christakopoulou et al. (2017) and Makhijani et al. (2019) use the notion of constrained resources to model and solve the problem of recommendation with capacity constraints. However, these works only consider optimizing the recommendation accuracy subject to capacity constraints without any consideration of the interactive learning mechanisms that discover user preferences through recommendations. The rotting bandit framework of Levine et al. (2017) can also model congestion in recommendation settings by allowing each arm's reward to depend on the number of times it was played in the past. However, their model makes recommendations to sequentially arriving users instead of considering a market that contains multiple users that simultaneously interact with the provider.

2 PROBLEM SETTING

We use bold font for vectors $\mathbf{x}$ and matrices $\mathbf{X}$, and calligraphic font $\mathcal{X}$ for sets. For a vector $\mathbf{x}$, we denote its i-th entry by x_i. For a matrix $\mathbf{X}$, we denote its (i, j)-th entry by X_{ij}, i-th row by $[\mathbf{X}]_{i,:}$ and j-th column by $[\mathbf{X}]_{:,j}$. We denote the inner product of two matrices by $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{trace}(\mathbf{A}^{\mathrm{T}}\mathbf{B})$ and the Hadamard (element-wise) product of two matrices by $\mathbf{A} \circ \mathbf{B}$. We denote the $L_{2,2}$ (Frobenius) norm of a matrix by $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle}$, and the $L_{2,\infty}$ norm of a matrix by $\|\mathbf{A}\|_{2,\infty} = \max_i \|[\mathbf{A}]_{i,:}\|_2$.

Suppose the market consists of a set of users $\mathcal{N}$ of size N and a set of items $\mathcal{I}$ of size M. The items are allocated to the users in multiple rounds denoted by $t \in \mathbb{N}$. Allocation of an item $i \in \mathcal{I}$ to a user $u \in \mathcal{N}$ results in a random reward that has a distribution unknown to the system provider. The expected reward obtained from allocating item i to user u is denoted by Θ^*_{ui}, and these values are collected into the reward matrix denoted by $\mathbf{\Theta}^* \in \mathbb{R}^{N \times M}$.

We assume that each item has (time-varying) capacity that corresponds to the maximum number of users it can be allocated to. We denote the capacity of item $i \in \mathcal{I}$ at round t by c_i^t, and collect them into vectors $\mathbf{c}_t \in \mathbb{R}^M$. Similarly, each user has a (time-varying) demand that corresponds to the maximum number of different items it can get allocated. We denote the demand of user $u \in \mathcal{N}$ at round t by d_u^t, and collect them into vectors $\mathbf{d}_t \in \mathbb{R}^N$. Therefore, each item can only be allocated to at most c_i^t different users, while each user can only get allocated at most d_u^t different types

of items in the round t. We shall call these the *allocation constraints*. We also denote the set of active users at round t by $\mathcal{N}_t := \{u \in \mathcal{N} : d_u^t > 0\}$ and denote the maximum number of simultaneously active users by $n := \max_t |\mathcal{N}_t|$.

Let $\mathbf{X}_t$ denote the *allocation matrix* for round t where the (u, i)-th entry is one if user u is allocated item i at round t, and zero otherwise. Due to the allocation constraints, all $\mathbf{X}_t$ must belong to the set of valid allocations defined as

$$\mathcal{X}_t := \{\mathbf{X} \in \{0,1\}^{N \times M} : \mathbf{X}\mathbb{1} \le \mathbf{d}_t \text{ and } \mathbf{X}^{\mathrm{T}}\mathbb{1} \le \mathbf{c}_t\},$$

where the inequalities are entry-wise and $\mathbb{1}$ denotes the all-ones vector of appropriate size.

2.1 Optimal and Stable Allocations

An allocation $\mathbf{X}$ is optimal if it maximizes the social welfare $\mathcal{W}(\mathbf{X}) := \langle \mathbf{X}, \mathbf{\Theta}^* \rangle$. Hence, if the provider had known the mean reward matrix $\mathbf{\Theta}^*$, the optimal allocation $\mathbf{X}_t^*$ at time t would be obtained by solving the integer program

$$\mathbf{X}_t^* \in \arg\max_{\mathbf{X} \in \mathcal{X}_t} \langle \mathbf{X}, \mathbf{\Theta}^* \rangle. \tag{2.1}$$

This integer program can be relaxed to a linear program by dropping the integral constraints (setting $X_{ui} \in [0, 1]$). Since the integrality gap of this relaxation is one [1], the relaxed problem can be used to obtain a solution for (2.1).

On the other hand, an allocation $\mathbf{X}$ is stable with prices $\mathbf{p}$ if for each individual user, the allocated set of items offers the maximum gains among all possible allocations. Here, gains are measured in terms of the surplus they receive from consuming the allocated items at the given prices. Thus, the notion of stability corresponds to the alignment between the market outcome and each user's preferences. Formally,

Definition 2.1. *Let* $\mathbf{x} \in \{0,1\}^M$ *denote the allocation made to a user and let* $\boldsymbol{\theta}_u^* = [\mathbf{\Theta}^*]_{u,:}$ *denote the u-th row of the mean reward matrix. Then, the surplus of user u from allocation* $\mathbf{x}$ *under prices* $\mathbf{p}$ *is given by*

$$\psi_u(\mathbf{x}, \mathbf{p}) := \langle \mathbf{x}, \boldsymbol{\theta}_u^* - \mathbf{p} \rangle.$$

An allocation $\mathbf{X} \in \mathcal{X}_t$ *with prices* $\mathbf{p}$ *is stable if every user's allocation maximizes their own surplus, i.e,*

$$\mathbf{x}_u \in \arg\max_{\mathbf{x} \in \mathcal{X}_u^t} \psi_u(\mathbf{x}, \mathbf{p}), \tag{2.2}$$

for all $u \in \mathcal{N}$, *where* $\mathbf{x}_u = [\mathbf{X}]_{u,:}$ *is the u-th row of* $\mathbf{X}$ *and* $\mathcal{X}_u^t := \{\mathbf{x} \in \{0,1\}^M : \mathbb{1}^{\mathrm{T}}\mathbf{x} \le d_u^t\}$ *is the set of allocation vectors that satisfy the constraints of user u at time t.*

A fundamental property of this market model shows that if the allocation and prices are set as the primal and dual optimal variables of the relaxed allocation problem, the market

---

[1]The integrality gap is the maximum ratio between the optimal values of the relaxation and integer program. (see Appendix A)

outcome becomes stable (Walras, 2014). To see this, we can consider the (partial) Lagrangian

$$\mathcal{L}(\mathbf{X}, \mathbf{p}; \boldsymbol{\Theta}) := \langle \mathbf{X}, \boldsymbol{\Theta} \rangle + \mathbf{p}^{\mathsf{T}}(\mathbf{c}_t - \mathbf{X}^{\mathsf{T}}\mathbb{1}),$$

where $\mathbf{p} \geq 0$ are the Lagrange multipliers associated with the capacity constraints. Then, the corresponding dual optimization problem is given by $\min_{\mathbf{p} \geq 0} g(\mathbf{p}; \boldsymbol{\Theta}^*)$ where

$$g(\mathbf{p}; \boldsymbol{\Theta}) := \max_{\mathbf{X} \in [0,1]^{N \times M}} \{\mathcal{L}(\mathbf{X}, \mathbf{p}; \boldsymbol{\Theta}) | \mathbf{X}\mathbb{1} \leq \mathbf{d}_t\} \quad (2.3)$$

is the dual function. Then, due to linear programming duality, the primal-dual optimal variables $(\mathbf{X}_t^*, \mathbf{p}_t^*)$ satisfy the stability condition given in (2.2). Therefore, if the provider chooses the allocation $\mathbf{X}_t^*$ with prices $\mathbf{p}_t^*$, the market outcome is both optimum and stable. However, since the mean reward matrix $\boldsymbol{\Theta}^*$ is unknown to the provider, it cannot directly solve the allocation and pricing problem as given.

Thus, we consider an online framework where the provider makes sequential recommendations to users and the users provide feedback about their recommendations so that the provider can learn user preferences. For clarity of exposition, we first present the problem without allowing the users to reject their offers in Section 2.2 and then extend our formulation to capture the user choices in Section 2.3.

2.2 Problem Formulation

In this section, we formulate the provider's problem and its objective. At each round t, the provider determines an allocation $\mathbf{X}_t \in \mathcal{X}_t$ along with a price vector $\mathbf{p}_t \in \mathbb{R}_+^M$. For convenience, we also define for each allocation $\mathbf{X}_t \in \mathcal{X}_t$ an equivalent set representation $\mathcal{A}_t \subseteq \mathcal{N} \times \mathcal{I}$ such that $(u, i) \in \mathcal{A}_t$ if and only if user u is allocated item i.

Since exact stability is unattainable when preferences are unknown, we need to formalize a measure of instability that captures how far we are from exact stability. Intuitively, the users want to obtain the largest possible surplus under the current condition of the market described by the prices $\mathbf{p}_t$. Therefore, the level of instability under prices $\mathbf{p}_t$ is related to the gap between the surplus the user obtained from allocation $\mathbf{X}_t$ and the maximum surplus the user could have obtained from any allocation. For this reason, we define the instability of a user as follows.

Definition 2.2. *Given an allocation $\mathbf{X} \in \mathcal{X}_t$ with prices $\mathbf{p}$, let $\mathbf{x}_u = [\mathbf{X}]_{u,:}$ denote the allocation made to user u. Then, the maximum surplus of user u at time t is*

$$\psi_{t,u}^*(\mathbf{p}) := \max_{\mathbf{x} \in \mathcal{X}_u^t} \psi_u(\mathbf{x}, \mathbf{p}),$$

and the instability of user u at time t is the difference

$$\mathcal{S}_u^t(\mathbf{x}_u, \mathbf{p}) := \psi_{t,u}^*(\mathbf{p}) - \psi_u(\mathbf{x}_u, \mathbf{p}).$$

Remark 2.3. Since $\psi_{t,u}^*(\mathbf{p}) \geq \psi_u(\mathbf{x}_u, \mathbf{p})$ for any $\mathbf{x}_u \in \mathcal{X}_u^t$, the instability $\mathcal{S}_u^t(\mathbf{x}_u, \mathbf{p})$ is always non-negative. Furthermore, we see that instability $\mathcal{S}_u^t(\mathbf{x}_u, \mathbf{p})$ is zero for all users $u \in \mathcal{N}$ if and only if the market outcome is (exactly) stable according to Definition 2.1.

After the provider decides on an allocation $\mathbf{X}_t$ with prices $\mathbf{p}_t$, it observes a random feedback R_{ui}^t for each user-item pair $(u, i) \in \mathcal{A}_t$ that is recommended. We denote by H_t the history $\{\mathbf{X}_\tau, \mathbf{p}_\tau, (R_{ui}^\tau)_{(u,i) \in \mathcal{A}_\tau}\}_{\tau=1}^{t-1}$ of observations available to the provider when choosing the next allocation $\mathbf{X}_t$ and prices $\mathbf{p}_t$. The provider employs a policy $\pi = \{\pi_t | t \in \mathbb{N}\}$, which is a sequence of functions, each mapping the history H_t to an action $(\mathbf{X}_t, \mathbf{p}_t)$.

The task of the provider is to repeatedly allocate the items to the users and choose the prices such that it can achieve two goals simultaneously: maximum *social welfare* and minimum *instability*. To measure the performance of policies in achieving the objectives of maximum social welfare and minimum instability, we define the following metrics:

Definition 2.4. *For a policy π, its social-welfare regret in T rounds is defined as*

$$\mathcal{R}^{SW}(T, \pi) = \sum_{t=1}^{T} \mathcal{W}(\mathbf{X}_t^*) - \sum_{t=1}^{T} \mathcal{W}(\mathbf{X}_t),$$

and its instability in T rounds is defined as

$$\mathcal{R}^I(T, \pi) = \sum_{t=1}^{T} \sum_{u \in \mathcal{N}} \mathcal{S}_u^t(\mathbf{x}_u^t, \mathbf{p}_t),$$

where $\mathbf{X}_t^* \in \arg\max_{\mathbf{X} \in \mathcal{X}_t} \langle \mathbf{X}, \boldsymbol{\Theta}^* \rangle$ denotes optimum allocation at time t and $\mathbf{x}_u^t = [\mathbf{X}_t]_{u,:}$ is the u-th row of $\mathbf{X}_t$.

2.3 Allowing Users to Accept or Reject the Offers

In this section, we model the users' ability to accept or reject the offers based on their valuations and the prices of the offered items. When an item i is offered to user u at price p_i, we assume that the user accepts the item and the system obtains a positive reward if and only if the user expects to get a positive surplus from this consumption, i.e. $\Theta_{ui}^* \geq p_i$. Consequently, we modify the definition of the social welfare of an allocation $\mathbf{X}$ with prices $\mathbf{p}$ as

$$\mathcal{W}(\mathbf{X}, \mathbf{p}) := \langle \mathbf{X}, \boldsymbol{\Theta}^* \circ \mathbb{1}\{\boldsymbol{\Theta}^* \geq \mathbb{1}\mathbf{p}^{\mathsf{T}}\} \rangle$$

and modify the definition of the surplus of user u as

$$\psi_u(\mathbf{x}, \mathbf{p}) := \langle \mathbf{x}, (\boldsymbol{\theta}_u^* - \mathbf{p}) \circ \mathbb{1}\{\boldsymbol{\theta}_u^* \geq \mathbf{p}\} \rangle.$$

The definitions of maximum surplus and instability of each user in Definition 2.2 as well as the definitions of social welfare regret and instability in Definition 2.4 are also modified accordingly. Note that these modifications do not affect the feedback model and the provider is assumed to observe the random feedback R_{ui}^t for each recommendation pair $(u, i) \in \mathcal{A}_t$ regardless of their acceptance.

3 METHODOLOGY

In order to facilitate our analysis, we start with assumptions that are standard in the online learning literature.

Assumption 1. *For all $(u,i) \in \mathcal{N} \times \mathcal{I}$ and $t \in \mathbb{N}$, the rewards R_{ui}^t are independent and η-sub-Gaussian with mean $\Theta_{ui}^* \in [0,1]$, i.e., $\mathbb{E}[\exp(\lambda(R_{ui}^t - \Theta_{ui}^*))] \leq \exp(\lambda\eta^2/2)$ almost surely for any λ.*

As is common in recommendation settings, we model the correlation between users and items through a linear latent factor model. Each user u (item i) is associated with a feature vector $\mathbf{f}_u \in \mathbb{R}^R$ ($\boldsymbol{\phi}_i \in \mathbb{R}^R$) in a shared R-dimensional space (typically $R \ll N, M$), and the mean reward of each user-item allocation pair is given by a linear form.

Assumption 2. *For all $(u,i) \in \mathcal{N} \times \mathcal{I}$, and $t \in \mathbb{N}$, the expected reward is given by $\Theta_{ui}^* = \langle \mathbf{f}_u, \boldsymbol{\phi}_i \rangle$ with feature vectors satisfying $\|\mathbf{f}_u\|_2 \leq 1$ and $\|\boldsymbol{\phi}_i\|_2 \leq 1$.*

In order to make use of initial historical data, we assume that the algorithm has access to an initial estimate $\boldsymbol{\Theta}_\circ$ that satisfies $\|\boldsymbol{\Theta}_\circ - \boldsymbol{\Theta}^*\|_F \leq G$. Note that one can also set $\boldsymbol{\Theta}_\circ = \mathbf{0}$ and let G be a number satisfying $\|\boldsymbol{\Theta}^*\|_F \leq G$.

3.1 Contextual Interactive Learning for Allocation and Pricing (CX-ILAP)

We start our analysis by considering the case where the feature vectors $\boldsymbol{\phi}_i$ for the items $i \in \mathcal{I}$ are given as contextual information, but the feature vectors $\mathbf{f}_i$ for the users $u \in \mathcal{N}$ are unknown. Letting $\boldsymbol{\Phi} \in \mathbb{R}^{M \times R}$ be the matrix with rows $\boldsymbol{\phi}_i$, this assumption corresponds to asserting that the mean reward matrix $\boldsymbol{\Theta}^*$ belongs to the set

$$\mathcal{F} := \{\boldsymbol{\Theta} : \exists \mathbf{F} \in \mathbb{R}^{N \times R}, \boldsymbol{\Theta} = \mathbf{F}\boldsymbol{\Phi}^T, \|\mathbf{F}\|_{2,\infty} \leq 1\}.$$

Our method, summarized in Algorithm 1, follows the standard OFU (Optimism in the Face of Uncertainty) principle of Abbasi-Yadkori (2011). It maintains a confidence set $\mathcal{C}_t$ which contains the true parameter $\boldsymbol{\Theta}^*$ with high probability and chooses the allocation $\mathbf{X}_t$ according to

$$(\mathbf{X}_t, \boldsymbol{\Theta}_t) \in \underset{(\mathbf{X}, \boldsymbol{\Theta}) \in \mathcal{X}_t \times \mathcal{C}_t}{\arg\max} \langle \mathbf{X}, \boldsymbol{\Theta} \rangle. \quad (3.1)$$

Then, in order to construct its pricing strategy, it computes the optimal dual variables $\bar{\mathbf{p}}_t$ for the allocation problem described by $\boldsymbol{\Theta}_t$, i.e., it solves the dual problem

$$\bar{\mathbf{p}}_t \in \underset{\mathbf{p} \geq 0}{\arg\min} \, g(\mathbf{p}; \boldsymbol{\Theta}_t), \quad (3.2)$$

where the dual function is as defined in (2.3). Typically, the faster the confidence set $\mathcal{C}_t$ shrinks, the lower regret and instability we have. However, the main difficulty is to construct a series of $\mathcal{C}_t$ that leverage the combinatorial observation model as well as the structure of the mean reward matrix so that we have low regret bounds. In this work,

---

Algorithm 1 Interactive Learning for Allocation and Pricing

Require: parameters $T, \delta, \alpha, \gamma, \nu$, and initial estimate $\boldsymbol{\Theta}_\circ$
1: **for** $t = 1, 2, \ldots, T$ **do**
    ▷ Find the regularized least squares estimate:
2:    $\widehat{\boldsymbol{\Theta}}_t = \arg\min_{\boldsymbol{\Theta} \in \mathcal{F}} \{L_{2,t}(\boldsymbol{\Theta}) + \gamma\|\boldsymbol{\Theta} - \boldsymbol{\Theta}_\circ\|_F^2\}$
    ▷ Construct the confidence set:
3:    $\mathcal{C}_t := \{\boldsymbol{\Theta} \in \mathcal{F} : \|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}_t\|_{E_{2,\infty}^t} \leq \sqrt{\rho_t(\delta, \alpha, \gamma)}\}$
    ▷ Compute the OFU allocation:
4:    $(\mathbf{X}_t, \boldsymbol{\Theta}_t) \in \arg\max_{(\mathbf{X}, \boldsymbol{\Theta}) \in \mathcal{X}_t \times \mathcal{C}_t} \langle \mathbf{X}, \boldsymbol{\Theta} \rangle$
    ▷ Compute the prices:
5:    $\bar{\mathbf{p}}_t \in \arg\min_{\mathbf{p} \geq 0} g(\mathbf{p}; \boldsymbol{\Theta}_t)$
6:    $\mathbf{p}_t = \bar{\mathbf{p}}_t - \nu\sqrt{w_t}$ for w_t defined in (3.4)
7:    Offer the allocation $\mathbf{X}_t$ together with prices $\mathbf{p}_t$
8:    Observe R_{ui}^t for all $(u,i) \in \mathcal{A}_t$
9: **end for**

---

we consider constructing confidence sets that are centered around the regularized least square estimates. We let the cumulative squared prediction error of $\boldsymbol{\Theta}$ at time t be

$$L_{2,t}(\boldsymbol{\Theta}) = \sum_{\tau=1}^{t-1} \sum_{(u,i) \in \mathcal{A}_\tau} (\Theta_{ui} - R_{ui}^\tau)^2,$$

and calculate the regularized least squares estimate as given in line 2 of Algorithm 1. Then, the confidence sets take the form $\mathcal{C}_t := \{\boldsymbol{\Theta} \in \mathcal{F} : \|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}_t\|_{E_{2,\infty}^t} \leq \sqrt{\rho_t}\}$, where ρ_t is an appropriately chosen confidence parameter, and $\|\cdot\|_{E_{2,\infty}^t}$ is the regularized empirical $L_{2,\infty}$ norm, namely,

$$\|\boldsymbol{\Delta}\|_{E_{2,\infty}^t}^2 := \max_{u \in \mathcal{N}} \sum_{i \in \mathcal{I}} (n_{ui}^t + \gamma)\Delta_{ui}^2,$$

where $n_{ui}^t := \sum_{\tau=1}^{t-1} \mathbb{1}\{(u,i) \in \mathcal{A}_\tau\}$ is the number of times item i has been allocated to user u before time t (excluding time t). Hence, the empirical norm is a measure of discrepancy that weighs the entries depending on how much they have been explored. Roughly speaking, since the confidence region is wider in directions that are not yet well-explored, the OFU step described in (3.1) is more inclined to make allocations that include the corresponding user-item pairs. In order to obtain low-regret guarantees for the allocations, the first step is to choose the correct ρ_t parameter such that $\mathcal{C}_t$ will contain the true parameter $\boldsymbol{\Theta}^*$ for all t with high probability. The following Lemma establishes that, if we set $\rho_t = \rho_t^*(\delta, \alpha, \gamma)$, the resulting confidence sets have the desired properties.

Lemma 3.1. *For any $\delta, \alpha, \gamma > 0$, if the confidence sets are*

$$\mathcal{C}_t := \left\{\boldsymbol{\Theta} \in \mathcal{F} : \|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}_t\|_{E_{2,\infty}^t} \leq \sqrt{\rho_t^*(\delta, \alpha, \gamma)}\right\} \quad (3.3)$$

with the confidence parameter

$$\begin{aligned}\rho_t^*(\delta, \alpha, \gamma) &:= 8\eta^2 R \log(3N/(\alpha\delta)) + 4\gamma G^2 \\ &+ 2\alpha t\sqrt{M}\left[8 + \sqrt{8\eta^2 \log(4MNt^2/\delta)}\right],\end{aligned}$$

then, with probability at least $1 - 2\delta$, $\mathcal{C}_t \ni \boldsymbol{\Theta}^*$, for all t.

If the users do not have the option to reject the recommendations, then the algorithm can offer the items at the estimated equilibrium price by setting $\mathbf{p}_t = \bar{\mathbf{p}}_t$, and it can provably achieve low regret and instability. However, when the users are allowed to reject the offers based on their expected surplus, such a pricing strategy cannot achieve provable optimality guarantees for the social welfare since it does not alleviate the possibility of rejections. Instead, our approach is to reduce the likelihood of rejections by choosing the prices as $\mathbf{p}_t = \bar{\mathbf{p}}_t - \nu \sqrt{w_t}$, where w_t is the *confidence width* of the allocation $\mathbf{X}_t$, defined as

$$w_t := \sum_{(u,i) \in \mathcal{A}_t} (n_{ui}^t + \gamma)^{-1}, \qquad (3.4)$$

and $\nu > 0$ is a parameter that depends on the problem size. As we show in the appendix, this pricing strategy can guarantee the number of rejections to grow only sublinearly in T, allowing us to achieve sub-linear regret and instability even when the users are allowed to reject.

Finally, we can show that if the algorithm follows the OFU allocations in (3.1) and prices in (3.2), while constructing the confidence sets according to (3.3), it obtains the following overall regret guarantee.

Theorem 3.2. *Under Assumptions 1 and 2 with known item feature matrix $\mathbf{\Phi}$, set $\nu = \left(4\rho_T^*(\delta, \alpha, \gamma)/(nM^2)\right)^{1/4}$, $\delta > 0$, $\alpha > 0$ and $\gamma \geq 1$. Then, with probability $1 - 2\delta$, the social welfare regret and instability of Algorithm 1 satisfy*

$$\mathcal{R}^{SW}(T, \pi) \leq \sqrt{\kappa_T nMT} + (\kappa_T)^{\frac{1}{4}} (nMT)^{\frac{3}{4}},$$
$$\mathcal{R}^{I}(T, \pi) \leq \sqrt{\kappa_T nMT} + (\kappa_T)^{\frac{1}{4}} (nMT)^{\frac{3}{4}},$$

where $\kappa_T = 8N\rho_T^*(\delta, \alpha, \gamma) \log(1 + T/\gamma)$. Instead, if we assume that the users are not allowed to reject the allocations and set $\nu = 0$, we have $\mathcal{R}^{SW}(T, \pi) \leq \sqrt{\kappa_T nMT}$ and $\mathcal{R}^{I}(T, \pi) \leq \sqrt{\kappa_T nMT}$, with probability $1 - 2\delta$.

Thus, by correctly selecting the parameters of the algorithm, we can simultaneously achieve social welfare regret and instability of order $\widetilde{\mathcal{O}}((NR)^{1/4}(nMT)^{3/4})$, where $\widetilde{\mathcal{O}}$ is the big-O notation ignoring the poly-logarithmic factors of N, M, T, R. On the other hand, if the users are not given the option to reject the offers, then we can obtain social welfare regret and instability of order $\widetilde{\mathcal{O}}(\sqrt{NMnRT})$.

If we were to ignore the structure between the mean rewards of user-item pairs and apply the standard combinatorial bandit algorithms to decide on the allocations, the social welfare regret (in the case where users cannot reject the offers) would be of order $\widetilde{\mathcal{O}}(M\sqrt{NnT})$ (Kveton et al., 2015). Since $n, R \ll M, N$ in many applications of our consideration, our algorithm significantly outperforms this naive approach by leveraging the structure of the rewards.

3.2 Low-Rank Interactive Learning for Allocation and Pricing (LR-ILAP)

In this section, we consider removing the assumption that the provider has access to the feature vectors ϕ_i of the items. In order to use the structure imposed by Assumption 2 without having the knowledge of any of the factors ϕ_i or $\mathbf{f}_i$, we can write that $\mathbf{\Theta}^*$ belongs to the set

$$\mathcal{L} = \{\mathbf{\Theta} \in \mathbb{R}^{N \times M} : \|\mathbf{\Theta}\|_\infty \leq 1, \mathrm{rank}(\mathbf{\Theta}) \leq R\}. \quad (3.5)$$

To adapt Algorithm 1 for this setting, we modify the construction of our confidence regions. In order to take advantage of the structure of the arms, we let $\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_F)$ denote the α-covering number of $\mathcal{L}$ in the Frobenious-norm $\|\cdot\|_F$, and define empirical L_2 norm at time t as

$$\|\mathbf{\Delta}\|_{E_2^t}^2 := \sum_{u \in \mathcal{N}} \sum_{i \in \mathcal{I}} (n_{ui}^t + \gamma) \Delta_{ui}^2.$$

Then, similar to Lemma 3.1, we can show that, with probability $1 - 2\delta$, $\mathbf{\Theta}^*$ belongs to the confidence region

$$\mathcal{Q}_t := \left\{\mathbf{\Theta} \in \mathcal{L} : \|\mathbf{\Theta} - \widehat{\mathbf{\Theta}}_t\|_{E_2^t} \leq \sqrt{\beta_t^*(\delta, \alpha, \gamma)}\right\}, \quad (3.6)$$

where $\widehat{\mathbf{\Theta}}_t = \arg\min_{\mathbf{\Theta} \in \mathcal{L}} \left\{L_{2,t}(\mathbf{\Theta}) + \gamma \|\mathbf{\Theta} - \mathbf{\Theta}_\circ\|_F^2\right\}$ and the confidence parameter is

$$\beta_t^*(\delta, \alpha, \gamma) := 8\eta^2 \log\left(\mathcal{N}(\mathcal{L}, \alpha, \|\cdot\|_F)/\delta\right) + 4\gamma G^2$$
$$+ 2\alpha t\sqrt{NM}\left[8 + \sqrt{8\eta^2 \log(4NMt^2/\delta)}\right].$$

Then, we show that the covering number for $\mathcal{L}$ given in equation (3.5) is upper bounded: $\log \mathcal{N}(\mathcal{L}, \alpha, \|\cdot\|_F) \leq (N + M + 1)R\log(9\sqrt{NM}/\alpha)$. Consequently, the following theorem shows that if the algorithm constructs the confidence sets according to (3.6), it can achieve $\widetilde{\mathcal{O}}(((N + M)R)^{1/4}(NMT)^{3/4})$ social welfare regret and instability. Moreover, if the users are not given the option to reject our offers, then the order improves to $\widetilde{\mathcal{O}}(\sqrt{NM(N + M)RT})$.

Theorem 3.3. *Under Assumptions 1 and 2 with unknown feature matrices, set $\nu = \left(4\beta_T^*(\delta, \alpha, \gamma)/(N^2M^2)\right)^{1/4}$, $\delta > 0$, $\alpha > 0$, $\gamma \geq 1$ and construct the confidence sets according to (3.6). Then, with probability $1 - 2\delta$, the social welfare regret and instability of Algorithm 1 satisfy*

$$\mathcal{R}^{SW}(T, \pi) \leq \sqrt{\lambda_T NMT} + (\lambda_T)^{\frac{1}{4}} (NMT)^{\frac{3}{4}},$$
$$\mathcal{R}^{I}(T, \pi) \leq \sqrt{\lambda_T NMT} + (\lambda_T)^{\frac{1}{4}} (NMT)^{\frac{3}{4}},$$

where $\lambda_T = 8\beta_T^*(\delta, \alpha, \gamma) \log(1 + T/\gamma)$. Instead, if we assume that the users are not allowed to reject our allocations and set $\nu = 0$, we have $\mathcal{R}^{SW}(T, \pi) \leq \sqrt{\lambda_T NMT}$ and $\mathcal{R}^{I}(T, \pi) \leq \sqrt{\lambda_T NMT}$, with probability $1 - 2\delta$.
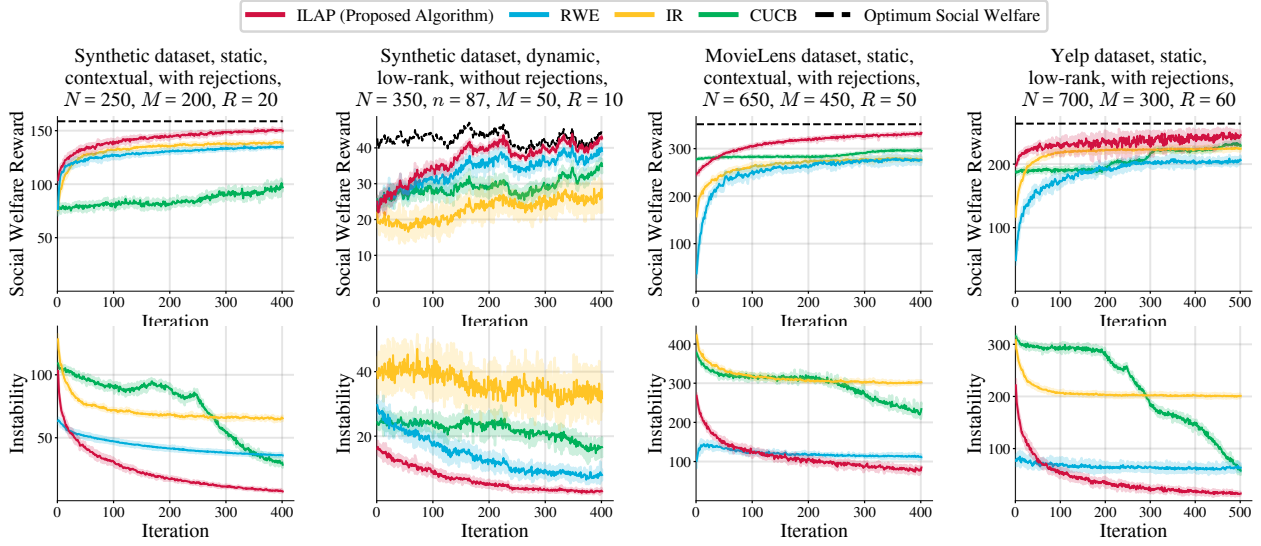
Figure 2: Instantaneous reward and instability at each round in different experimental settings. ILAP is able to achieve lower regret and instability than all other baseline methods in all experimental settings. The experiments are run on 20 problem instances and means are reported together with one standard deviation of uncertainty regions.

4 EXPERIMENTS

In this section, we demonstrate the efficacy of our proposed algorithm through an experimental study over synthetic and real-world datasets. The goal of our experimental evaluation is to evaluate our algorithm for making online recommendations in various market settings.

Baseline Algorithms: To the best of our knowledge, there are no current approaches to online learning for allocations and pricing to maximize social welfare and minimize instability. Therefore, we show the importance of different aspects in our framework by comparing it with algorithms that do not capture all these aspects: a non-bandit algorithm that only aims for momentary performance, a bandit-based algorithm that is unaware of the capacities, and another bandit-based algorithm that is unaware of the correlation between the arms.

(1) RWE (Recommendations without Exploration): It solves for the least squares problem given in line 2 of Algorithm 1 to estimate the mean rewards of user-item allocation pairs. Then, without deliberate exploration, it decides on the allocation and prices as the optimum values for the allocation problem with the least square estimation.

(2) IR (Interactive Recommendations): It runs the interactive recommendation algorithm of Zhao et al. (2013) that finds the best items for each user without considering the capacity constraints. Being oblivious of the capacities, it also does not set any price for the items. Since these recommendations do not necessarily satisfy the constraints, it needs to perform a post-processing step: If it decides to recommend an item to more users than its capacity, it only

chooses a random subset of the assigned users. Heuristically, we let the algorithm observe zero rewards ($R_{ui}^t = 0$) for all the user-item allocations that are eliminated during this post-processing step because otherwise, the allocations would continue to violate the capacity constraints and the algorithm would not reduce its regret.

(3) CUCB (Combinatorial-UCB): It runs the CUCB algorithm of Chen et al. (2013) to decide on allocations without assuming any structure between the users and items. It views the user-item allocation pairs as arms that have no correlation between them. At each round, it decides on the allocation and prices by solving the allocation problem using the upper confidence bounds of the reward parameters.

4.1 Experimental Setup and Datasets

We use a synthetic dataset and two real-world datasets. For the synthetic data, we set the mean reward values $\Theta_{ui}^* = \langle \mathbf{f}_u, \boldsymbol{\phi}_i \rangle$ by randomly choosing length-R feature vectors that satisfy $\|\mathbf{f}_u\|_2 \leq 1$, for all $u \in \mathcal{N}$, and $\|\boldsymbol{\phi}_i\|_2 \leq 1$, for all $i \in \mathcal{I}$. For the real-world data, we consider the following publicly available data sets: a subset of Movielens 100k (Harper and Konstan, 2015) which includes ratings from 650 users on 450 movies and a subset of Yelp Dataset (Yelp, 2021) that corresponds to ratings from 700 users on 300 restaurants. The subsets of users and items are selected such that the datasets contain ratings for most of the user-item pairs. To set the problem up, the missing ratings are completed with an off-the-shelf matrix completion algorithm with matrix factorization (Hastie et al., 2015). We consider problem settings with known item features as well as with unknown features. In the setting with known fea-

tures for the items (contextual), all algorithms have access to the item feature matrix $\mathbf{\Phi}$. In the other case (low-rank), none of the algorithms are provided with any contextual information. To the best of our knowledge, the capacity information is not available in any of the publicly available recommendation datasets. Therefore, we consider instantiating random capacities. We consider settings with static and dynamic (time-varying) capacities/demands. For the static case, we assume that all users request one item at all rounds, and the capacity of each item remains unchanged with time. In the dynamic setting, we allow both the demands $\mathbf{d}_t$ and capacities $\mathbf{c}_t$ vary with time t. We consider that each entry of $\mathbf{d}_t$ is independently sampled from a distribution over $\{0, 1\}$ so that each user u is either active ($d_u^t = 1$) or inactive ($d_u^t = 0$). Similarly, each entry of $\mathbf{c}_t$ is independently sampled from a distribution over nonnegative integers. Lastly, the rewards are assumed to have distribution $R_{ui}^t \sim \mathcal{N}(\Theta_{ui}^*, \eta^2)$. For additional details on the experimental setup, please refer to Appendix G.

4.2 Experimental Results

We provide a summary of our results in Figure 2. The results for other experimental settings are left to Appendix G. The observations can be summed up into the following points: (1) ILAP algorithm (our proposed approach) is able to achieve lower regret and instability than all other baseline methods in all experimental settings. (2) Even though RWE can obtain slightly larger rewards than our method in the initial rounds, it often gets stuck at suboptimal allocations. So, it cannot achieve *no-regret* because it tries to directly exploit the information it acquired so far without making any deliberate explorations. (3) Since IR does not consider the capacities while making the allocations, the recommendations exceed the capacities of the items and lead to very large regrets. Furthermore, since it does not set any prices for the items, it incurs high instability. (4) Since CUCB does not leverage the low-rank structure of the parameters, it needs to sample and learn about each user-item allocation pair separately. Hence, it takes much longer for it to learn the optimum allocations.

5 SOCIETAL IMPACT

Our design objective in this work is to allocate multiple items to multiple users in a way that will maximize social welfare while ensuring stability. Therefore, applications of algorithms developed here would benefit the cumulative welfare of society. However, since our objectives measure only the total wellness in the system, they may raise concerns about the fair distribution of resources. Further, since our work relies on assumptions about user behavior and preferences, our results may not apply when these assumptions do not hold.

6 CONCLUSION

In this paper, we have studied the setting of interactive recommendations that achieve socially optimal and stable allocations under capacity constraints. We considered an online learning setting where the users are given sequential recommendations which they can accept or reject based on the prices determined by the provider. To leverage the correlation between different user-item pairs, we considered linear latent factor models with partially-known and unknown features. We have proposed a sequential allocation and pricing algorithm that enjoys sub-linear social welfare regret and sub-linear instability.

Acknowledgements

References

Abbasi-Yadkori, Y. (2011). Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, pages 2312–2320.

Agrawal, S. and Devanur, N. R. (2016). Linear contextual bandits with knapsacks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3458–3467, Red Hook, NY, USA. Curran Associates Inc.

Aridor, G., Liu, K., Slivkins, A., and Wu, Z. S. (2019). The perils of exploration under competition. *Proceedings of the 2019 ACM Conference on Economics and Computation*.

Arrow, K. J. (1951). An extension of the basic theorems of classical welfare economics. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*.

Audibert, J.-Y., Bubeck, S., and Lugosi, G. (2011). Minimax policies for combinatorial prediction games. *Proceedings of the Twenty-Fourth Annual Conference on Learning Theory*, 19:107–132.

Barraza-Urbina, A. (2017). The exploration-exploitation trade-off in interactive recommender systems. *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 431–435.

Bennett, J. and Lanning, S. (2007). The netflix prize. *Proceedings of KDD Cup Workshop*, pages 3–6.

Bertsekas, D. P. (1991). *Linear network optimization: algorithms and codes*. MIT Press.

Candes, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359.

Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. *Proceedings of the 30th International Conference on Machine Learning*, 28(1):151–159.

Christakopoulou, K., Kawale, J., and Banerjee, A. (2017). Recommendation with capacity constraints. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1439–1448.

Combes, R., Magureanu, S., and Proutiere, A. (2017). Minimal exploration in structured stochastic bandits. *Advances in Neural Information Processing Systems*, 30:1761–1769.

Dani, V., Hayes, T., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland*, pages 355–366.

Frazier, P., Kempe, D., Kleinberg, J., and Kleinberg, R. (2014). Incentivizing exploration. *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, page 5–22.

Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4).

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015). Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16(1):3367–3402.

Heller, I. and Tompkins, C. B. (1957). An extension of a theorem of Dantzig's. *Linear Inequalities and Related Systems, Annals of Mathematics Studies*, 38:247–254.

Jagadeesan, M., Wei, A., Wang, Y., Jordan, M. I., and Steinhardt, J. (2021). Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*.

Johari, R., Manshadi, V., and Walton, N. (2021). Matching while learning. *Operations Research*.

Johnson, N., Sivakumar, V., and Banerjee, A. (2016). Structured stochastic linear bandits. *arXiv preprint arXiv:1606.05693*.

Kawale, J., Bui, H. H., Kveton, B., Tran-Thanh, L., and Chawla, S. (2015). Efficient thompson sampling for online matrix-factorization recommendation. *Advances in Neural Information Processing Systems*, 28:1297–1305.

Kelly, F. (1997). Charging and rate control for elastic traffic. *European Transactions on Telecommunications*, 8(1):33–37.

Kleinberg, R. and Leighton, T. (2003). The value of knowing a demand curve: bounds on regret for online posted-price auctions. *Proceedings of 44th Annual IEEE Symposium on Foundations of Computer Science*, pages 594–605.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. (2015). Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 38:535–543.

Levine, N., Crammer, K., and Mannor, S. (2017). Rotting bandits. *Advances in Neural Information Processing Systems*, 30.

Liu, L., Mania, H., and Jordan, M. I. (2020). Competing bandits in matching markets. *Proceedings of the Twenty-Third Conference on Artificial Intelligence and Statistics*, pages 1618–1628.

Lu, Y., Meisami, A., and Tewari, A. (2021). Low-rank generalized linear bandit problems. *Proceedings of the Twenty-Fourth International Conference on Artificial Intelligence and Statistics*, 130:460–468.

Makhijani, R., Chakrabarti, S., Struble, D., and Liu, Y. (2019). Lore: A large-scale offer recommendation engine with eligibility and capacity constraints. *Proceedings of the Thirteenth ACM Conference on Recommender Systems*, pages 160–168.

Resnick, P. and Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3):56–58.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. *Proceedings Tenth International Conference on World Wide Web*, pages 285–295.

Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S. (2007). Collaborative filtering recommender systems. *The Adaptive Web*, 4321:291–324.

Smith, V. L. (1965). Experimental auction markets and the walrasian hypothesis. *Journal of Political Economy*, 73(4):64–70.

Walras, L. (2014). *Leon Walras: Elements of theoretical economics*. Cambridge University Press.

Wang, Q., Zeng, C., Zhou, W., Li, T., Iyengar, S. S., Shwartz, L., and Grabarnik, G. Y. (2019). Online interactive collaborative filtering using multi-armed bandit with dependent arms. *IEEE Transactions on Knowledge and Data Engineering*, 31(8):1569–1580.

Wu, H., Srikant, R., Liu, X., and Jiang, C. (2015). Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Yang, M., Li, Q., Qin, Z., and Ye, J. (2020). Hierarchical adaptive contextual bandits for resource constraint based recommendation. In *Proceedings of The Web Conference*

2020, WWW '20, page 292–302, New York, NY, USA. Association for Computing Machinery.

Yelp (2021). Yelp open dataset.

Zhao, X., Zhang, W., and Wang, J. (2013). Interactive collaborative filtering. *Proceedings of the Twenty-Second ACM International Conference on Information & Knowledge Management*, pages 1411–1420.

A   RELAXATION OF INTEGER PROGRAM

A traditional linear integer program (IP) in matrix form is formulated as

$$
\begin{aligned}
\max_{\mathbf{x}} \ & \mathbf{t}^{\mathrm{T}}\mathbf{x} \\
\text{s.t. } & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\
& \mathbf{x} \in \mathbb{Z}_+^d.
\end{aligned}
\tag{A.1}
$$

This problem can be relaxed to a linear program by dropping the integral constraints (setting $\mathbf{x} \in \mathbb{R}_+^d$). The integrality gap of an integer maximization program is defined as the maximum ratio of the optimal values for the relaxed linear program and the integer program. When the vector $\mathbf{b}$ is integral and the matrix $\mathbf{A}$ is totally unimodular (all entries are 1, 0, or -1 and every square sub-minor has determinant of +1 or -1) then the integrality gap is one and the solution of the relaxed linear program is integer valued (Bertsekas, 1991).

Hence, we can solve (A.1) by instead solving the following relaxed linear program:

$$
\begin{aligned}
\max_{\mathbf{x}} \ & \mathbf{t}^{\mathrm{T}}\mathbf{x} \\
\text{s.t. } & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\
& \mathbf{x} \in \mathbb{R}_+^d.
\end{aligned}
\tag{A.2}
$$

For a matrix $\mathbf{A}$ whose rows can be partitioned into two disjoint sets $\mathcal{C}$ and $\mathcal{D}$, the following four conditions together are sufficient for $\mathbf{A}$ to be totally unimodular (Heller and Tompkins, 1957):

1. Every entry in $\mathbf{A}$ is $0, +1$, or -1.
2. Every column of $\mathbf{A}$ contains at most two non-zero entries.
3. If two non-zero entries in a column of $\mathbf{A}$ have the same sign, then the row of one is in $\mathcal{C}$, and the other in $\mathcal{D}$.
4. If two non-zero entries in a column of $\mathbf{A}$ have opposite signs, then the rows of both are in $\mathcal{C}$, or both in $\mathcal{D}$.

In the setting of resource allocation, we can write problem (2.1) equivalently as problem (A.1) where $\mathbf{x} = \mathrm{vec}(\mathbf{X})$, $\mathbf{t} = \mathrm{vec}(\mathbf{\Theta}^*)$, $\mathbf{A}$ and $\mathbf{b}$ are given as

$$
\mathbf{A} = \begin{bmatrix} \mathbb{1}_N^{\mathrm{T}} \otimes \mathbf{I}_M \\ \mathbf{I}_N \otimes \mathbb{1}_M^{\mathrm{T}} \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} \mathbf{c}_t \\ \mathbf{d}_t \end{bmatrix}.
\tag{A.3}
$$

For matrix $\mathbf{A}$ given in (A.3), we can set $\mathcal{C}$ to be the set of first M rows corresponding to the capacity constraints, and $\mathcal{D}$ to be the set of remaining rows corresponding to the demand constraints. Since this $\mathbf{A}$ matrix satisfies the conditions of the proposition for sets $\mathcal{C}$ and $\mathcal{D}$, we obtain that $\mathbf{A}$ is totally unimodular. Finally, since the vector $\mathbf{b}$ is integral and the matrix $\mathbf{A}$ is totally unimodular, the integrality gap is one.

B   DEFINITIONS

We start with definitions that we will use throughout the proofs.

Definition B.1. *For a symmetric positive definite matrix* $\mathbf{W} \in \mathbb{R}^{d \times d}$, *we define*

- $\mathbf{W}$-*inner product of two vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ *as* $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{W}} := \langle \mathbf{W}\mathbf{x}, \mathbf{y} \rangle$,
- $\mathbf{W}$-*norm of a vector* $\mathbf{x} \in \mathbb{R}^d$ *as* $\|\mathbf{x}\|_{\mathbf{W}} := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{W}}}$.

Definition B.2. $\mathbf{A}_t \in \mathbb{R}^{NM \times NM}$ *is a diagonal matrix with* $(N(i-1)+u)$*th diagonal entry equal to* $n_{ui}^t + \gamma$. *Recall that* $n_{ui}^t = \sum_{\tau=1}^{t-1} \mathbb{1}\{(u,i) \in \mathcal{A}_t\}$ *denotes the number of times pair* (u,i) *has been sampled before time* t *(excluding time* t).

Then, the regularized empirical L_2-norm of a matrix $\mathbf{Z} \in \mathbb{R}^{N \times M}$ can be written as

$$
\|\mathbf{Z}\|_{E_2^t} = \|\mathrm{vec}(\mathbf{Z})\|_{\mathbf{A}_t}.
\tag{B.1}
$$

For the ease of exposition in the following sections, we will use the shorthand $\|\mathbf{Z}\|_{\mathbf{A}_t} = \|\mathrm{vec}(\mathbf{Z})\|_{\mathbf{A}_t}$.

Definition B.3. *Define* $\mathbf{E}_{ui} := \mathbf{e}_u \mathbf{e}_i^T$ *to be the indicator matrix for* (u, i) *and* $\mathbf{e}_{ui} := \text{vec}(\mathbf{E}_{ui})$.

Definition B.4. *Define the "confidence width" of the allocations as*

$$w_t := \|\mathbf{X}_t\|_{\mathbf{A}_t^{-1}} \qquad and \qquad w_{ui}^t := \|\mathbf{E}_{ui}\|_{\mathbf{A}_t^{-1}}. \tag{B.2}$$

Definition B.5. *Define the regularized empirical* L_2-*norm* $\|\cdot\|_{E_2^t}$ *as*

$$\|\mathbf{\Delta}\|_{E_2^t}^2 := \sum_{\tau=1}^{t-1} \sum_{(u,i)\in\mathcal{A}_\tau} \langle \mathbf{\Delta}, \mathbf{E}_{ui} \rangle^2 + \gamma\|\mathbf{\Delta}\|_{\text{F}}^2 = \sum_{u\in\mathcal{N}} \sum_{i\in\mathcal{I}} (n_{ui}^t + \gamma)(\Delta_{ui})^2.$$

For future reference, we also define the (non-regularized) empirical L_2-*norm* $\|\cdot\|_{\widetilde{E}_2^t}$ *by*

$$\|\mathbf{\Delta}\|_{\widetilde{E}_2^t}^2 := \sum_{\tau=1}^{t-1} \sum_{(u,i)\in\mathcal{A}_\tau} \langle \mathbf{\Delta}, \mathbf{E}_{ui} \rangle^2 = \sum_{i=1}^{d} (n_{ui}^t)(\Delta_{ui})^2.$$

Note that the regularized empirical L_2-norm is related to (non-regularized) empirical L_2-norm as

$$\|\mathbf{\Delta}\|_{E_2^t}^2 = \|\mathbf{\Delta}\|_{\widetilde{E}_2^t}^2 + \gamma\|\mathbf{\Delta}\|_{\text{F}}^2.$$

Definition B.6. *Define the social welfare regret* r_t *at time* t *as*

$$r_t := \langle \mathbf{X}_t^*, \mathbf{\Theta}^* \rangle - \langle \mathbf{X}_t, \mathbf{\Theta}^* \rangle.$$

Definition B.7. *The OFU estimate of the allocation and the parameters are*

$$(\mathbf{X}_t, \mathbf{\Theta}_t) \in \underset{(\mathbf{X},\mathbf{\Theta})\in\mathcal{X}_t\times\mathcal{C}_t}{\arg\max} \langle \mathbf{X}, \mathbf{\Theta} \rangle.$$

The entries of $\mathbf{X}_t$ *are denoted by* X_{ui}^t *and the entries of* $\mathbf{\Theta}_t$ *are denoted by* Θ_{ui}^t.

Definition B.8. *We define* $\mathbf{\Delta}_t = \mathbf{\Theta}_t - \mathbf{\Theta}^*$ *with entries* $\Delta_{ui}^t = \Theta_{ui}^t - \Theta_{ui}^*$.

C PROOFS FOR LOW-RANK ALLOCATION AND PRICING

C.1 Construction of Confidence Sets

Lemma C.1. *For any* $\delta > 0$ *and* $\mathbf{\Theta} \in \mathbb{R}^d$,

$$\mathbb{P}\left(L_{2,t}(\mathbf{\Theta}) \geq L_{2,t}(\mathbf{\Theta}^*) + \frac{1}{2}\|\mathbf{\Theta}^* - \mathbf{\Theta}\|_{\widetilde{E}_2^t}^2 - 4\eta^2 \log(1/\delta) \quad, \forall t \in \mathbb{N}\right) \geq 1 - \delta. \tag{C.1}$$

Proof. Let $\mathcal{H}_{t-1}$ be the σ-algebra generated by $(H_t, \mathcal{A}_t)$ and let $\mathcal{H}_0 = \sigma(\emptyset, \Omega)$. Then, define $\epsilon_{ui}^t := R_{ui}^t - \Theta_{ui}^*$ for all $t \in \mathbb{N}$ and $(u, i) \in \mathcal{A}_t$. By previous assumptions, $\mathbb{E}[\epsilon_{ui}^t | \mathcal{H}_{t-1}] = 0$ and $\mathbb{E}[\exp(\lambda\epsilon_{ui}^t)|\mathcal{H}_{t-1}] \leq \exp\left(\frac{\lambda^2\eta^2}{2}\right)$ for all t.

Define $Z_{ui}^t := (R_{ui}^t - \Theta_{ui}^*)^2 - (R_{ui}^t - \Theta_{ui})^2$. Then, we have

$$Z_{ui}^t = -(\Theta_{ui} - \Theta_{ui}^*)^2 + 2\epsilon_{ui}^t(\Theta_{ui} - \Theta_{ui}^*).$$

Therefore, the conditional mean and conditional cumulant generating function satisfy

$$\mu_{ui}^t := \mathbb{E}[Z_{ui}^t | \mathcal{H}_{t-1}] = -(\Theta_{ui} - \Theta_{ui}^*)^2$$
$$\psi_{ui}^t(\lambda) := \log \mathbb{E}[\exp(\lambda[Z_{ui}^t - \mu_{ui}^t])|\mathcal{H}_{t-1}]$$
$$= \log \mathbb{E}[\exp(2\lambda(\Theta_{ui} - \Theta_{ui}^*)\epsilon_{ui}^t)|\mathcal{H}_{t-1}]$$
$$\leq \frac{(2\lambda(\Theta_{ui} - \Theta_{ui}^*))^2\eta^2}{2}.$$

Then, define $Z^t := \sum_{(u,i) \in \mathcal{A}_t} Z_{ui}^t$. Since each Z_{ui}^t term is conditionally independent, the conditional mean and conditional cumulant generating functions of Z^t satisfy

$$\mu^t := \mathbb{E}[Z^t | \mathcal{H}_{t-1}] = - \sum_{(u,i) \in \mathcal{A}_t} (\Theta_{ui} - \Theta_{ui}^*)^2$$

$$\psi^t(\lambda) := \sum_{(u,i) \in \mathcal{A}_t} \psi_{ui}^t(\lambda)$$

$$\leq \sum_{(u,i) \in \mathcal{A}_t} \frac{(2\lambda(\Theta_{ui} - \Theta_{ui}^*))^2 \eta^2}{2}.$$

Applying Lemma E.2 to $(Z^t)_{t \in \mathbb{N}}$ shows that for all $x \geq 0$ and $\lambda \geq 0$,

$$\mathbb{P}\left(\sum_{\tau=1}^{t-1} Z^\tau \leq \frac{x}{\lambda} + \sum_{\tau=1}^{t-1} \sum_{(u,i) \in \mathcal{A}_\tau} (\Theta_{ui} - \Theta_{ui}^*)^2 (2\lambda\eta^2 - 1) \quad , \forall t \in \mathbb{N} \right) \geq 1 - e^{-x}.$$

Note that we have $\sum_{\tau=1}^{t-1} Z^\tau = L_{2,t}(\boldsymbol{\Theta}^*) - L_{2,t}(\boldsymbol{\Theta})$, and $\sum_{\tau=1}^{t-1} \sum_{(u,i) \in \mathcal{A}_\tau} (\Theta_{ui} - \Theta_{ui}^*)^2 = \|\boldsymbol{\Theta}^* - \boldsymbol{\Theta}\|_{\widetilde{E}_2^t}^2$.

Then, choosing $\lambda = \frac{1}{4\eta^2}$ and $x = \log \frac{1}{\delta}$ gives

$$\mathbb{P}\left(L_{2,t}(\boldsymbol{\Theta}) \geq L_{2,t}(\boldsymbol{\Theta}^*) + \frac{1}{2} \|\boldsymbol{\Theta}^* - \boldsymbol{\Theta}\|_{\widetilde{E}_2^t}^2 - 4\eta^2 \log(1/\delta) \quad , \forall t \in \mathbb{N} \right) \geq 1 - \delta.$$

$\square$

Lemma C.2. *If $\boldsymbol{\Theta}, \boldsymbol{\Theta}^\alpha \in \mathcal{L}$ satisfy $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^\alpha\|_F \leq \alpha$, then with probability at least $1 - \delta$,*

$$\left| \frac{1}{2} \|\boldsymbol{\Theta}^* - \boldsymbol{\Theta}^\alpha\|_{\widetilde{E}_2^t}^2 - \frac{1}{2} \|\boldsymbol{\Theta}^* - \boldsymbol{\Theta}\|_{\widetilde{E}_2^t}^2 + L_{2,t}(\boldsymbol{\Theta}) - L_{2,t}(\boldsymbol{\Theta}^\alpha) \right| \leq \alpha t \sqrt{NM} \left[8 + 2\sqrt{2\eta^2 \log(4NMt^2/\delta)} \right]. \tag{C.2}$$

Proof. Let $\mathbf{E}_{ui} = \mathbf{e}_u \mathbf{e}_i^\mathsf{T}$ be the indicator matrix for (u,i). Since any two $\boldsymbol{\Theta}, \boldsymbol{\Theta}^\alpha \in \mathcal{L}$ satisfy $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^\alpha\|_F \leq \sqrt{NM}$, it is enough to consider $\alpha \leq \sqrt{NM}$. We find

$$\sum_{u=1}^{N} \sum_{i=1}^{M} |\langle \boldsymbol{\Theta}, \mathbf{E}_{ui} \rangle^2 - \langle \boldsymbol{\Theta}^\alpha, \mathbf{E}_{ui} \rangle^2| \leq \max_{\|\boldsymbol{\Delta}\|_F \leq \alpha} \left\{ \sum_{u=1}^{N} \sum_{i=1}^{M} |\Theta_{ui}^2 - (\Theta_{ui} + \Delta_{ui})^2| \right\}$$

$$= \max_{\|\boldsymbol{\Delta}\|_F \leq \alpha} \left\{ \sum_{u=1}^{N} \sum_{i=1}^{M} |2\Theta_{ui}\Delta_{ui} + \Delta_{ui}^2| \right\}$$

$$\leq \max_{\|\boldsymbol{\Delta}\|_F \leq \alpha} \left\{ 2 \sum_{u=1}^{N} \sum_{i=1}^{M} |\Theta_{ui}\Delta_{ui}| + \sum_{u=1}^{N} \sum_{i=1}^{M} \Delta_{ui}^2 \right\}$$

$$\leq \max_{\|\boldsymbol{\Delta}\|_F \leq \alpha} \left\{ 2\|\boldsymbol{\Delta}\|_{1,1} + \|\boldsymbol{\Delta}\|_F^2 \right\}$$

$$\leq 2\alpha\sqrt{NM} + \alpha^2.$$

Therefore, it implies

$$\sum_{u=1}^{N} \sum_{i=1}^{M} |\langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^*, \mathbf{E}_{ui} \rangle^2 - \langle \boldsymbol{\Theta}^\alpha - \boldsymbol{\Theta}^*, \mathbf{E}_{ui} \rangle^2| = \sum_{u=1}^{N} \sum_{i=1}^{M} |\langle \boldsymbol{\Theta}, \mathbf{E}_{ui} \rangle^2 - \langle \boldsymbol{\Theta}^\alpha, \mathbf{E}_{ui} \rangle^2 + 2\langle \boldsymbol{\Theta}^*, \mathbf{E}_{ui} \rangle\langle \boldsymbol{\Theta}^\alpha - \boldsymbol{\Theta}, \mathbf{E}_{ui} \rangle|$$

$$\leq 2\alpha\sqrt{NM} + \alpha^2 + 2\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^\alpha\|_{1,1}$$

$$\leq 4\alpha\sqrt{NM} + \alpha^2.$$

Similarly, for any t, we have

$$
\sum_{u=1}^{N} \sum_{i=1}^{M} \left| \left(R_{ui}^t - \langle \mathbf{\Theta}, \mathbf{E}_{ui} \rangle \right)^2 - \left(R_{ui}^t - \langle \mathbf{\Theta}^\alpha, \mathbf{E}_{ui} \rangle \right)^2 \right| = \sum_{u=1}^{N} \sum_{i=1}^{M} \left| 2 R_{ui}^t \langle \mathbf{\Theta}^\alpha - \mathbf{\Theta}, \mathbf{E}_{ui} \rangle + \langle \mathbf{\Theta}, \mathbf{E}_{ui} \rangle^2 - \langle \mathbf{\Theta}^\alpha, \mathbf{E}_{ui} \rangle^2 \right|
$$

$$
\leq 2 \sum_{u=1}^{N} \sum_{i=1}^{M} \left| R_{ui}^t \right| \left| \langle \mathbf{\Theta}^\alpha - \mathbf{\Theta}, \mathbf{E}_{ui} \rangle \right| + 2\alpha \sqrt{NM} + \alpha^2
$$

$$
\leq 2 \| \mathbf{\Theta}^\alpha - \mathbf{\Theta} \|_{\mathrm{F}} \left(\sum_{u=1}^{N} \sum_{i=1}^{M} |R_{ui}^t|^2 \right)^{1/2} + 2\alpha \sqrt{NM} + \alpha^2
$$

$$
\leq 2\alpha \left(\sum_{u=1}^{N} \sum_{i=1}^{M} |R_{ui}^t|^2 \right)^{1/2} + 2\alpha \sqrt{NM} + \alpha^2.
$$

Summing over t and noting that $\mathcal{A}_t \subseteq [N] \times [M]$, the left hand side of (C.2) is bounded by

$$
\sum_{\tau=1}^{t-1} \left(\frac{1}{2} \left[4\alpha \sqrt{NM} + \alpha^2 \right] + 2\alpha \left(\sum_{u=1}^{N} \sum_{i=1}^{M} |R_{ui}^t|^2 \right)^{1/2} + 2\alpha \sqrt{NM} + \alpha^2 \right) \leq \alpha \sum_{\tau=1}^{t-1} \left(6\sqrt{NM} + 2 \left(\sum_{u=1}^{N} \sum_{i=1}^{M} |R_{ui}^t|^2 \right)^{1/2} \right).
$$

Because ϵ_{ui}^τ is η-sub-Gaussian, $\mathbb{P} \left(|\epsilon_{ui}^\tau| > \sqrt{2\eta^2 \log(2/\delta)} \right) \leq \delta$. By a union bound, $\mathbb{P} \left(\exists \tau \in \mathbb{N}, u \in [N], i \in [M] \text{ s.t. } |\epsilon_{ui}^\tau| > \sqrt{2\eta^2 \log(4NM\tau^2/\delta)} \right) \leq \frac{\delta NM}{2} \sum_{\tau=1}^{\infty} \frac{1}{NM\tau^2} \leq \delta$. Since $|R_{ui}^\tau| \leq 1 + |\epsilon_{ui}^\tau|$, we have $|R_{ui}^\tau| \leq 1 + \sqrt{2\eta^2 \log(4NM\tau^2/\delta)}$ with probability at least $1 - \delta$. Consequently, the bound for the discretization error becomes

$$
\alpha t \sqrt{NM} \left[8 + 2\sqrt{2\eta^2 \log(4NMt^2/\delta)} \right].
$$

$\square$

Lemma C.3. *For any $\delta > 0$, $\alpha > 0$ and $\gamma > 0$, if*

$$
\mathcal{Q}_t = \left\{ \mathbf{\Theta} \in \mathcal{F} : \| \mathbf{\Theta} - \widehat{\mathbf{\Theta}}_t \|_{E_2^t} \leq \sqrt{\beta_t^*(\delta, \alpha, \gamma)} \right\} \tag{C.3}
$$

for all $t \in \mathbb{N}$, then

$$
\mathbb{P} \left(\mathbf{\Theta}^* \in \mathcal{Q}_t \quad , \forall t \in \mathbb{N} \right) \geq 1 - 2\delta. \tag{C.4}
$$

Proof. Let $\mathcal{F}^\alpha \subset \mathcal{F}$ be an α-cover of $\mathcal{F}$ in the Frobenius norm so that for any $\mathbf{\Theta} \in \mathcal{F}$, there exists $\mathbf{\Theta}^\alpha \in \mathcal{F}^\alpha$ such that $\| \mathbf{\Theta} - \mathbf{\Theta}^\alpha \|_{\mathrm{F}} \leq \alpha$. By a union bound applied to Lemma C.1, with probability at least $1 - \delta$,

$$
L_{2,t}(\mathbf{\Theta}^\alpha) - L_{2,t}(\mathbf{\Theta}^*) \geq \frac{1}{2} \| \mathbf{\Theta}^* - \mathbf{\Theta}^\alpha \|_{\widetilde{E}_2^t}^2 - 4\eta^2 \log(|\mathcal{F}^\alpha|/\delta) \quad , \forall \mathbf{\Theta}^\alpha \in \mathcal{F}^\alpha, t \in \mathbb{N}.
$$

Therefore, with probability at least $1 - \delta$, for all $\mathbf{\Theta} \in \mathcal{F}, t \in \mathbb{N}$,

$$
L_{2,t}(\mathbf{\Theta}) - L_{2,t}(\mathbf{\Theta}^*) \geq \frac{1}{2} \| \mathbf{\Theta}^* - \mathbf{\Theta} \|_{\widetilde{E}_2^t}^2 - 4\eta^2 \log(|\mathcal{F}^\alpha|/\delta)
$$
$$
+ \min_{\mathbf{\Theta}^\alpha \in \mathcal{F}^\alpha} \left\{ \frac{1}{2} \| \mathbf{\Theta}^* - \mathbf{\Theta}^\alpha \|_{\widetilde{E}_2^t}^2 - \frac{1}{2} \| \mathbf{\Theta}^* - \mathbf{\Theta} \|_{\widetilde{E}_2^t}^2 + L_{2,t}(\mathbf{\Theta}) - L_{2,t}(\mathbf{\Theta}^\alpha) \right\}.
$$

By Lemma C.2, with probability at least $1 - 2\delta$,

$$
L_{2,t}(\mathbf{\Theta}) - L_{2,t}(\mathbf{\Theta}^*) \geq \frac{1}{2} \| \mathbf{\Theta}^* - \mathbf{\Theta} \|_{\widetilde{E}_2^t}^2 - D_t,
$$

where $D_t := 4\eta^2 \log(|\mathcal{F}^\alpha|/\delta) + \alpha t \sqrt{NM} \left[8 + 2\sqrt{2\eta^2 \log(4NMt^2/\delta)} \right]$.

Adding the regularization terms to both sides, we obtain

$$L_{2,t}(\boldsymbol{\Theta}) + \gamma\|\boldsymbol{\Theta} - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2 - L_{2,t}(\boldsymbol{\Theta}^*) - \gamma\|\boldsymbol{\Theta}^* - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2 \geq \frac{1}{2}\|\boldsymbol{\Theta}^* - \boldsymbol{\Theta}\|_{\widehat{E}_2^t}^2 + \gamma\|\boldsymbol{\Theta} - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2 - D_t - \gamma\|\boldsymbol{\Theta}^* - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2.$$

Note the definition of the least square estimate $\widehat{\boldsymbol{\Theta}}_t = \arg\min_{\boldsymbol{\Theta} \in \mathcal{F}}\{L_{2,t}(\boldsymbol{\Theta}) + \gamma\|\boldsymbol{\Theta} - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2\}$. By letting $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}_t$, the left hand side becomes non-positive, and hence

$$\frac{1}{2}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|_{\widehat{E}_2^t}^2 \leq D_t + \gamma\left(\|\boldsymbol{\Theta}^* - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2 - \|\widehat{\boldsymbol{\Theta}}_t - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2\right).$$

Then,

$$\frac{1}{2}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|_{\widehat{E}_2^t}^2 + \gamma\left(\|\widehat{\boldsymbol{\Theta}}_t - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2 + \|\boldsymbol{\Theta}^* - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2\right) \leq D_t + 2\gamma\|\boldsymbol{\Theta}^* - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2.$$

By triangle inequality we have $\|\widehat{\boldsymbol{\Theta}}_t - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}} + \|\boldsymbol{\Theta}^* - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}} \geq \|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|_{\mathrm{F}}$. Taking squares on both sides, we obtain $\|\widehat{\boldsymbol{\Theta}}_t - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2 + \|\boldsymbol{\Theta}^* - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2 \geq \frac{1}{2}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|_{\mathrm{F}}^2$. Then, noting that $\|\boldsymbol{\Delta}\|_{2,E_t^2}^2 = \|\boldsymbol{\Delta}\|_{\widehat{E}_2^t}^2 + \gamma\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2$, we have

$$\frac{1}{2}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|_{E_2^t}^2 \leq D_t + 2\gamma\|\boldsymbol{\Theta}^* - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2.$$

Lastly, using the inequality $\|\boldsymbol{\Theta}^* - \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2 \leq G^2$,

$$\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|_{E_2^t}^2 \leq 8\eta^2 \log(|\mathcal{F}^\alpha|/\delta) + 2\alpha t\sqrt{NM}\left[8 + 2\sqrt{2\eta^2\log(4NMt^2/\delta)}\right] + 4\gamma G^2.$$

Taking the infimum over the size of α-covers, we obtain the final result.

$\square$

C.2 Regret Bounds

Throughout this section we will use the shorthand $\beta_t = \beta_t^*(\delta, \alpha, \gamma)$ for the parameter defined as

$$\beta_t^*(\delta, \alpha, \gamma) := 8\eta^2 \log\left(\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_2)/\delta\right) + 2\alpha t\sqrt{NM}\left(8 + \sqrt{8\eta^2\log(4NMt^2/\delta)}\right) + 4\gamma G^2. \tag{C.5}$$

Lemma C.4. *For any* $\mathbf{X}_t \in \mathcal{X}_t$ *and* $\boldsymbol{\Theta} \in \mathcal{Q}_t$, *we have*

$$|\langle \boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}^t, \mathbf{X}\rangle| \leq w_t\sqrt{\beta_t}. \tag{C.6}$$

Proof. Let $\boldsymbol{\Delta} = \boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}^t$. Then,

$$\begin{aligned}
|\langle \boldsymbol{\Delta}, \mathbf{X}_t\rangle| &= |\mathrm{vec}(\boldsymbol{\Delta})^{\mathsf{T}}\mathrm{vec}(\mathbf{X}_t)| \\
&= |\mathrm{vec}(\boldsymbol{\Delta})^{\mathsf{T}}\mathbf{A}_t^{1/2}\mathbf{A}_t^{-1/2}\mathrm{vec}(\mathbf{X}_t)| \\
&= |(\mathbf{A}_t^{1/2}\mathrm{vec}(\boldsymbol{\Delta}))^{\mathsf{T}}\mathbf{A}_t^{-1/2}\mathrm{vec}(\mathbf{X}_t)| \\
&\leq \|\mathbf{A}_t^{1/2}\mathrm{vec}(\boldsymbol{\Delta})\|\|\mathbf{A}_t^{-1/2}\mathrm{vec}(\mathbf{X}_t)\| \\
&= \|\mathrm{vec}(\boldsymbol{\Delta})\|_{\mathbf{A}_t}\|\mathrm{vec}(\mathbf{X}_t)\|_{\mathbf{A}_t^{-1}} \\
&= \|\boldsymbol{\Delta}\|_{\mathbf{A}_t}\|\mathbf{X}_t\|_{\mathbf{A}_t^{-1}} \\
&\leq w_t\sqrt{\beta_t}.
\end{aligned}$$

$\square$

Lemma C.5. *For any $t \in \mathbb{N}$, we have the identity*

$$w_t^2 = \sum_{(u,i) \in \mathcal{A}_t} (w_{ui}^t)^2.$$

Proof.

$$\begin{aligned}
w_t^2 &= \langle \text{vec}(\mathbf{X}_t), \text{vec}(\mathbf{X}_t) \rangle_{\mathbf{A}_t^{-1}} \\
&= \left\langle \mathbf{A}_t^{-1} \sum_{(u,i) \in \mathcal{A}_t} \mathbf{e}_{ui}, \sum_{(u,i) \in \mathcal{A}_t} \mathbf{e}_{ui} \right\rangle \\
&= \sum_{(u,i) \in \mathcal{A}_t} \sum_{(r,j) \in \mathcal{A}_t} \left\langle \mathbf{A}_t^{-1} \mathbf{e}_{ui}, \mathbf{e}_{rj} \right\rangle \\
&= \sum_{(u,i) \in \mathcal{A}_t} \left\langle \mathbf{A}_t^{-1} \mathbf{e}_{ui}, \mathbf{e}_{ui} \right\rangle \\
&= \sum_{(u,i) \in \mathcal{A}_t} (w_{ui}^t)^2,
\end{aligned}$$

where the penultimate step follows because $\left\langle \mathbf{A}_t^{-1} \mathbf{e}_{ui}, \mathbf{e}_{rj} \right\rangle = 0$ for $(u,i) \neq (r,j)$. $\qquad \square$

Lemma C.6. *If $\mathbf{\Theta}^* \in \mathcal{Q}_t$, then*

$$r_t \leq \langle \mathbf{X}_t, \mathbf{\Theta}_t - \mathbf{\Theta}^* \rangle \leq 2 w_t \sqrt{\beta_t}.$$

Proof. By the choice of $(\mathbf{X}_t, \mathbf{\Theta}_t)$, we have

$$\langle \mathbf{X}_t, \mathbf{\Theta}_t \rangle = \max_{(\mathbf{X}, \mathbf{\Theta}) \in \mathcal{X}_t \times \mathcal{Q}_t} \langle \mathbf{X}, \mathbf{\Theta} \rangle \geq \langle \mathbf{X}_t^*, \mathbf{\Theta}^* \rangle,$$

where the inequality uses $\mathbf{\Theta}^* \in \mathcal{Q}_t$. Hence,

$$\begin{aligned}
r_t &= \langle \mathbf{X}_t^*, \mathbf{\Theta}^* \rangle - \langle \mathbf{X}_t, \mathbf{\Theta}^* \rangle \\
&\leq \langle \mathbf{X}_t, \mathbf{\Theta}_t - \mathbf{\Theta}^* \rangle \\
&= \langle \mathbf{X}_t, \mathbf{\Theta}_t - \widehat{\mathbf{\Theta}}^t \rangle + \langle \mathbf{X}_t, \widehat{\mathbf{\Theta}}^t - \mathbf{\Theta}^* \rangle \\
&\leq 2 w_t \sqrt{\beta_t},
\end{aligned}$$

where the last step follows from Lemma $C.4$. $\qquad \square$

Next, we show that the confidence widths do not grow too fast.

Lemma C.7. *For every t,*

$$\log \det \mathbf{A}_{t+1} = NM \log \gamma + \sum_{\tau=1}^{t} \sum_{(u,i) \in \mathcal{A}_\tau} \log(1 + (w_{ui}^\tau)^2). \tag{C.7}$$

Proof. By the definition of $\mathbf{A}_t$, we have

$$\begin{aligned}
\det \mathbf{A}_{t+1} &= \det \left(\mathbf{A}_t + \sum_{(u,i) \in \mathcal{A}_t} \mathbf{e}_{ui} \mathbf{e}_{ui}^{\mathsf{T}} \right) \\
&= \det \left(\mathbf{A}_t^{1/2} \left(\mathbf{I} + \mathbf{A}_t^{-1/2} \left(\sum_{(u,i) \in \mathcal{A}_t} \mathbf{e}_{ui} \mathbf{e}_{ui}^{\mathsf{T}} \right) \mathbf{A}_t^{-1/2} \right) \mathbf{A}_t^{1/2} \right) \\
&= \det(\mathbf{A}_t) \det \left(\mathbf{I} + \sum_{(u,i) \in \mathcal{A}_t} \mathbf{A}_t^{-1/2} \mathbf{e}_{ui} \mathbf{e}_{ui}^{\mathsf{T}} \mathbf{A}_t^{-1/2} \right).
\end{aligned}$$

Each $\mathbf{A}_t^{-1/2} \mathbf{e}_{ui} \mathbf{e}_{ui}^{\mathsf{T}} \mathbf{A}_t^{-1/2}$ term has zeros everywhere except one entry on the diagonal and that non-zero entry is equal to $(w_{ui}^t)^2$. Furthermore, the location of the non-zero entry is different in for each term. Hence,

$$\det \left(\mathbf{I} + \sum_{(u,i) \in \mathcal{A}_t} \mathbf{A}_t^{-1/2} \mathbf{e}_{ui} \mathbf{e}_{ui}^{\mathsf{T}} \mathbf{A}_t^{-1/2} \right) = \prod_{(u,i) \in \mathcal{A}_t} (1 + (w_{ui}^t)^2).$$

Therefore, we have

$$\log \det \mathbf{A}_{t+1} = \log \det \mathbf{A}_t + \sum_{(u,i) \in \mathcal{A}_t} \log(1 + (w_{ui}^t)^2).$$

Since $\mathbf{A}_1 = \gamma \mathbf{I}$, we have $\log \det \mathbf{A}_1 = NM \log \gamma$ and the result follows by induction. □

Lemma C.8. *For all $t \in \mathbb{N}$, $\log \det \mathbf{A}_t \le NM \log(t + \gamma - 1)$.*

Proof. Noting that $\mathbf{A}_t$ is a diagonal matrix with diagonals $(n_{ui}^t + \gamma)$,

$$\text{trace } \mathbf{A}_t = NM\gamma + \sum_{u=1}^{N} \sum_{i=1}^{M} n_{ui}^t$$
$$\le NM\gamma + NM(t - 1)$$
$$= NM(t + \gamma - 1).$$

Now, recall that $\text{trace } \mathbf{A}_t$ equals the sum of the eigenvalues of $\mathbf{A}_t$. On the other hand, $\det(\mathbf{A}_t)$ equals the product of the eigenvalues. Since $\mathbf{A}_t$ is positive definite, its eigenvalues are all positive. Subject to these constraints, $\det(\mathbf{A}_t)$ is maximized when all the eigenvalues are equal; the desired bound follows. □

Lemma C.9. *Let $\gamma \ge 1$. Then, for all t, we have*

$$\sum_{\tau=1}^{t} w_\tau^2 \le 2NM \log \left(1 + \frac{t}{\gamma} \right).$$

Proof. Note that $0 \le (w_{ui}^\tau)^2 \le 1$, if $\gamma \ge 1$. Using the inequality $y \le 2 \log(1 + y)$ for $0 \le y \le 1$, we have

$$\sum_{\tau=1}^{t} w_\tau^2 = \sum_{\tau=1}^{t} \sum_{(u,i) \in \mathcal{A}_\tau} (w_{ui}^\tau)^2$$
$$\le 2 \sum_{\tau=1}^{t} \sum_{(u,i) \in \mathcal{A}_\tau} \log(1 + (w_{ui}^\tau)^2)$$
$$= 2 \log \det \mathbf{A}_{t+1} - 2NM \log \gamma$$
$$\le 2NM \log \left(1 + \frac{t}{\gamma} \right),$$

where the last two lines follow from Lemmas C.7 and C.8 respectively. □

Theorem C.10. *If $\gamma \ge 1$ and $\mathbf{\Theta}^* \in \mathcal{Q}_t$ for all $t \le T$, then the T period social welfare regret is bounded by*

$$\mathcal{R}^{SW}(T, \pi) \le \sqrt{8NM\beta_T^*(\delta, \alpha, \gamma)T \log \left(1 + \frac{T}{\gamma} \right)}.$$

Proof. Recall the definition of instantaneous regret $r_t = \langle \mathbf{X}_t^*, \mathbf{\Theta}^* \rangle - \langle \mathbf{X}_t, \mathbf{\Theta}^* \rangle$. Assuming that $\mathbf{\Theta}^* \in \mathcal{Q}_t$ for all $t \leq T$, we have

$$
\mathcal{R}^{SW}(T, \pi) = \sum_{t=1}^{T} r_t
$$

$$
\leq \left(T \sum_{t=1}^{T} r_t^2 \right)^{1/2}
$$

$$
\leq \left(4\beta_T T \sum_{t=1}^{T} w_t^2 \right)^{1/2},
$$

where the last step follows from Lemma C.6. Then, using Lemma C.9, we prove the result. $\qquad\square$

In order to prove our instability bounds, we start with the following lemma.

Lemma C.11. *If $\mathbf{\Theta}^* \in \mathcal{Q}_t$, we have*

$$
\sum_{u=1}^{N} \left\{ \max_{\mathbf{x}_u \in \mathcal{X}_u^t} \sum_{i=1}^{M} x_{ui}(\Theta_{ui}^* - \bar{p}_i^t) \right\} \leq \sum_{u=1}^{N} \sum_{i=1}^{M} x_{ui}^t(\Theta_{ui}^t - \bar{p}_i^t)
$$

, where $\mathcal{X}_u^t = \{x \in \mathbb{R}_+^M : \mathbb{1}^T x \leq d_u^t\}$.

Proof. From the definition of $\mathbf{\Theta}_t$ and $\mathbf{X}_t$, we have

$$
\sum_{u=1}^{N} \sum_{i=1}^{M} x_{ui}^t \Theta_{ui}^t = \langle \mathbf{X}_t, \mathbf{\Theta}_t \rangle
$$

$$
= \max_{(\mathbf{X}, \mathbf{\Theta}) \in \mathcal{X}_t \times \mathcal{Q}_t} \langle \mathbf{X}, \mathbf{\Theta} \rangle
$$

$$
= \max_{\mathbf{\Theta} \in \mathcal{Q}_t} \max_{\mathbf{X} \in \mathcal{X}_t} \langle \mathbf{X}, \mathbf{\Theta} \rangle
$$

$$
= \max_{\mathbf{\Theta} \in \mathcal{Q}_t} \max_{\mathbf{X} \in \mathbb{R}_+^{N \times M}} \min_{\boldsymbol{\lambda} \in \mathbb{R}_+^M} \langle \mathbf{X}, \mathbf{\Theta} \rangle + \boldsymbol{\lambda}^T (\mathbf{c}_t - \mathbf{X}^T \mathbb{1})
$$
$$
\text{s.t. } \mathbf{X}\mathbb{1} \leq \mathbf{d}^t
$$

$$
= \max_{\mathbf{\Theta} \in \mathcal{Q}_t} \max_{\mathbf{X} \in \mathbb{R}_+^{N \times M}} \min_{\boldsymbol{\lambda} \in \mathbb{R}_+^M} \langle \mathbf{X}, \mathbf{\Theta} - \mathbb{1}\boldsymbol{\lambda}^T \rangle + \boldsymbol{\lambda}^T \mathbf{c}_t
$$
$$
\text{s.t. } \mathbf{X}\mathbb{1} \leq \mathbf{d}^t
$$

$$
= \max_{\mathbf{\Theta} \in \mathcal{Q}_t} \max_{\mathbf{X} \in \mathbb{R}_+^{N \times M}} \min_{\boldsymbol{\lambda} \in \mathbb{R}_+^M} \sum_{u=1}^{N} \sum_{i=1}^{M} x_{ui}(\Theta_{ui} - \lambda_i) + \sum_{i=1}^{M} \lambda_i c_i^t.
$$
$$
\text{s.t. } \mathbf{X}\mathbb{1} \leq \mathbf{d}^t
$$

Note that the optimum value of this optimization problem is achieved when $\boldsymbol{\lambda} = \bar{\mathbf{p}}_t$, $\mathbf{\Theta} = \mathbf{\Theta}_t$ and $\mathbf{X} = \mathbf{X}_t$. Therefore, if we fix $\boldsymbol{\lambda} = \bar{\mathbf{p}}_t$ and optimize over $\mathbf{\Theta}$ and $\mathbf{X}$, the optimum value of the objective does not change. Consequently, we can write

$$
\sum_{u=1}^{N} \sum_{i=1}^{M} x_{ui}^t \Theta_{ui}^t = \max_{\mathbf{\Theta} \in \mathcal{Q}_t} \max_{\mathbf{X} \in \mathbb{R}_+^{N \times M}} \sum_{u=1}^{N} \sum_{i=1}^{M} x_{ui}(\Theta_{ui} - \bar{p}_i^t) + \sum_{i=1}^{M} \bar{p}_i^t c_i^t
$$
$$
\text{s.t. } \mathbf{X}\mathbb{1} \leq \mathbf{d}^t
$$

$$
= \max_{\mathbf{\Theta} \in \mathcal{Q}_t} \sum_{u=1}^{N} \left\{ \max_{\mathbf{x}_u \in \mathcal{X}_u^t} \sum_{i=1}^{M} x_{ui}(\Theta_{ui} - \bar{p}_i^t) \right\} + \sum_{i=1}^{M} \bar{p}_i^t c_i^t
$$

$$
\geq \sum_{u=1}^{N} \left\{ \max_{\mathbf{x}_u \in \mathcal{X}_u^t} \sum_{i=1}^{M} x_{ui}(\Theta_{ui}^* - \bar{p}_i^t) \right\} + \sum_{i=1}^{M} \bar{p}_i^t c_i^t,
$$

where the last step uses the condition that $\boldsymbol{\Theta}^* \in \mathcal{Q}_t$. Noting that $\left(c_i^t - \sum_{u=1}^N x_{ui}^t \right) \bar{p}_i^t = 0$ by complementary slackness, we can write

$$\sum_{u=1}^N \left\{ \max_{\mathbf{x}_u \in \mathcal{X}_u^t} \sum_{i=1}^M x_{ui}(\Theta_{ui}^* - \bar{p}_i^t) \right\} \leq \sum_{u=1}^N \sum_{i=1}^M x_{ui}^t \Theta_{ui}^t - \sum_{i=1}^M \bar{p}_i^t c_i^t$$
$$= \sum_{u=1}^N \sum_{i=1}^M x_{ui}^t(\Theta_{ui}^t - \bar{p}_i^t).$$

$\square$

Lemma C.12. *In the model without acceptance/rejection options, if the offered prices are $\mathbf{p}_t = \bar{\mathbf{p}}_t$, and $\boldsymbol{\Theta}^* \in \mathcal{Q}_t$ for all $t \leq T$, then*

$$\sum_{u=1}^N \mathcal{S}_u^t \leq \langle \mathbf{X}_t, \boldsymbol{\Theta}_t - \boldsymbol{\Theta}^* \rangle.$$

Proof. Using the definition of the instability and Lemma C.11, we have

$$\sum_{u=1}^N \mathcal{S}_u^t = \sum_{u=1}^N \left\{ \max_{\mathbf{x} \in \mathcal{X}_u^t} \mathbb{E}\left[\sum_{i=1}^M x_{ui}(R_{ui}^t - p_i^t) \right] - \mathbb{E}\left[\sum_{i=1}^M x_{ui}^t(R_{ui}^t - p_i^t) \right] \right\}$$
$$\leq \sum_{u=1}^N \left\{ \max_{\mathbf{x} \in \mathcal{X}_u^t} \sum_{i=1}^M x_{ui}(\Theta_{ui}^* - p_i^t) - \sum_{i=1}^M x_{ui}^t(\Theta_{ui}^* - p_i^t) \right\}$$
$$\leq \sum_{u=1}^N \left\{ \sum_{i=1}^M x_{ui}^t(\Theta_{ui}^t - p_i^t) - \sum_{i=1}^M x_{ui}^t(\Theta_{ui}^* - p_i^t) \right\}$$
$$\leq \sum_{u=1}^N \sum_{i=1}^M x_{ui}^t(\Theta_{ui}^t - \Theta_{ui}^*)$$
$$= \langle \mathbf{X}_t, \boldsymbol{\Theta}_t - \boldsymbol{\Theta}^* \rangle.$$

$\square$

Theorem C.13. *In the model without acceptance/rejection options, if the offered prices are $\mathbf{p}_t = \bar{\mathbf{p}}_t$, and $\boldsymbol{\Theta}^* \in \mathcal{Q}_t$ for all $t \leq T$, then the T period instability is bounded as*

$$\mathcal{R}^I(T, \pi) \leq \sqrt{8NM\beta_T^*(\delta, \alpha, \gamma)T \log\left(1 + \frac{T}{\gamma}\right)}. \tag{C.8}$$

Proof. Using Lemma C.12 and C.6,

$$\mathcal{R}^I(T, \pi) \leq \sum_{t=1}^T \langle \mathbf{X}_t, \boldsymbol{\Theta}_t - \boldsymbol{\Theta}^* \rangle$$
$$\leq 2\sqrt{\beta_T} \sum_{t=1}^T w_t$$
$$\leq 2\sqrt{\beta_T T \sum_{t=1}^T w_t^2}$$
$$\leq \sqrt{8\beta_T NMT \log\left(1 + \frac{T}{\gamma}\right)}.$$

$\square$

Corollary C.14. *In the model without acceptance/rejection options, letting $\delta = \mathcal{O}((NMT)^{-1})$, $\alpha = \mathcal{O}((NMT)^{-1})$ and $\gamma = 1$, Algorithm 1 achieves regret that satisfy*

$$\mathcal{R}^{SW}(T, \pi) = \tilde{\mathcal{O}}\left(\sqrt{NM(N+M)RT}\right), \tag{C.9}$$

$$\mathcal{R}^{I}(T, \pi) = \tilde{\mathcal{O}}\left(\sqrt{NM(N+M)RT}\right). \tag{C.10}$$

Proof. By Lemma C.3, $\mathbf{\Theta}^* \in \mathcal{Q}_t$ for all $t \leq T$ with probability at least $1 - 2\delta$. Therefore, the bounds of Theorem C.10 (or Theorem C.13 for instability) hold true with probability at least $1 - 2\delta$. Then, both for social welfare regret and instability regret,

$$\mathcal{R}(T, \pi) \leq (1 - \delta)\sqrt{8d\beta_T^*(\delta, \alpha, \gamma)T\log\left(1 + \frac{T}{\gamma}\right)} + 2\delta BdT.$$

Letting $\delta = \mathcal{O}(T^{-1})$, $\alpha = \mathcal{O} = (T^{-1})$ and $\gamma = 1$,

$$\mathcal{R}(T, \pi) \in \tilde{\mathcal{O}}\left(\sqrt{NM\beta_T^*(T, T^{-1}, 1)T}\right).$$

Then, we note that $\beta_T^*(T, T^{-1}, 1) \in \tilde{\mathcal{O}}\left(\eta^2 \log\left(\mathcal{N}(\mathcal{F}, T^{-1}, \|\cdot\|_2)\right)\right)$. Using Lemma C.21 to upper bound the covering number, we complete the proof.

$\square$

C.3 Extension to the Model with Acceptance/Rejection Options

Lemma C.15. *For any $\nu \geq 0$, if $\mathbf{\Theta}^* \in \mathcal{Q}_t$, then*

$$\sum_{(u,i)\in\mathcal{A}_t} \mathbb{1}\left\{\Delta_{ui}^t \geq \nu\right\} \leq \frac{2w_t\sqrt{\beta_t}}{\nu}.$$

Proof. Using Lemma C.6,

$$\sum_{(u,i)\in\mathcal{A}_t} \mathbb{1}\left\{\Delta_{ui}^t \geq \nu\right\} \leq \frac{1}{\nu}\sum_{(u,i)\in\mathcal{A}_t} \left|\Delta_{ui}^t\right|$$
$$\leq \frac{2w_t\sqrt{\beta_t}}{\nu}.$$

$\square$

Theorem C.16. *In the model with acceptance/rejection options, if the offered prices $\mathbf{p}_t$ satisfy $\mathbf{p}_t = \bar{\mathbf{p}}_t - \nu_t$ where $\nu_t = \nu\sqrt{w_t}$ for some $\nu \geq 0$, and $\mathbf{\Theta}^* \in \mathcal{Q}_t$ for all $t \leq T$, then the T period social welfare regret is bounded as*

$$\mathcal{R}^{SW}(T, \pi) \leq \left(8NM\beta_T T\log\left(1 + \frac{T}{\gamma}\right)\right)^{1/2} + \frac{2}{\nu}\left(2NMT^3\beta_T^2\log\left(1 + \frac{T}{\gamma}\right)\right)^{1/4}.$$

Proof. In the model with acceptance/rejection options, let the instantaneous regret at time t be defined as

$$\tilde{r}_t := \sum_{(u,i)\in\mathcal{A}_t^*} \Theta_{ui}^* - \sum_{(u,i)\in\mathcal{A}_t} \Theta_{ui}^*\mathbb{1}\{\Theta_{ui}^* \geq p_i^t\}.$$

Then, recalling the definition of r_t, we can write

$$\widetilde{r}_t = \sum_{(u,i)\in\mathcal{A}_t^*} \Theta_{ui}^* - \sum_{(u,i)\in\mathcal{A}_t} \Theta_{ui}^* + \sum_{(u,i)\in\mathcal{A}_t} \Theta_{ui}^* \mathbb{1}\{\Theta_{ui}^* < p_i^t\}$$

$$\le \langle \mathbf{X}_t^*, \mathbf{\Theta}^* \rangle - \langle \mathbf{X}_t, \mathbf{\Theta}^* \rangle + \sum_{(u,i)\in\mathcal{A}_t} \mathbb{1}\{\Theta_{ui}^* < p_i^t\}$$

$$\le r_t + \sum_{(u,i)\in\mathcal{A}_t} \mathbb{1}\{\Theta_{ui}^t - \Theta_{ui}^* > \Theta_{ui}^t - p_i^t\}$$

$$\le r_t + \sum_{(u,i)\in\mathcal{A}_t} \mathbb{1}\{\Delta_{ui}^t > \Theta_{ui}^t - \bar{p}_i^t + \nu_t\}$$

$$\le r_t + \sum_{(u,i)\in\mathcal{A}_t} \mathbb{1}\{\Delta_{ui}^t > \nu_t\},$$

where the last step uses the fact that $\Theta_{ui}^t \ge \bar{p}_i^t$ for $(u,i) \in \mathcal{A}_t$. Then, we have

$$\mathcal{R}^{SW}(T,\pi) = \sum_{t=1}^T \widetilde{r}_t$$

$$\le \sum_{t=1}^T r_t + \sum_{t=1}^T \sum_{(u,i)\in\mathcal{A}_t} \mathbb{1}\{\Delta_{ui}^t > \nu_t\}$$

$$\le \left(T \sum_{t=1}^T r_t^2\right)^{1/2} + 2\sqrt{\beta_T} \sum_{t=1}^T \frac{w_t}{\nu_t}$$

$$= \left(T \sum_{t=1}^T r_t^2\right)^{1/2} + \frac{2}{\nu}\sqrt{\beta_T} \sum_{t=1}^T \sqrt{w_t}$$

$$\le \left(T \sum_{t=1}^T r_t^2\right)^{1/2} + \frac{2}{\nu}\sqrt{\beta_T} \left(T^3 \sum_{t=1}^T w_t^2\right)^{1/4}$$

$$\le \left(8NM\beta_T T \log\left(1 + \frac{T}{\gamma}\right)\right)^{1/2} + \frac{2}{\nu}\left(2NMT^3\beta_T^2 \log\left(1 + \frac{T}{\gamma}\right)\right)^{1/4}.$$

$\square$

Lemma C.17. *In the model with acceptance/rejection options, if the offered prices $\mathbf{p}_t$ satisfy $\mathbf{p}_t = \bar{\mathbf{p}}_t - \nu_t$ where $\nu_t = \nu\sqrt{w_t}$ for some $\nu \ge 0$, and $\mathbf{\Theta}^* \in \mathcal{Q}_t$, then*

$$\sum_{u=1}^N \mathcal{S}_u^t \le \langle \mathbf{X}_t, \mathbf{\Theta}_t - \mathbf{\Theta}^* \rangle + \nu_t n M.$$

Proof. Using the definition of the instability and Lemma C.11, we have

$$
\begin{aligned}
\sum_{u=1}^{N} \mathcal{S}_u^t &= \sum_{u=1}^{N} \left\{ \max_{\mathbf{x} \in \mathcal{X}_u^t} \sum_{i=1}^{M} x_{ui}(\Theta_{ui}^* - p_i^t) \mathbb{1}\{\Theta_{ui}^* \geq p_i^t\} - \sum_{i=1}^{M} x_{ui}^t(\Theta_{ui}^* - p_i^t) \mathbb{1}\{\Theta_{ui}^* \geq p_i^t\} \right\} \\
&\leq \sum_{u=1}^{N} \left\{ \max_{\mathbf{x} \in \mathcal{X}_u^t} \sum_{i=1}^{M} x_{ui}(\Theta_{ui}^* - \bar{p}_i^t + \nu_t) \mathbb{1}\{\Theta_{ui}^* \geq \bar{p}_i^t - \nu_t\} - \sum_{i=1}^{M} x_{ui}^t(\Theta_{ui}^* - p_i^t) \mathbb{1}\{\Theta_{ui}^* \geq p_i^t\} \right\} \\
&\leq \sum_{u=1}^{N} \left\{ \max_{\mathbf{x} \in \mathcal{X}_u^t} \sum_{i=1}^{M} x_{ui}(\Theta_{ui}^* - \bar{p}_i^t) \mathbb{1}\{\Theta_{ui}^* \geq \bar{p}_i^t - \nu_t\} + \nu_t \max_{\mathbf{x} \in \mathcal{X}_u^t} \sum_{i=1}^{M} x_{ui} - \sum_{i=1}^{M} x_{ui}^t(\Theta_{ui}^* - p_i^t) \mathbb{1}\{\Theta_{ui}^* \geq p_i^t\} \right\} \\
&\leq \sum_{u=1}^{N} \left\{ \max_{\mathbf{x} \in \mathcal{X}_u^t} \sum_{i=1}^{M} x_{ui}(\Theta_{ui}^* - \bar{p}_i^t) \mathbb{1}\{\Theta_{ui}^* \geq \bar{p}_i^t\} + \nu_t d_u^t - \sum_{i=1}^{M} x_{ui}^t(\Theta_{ui}^* - p_i^t) \mathbb{1}\{\Theta_{ui}^* \geq p_i^t\} \right\} \\
&\leq \sum_{u=1}^{N} \left\{ \max_{\mathbf{x} \in \mathcal{X}_u^t} \sum_{i=1}^{M} x_{ui}(\Theta_{ui}^* - \bar{p}_i^t) + \nu_t d_u^t - \sum_{i=1}^{M} x_{ui}^t(\Theta_{ui}^* - p_i^t) \mathbb{1}\{\Theta_{ui}^* \geq p_i^t\} \right\} \\
&\leq \sum_{u=1}^{N} \left\{ \sum_{i=1}^{M} x_{ui}^t(\Theta_{ui}^t - \bar{p}_i^t) + \nu_t d_u^t - \sum_{i=1}^{M} x_{ui}^t(\Theta_{ui}^* - p_i^t) \right\} \\
&\leq \sum_{u=1}^{N} \left\{ \sum_{i=1}^{M} x_{ui}^t(\Theta_{ui}^t - \Theta_{ui}^*) + \nu_t d_u^t \right\} \\
&\leq \sum_{u=1}^{N} \sum_{i=1}^{M} x_{ui}^t \Delta_{ui}^t + \nu_t \sum_{u=1}^{N} d_u^t \\
&\leq \langle \mathbf{X}_t, \boldsymbol{\Theta}_t - \boldsymbol{\Theta}^* \rangle + \nu_t n M.
\end{aligned}
$$

$\square$

Theorem C.18. *In the model with acceptance/rejection options, if the offered prices $\mathbf{p}_t$ satisfy $\mathbf{p}_t = \bar{\mathbf{p}}_t - \nu_t$ where $\nu_t = \nu \sqrt{w_t}$ for some $\nu \geq 0$, and $\boldsymbol{\Theta}^* \in \mathcal{Q}_t$ for all $t \leq T$, then the T period instability is bounded as*

$$
\mathcal{R}^I(T, \pi) \leq \left(8\beta_T NMT \log\left(1 + \frac{T}{\gamma} \right) \right)^{1/2} + \nu nM \left(2NMT^3 \log\left(1 + \frac{T}{\gamma} \right) \right)^{1/4}.
$$

Proof. Using Lemma C.17 and C.6,

$$
\begin{aligned}
\mathcal{R}^I(T, \pi) &\leq \sum_{t=1}^{T} \left(\langle \mathbf{X}_t, \boldsymbol{\Theta}_t - \boldsymbol{\Theta}^* \rangle + \nu_t n M \right) \\
&\leq 2\sqrt{\beta_T} \sum_{t=1}^{T} w_t + nM \sum_{t=1}^{T} \nu_t \\
&\leq \sqrt{8\beta_T NMT \log\left(1 + \frac{T}{\gamma} \right)} + \nu nM \sum_{t=1}^{T} \sqrt{w_t} \\
&\leq \sqrt{8\beta_T NMT \log\left(1 + \frac{T}{\gamma} \right)} + \nu nM \sum_{t=1}^{T} \sqrt{w_t} \\
&\leq \left(8\beta_T NMT \log\left(1 + \frac{T}{\gamma} \right) \right)^{1/2} + \nu nM \left(2NMT^3 \log\left(1 + \frac{T}{\gamma} \right) \right)^{1/4}.
\end{aligned}
$$

$\square$

Theorem C.19. *In the model with acceptance/rejection options, if the offered prices $\mathbf{p}_t$ satisfy $\mathbf{p}_t = \bar{\mathbf{p}}_t - \nu_t$ where $\nu_t = \nu\sqrt{w_t}$ for $\nu = \left(\frac{4\beta_T}{N^2 M^2}\right)^{1/4}$, then, with probability $1 - 2\delta$, the algorithm results in regrets that satisfy*

$$\mathcal{R}^{SW}(T, \pi) \leq \sqrt{\lambda_T NMT} + (\lambda_T)^{\frac{1}{4}}(NMT)^{\frac{3}{4}},$$
$$\mathcal{R}^{I}(T, \pi) \leq \sqrt{\lambda_T NMT} + (\lambda_T)^{\frac{1}{4}}(NMT)^{\frac{3}{4}}.$$

where $\lambda_T = 8\beta_T \log(1 + T/\gamma)$.

Proof. By Theorem C.16,

$$\mathcal{R}^{SW}(T, \pi) \leq \left(8NM\beta_T T \log\left(1 + \frac{T}{\gamma}\right)\right)^{1/2} + \frac{2}{\nu}\left(2NMT^3\beta_T^2 \log\left(1 + \frac{T}{\gamma}\right)\right)^{1/4}$$
$$= (\lambda_T NMT)^{\frac{1}{2}} + (\lambda_T)^{\frac{1}{4}}(NMT)^{\frac{3}{4}}.$$

Similarly, we have

$$\mathcal{R}^{I}(T, \pi) \leq \left(8\beta_T NMT \log\left(1 + \frac{T}{\gamma}\right)\right)^{1/2} + \nu NM\left(2NMT^3 \log\left(1 + \frac{T}{\gamma}\right)\right)^{1/4}$$
$$= (\lambda_T NMT)^{\frac{1}{2}} + (\lambda_T)^{\frac{1}{4}}(NMT)^{\frac{3}{4}}.$$

$\square$

Theorem C.20. *In the model with acceptance/rejection options, Letting $\delta = \mathcal{O}((NMT)^{-1})$, $\alpha = \mathcal{O}((NMT)^{-1})$, $\gamma = 1$ and $\nu = \mathcal{O}\left(\left(\frac{\eta^2(N+M)R}{N^2 M^2}\right)^{1/4}\right)$ results in regret bounds that satisfy*

$$\mathcal{R}^{SW}(T, \pi) = \widetilde{\mathcal{O}}\left(\left(\eta^2(N+M)R\right)^{\frac{1}{4}}(NMT)^{\frac{3}{4}}\right),$$
$$\mathcal{R}^{I}(T, \pi) = \widetilde{\mathcal{O}}\left(\left(\eta^2(N+M)R\right)^{\frac{1}{4}}(NMT)^{\frac{3}{4}}\right).$$

Proof. By Theorem C.19, with probability 1,

$$\mathcal{R}^{SW}(T, \pi) \leq (1 - 2\delta)\left((\lambda_T NMT)^{\frac{1}{2}} + (\lambda_T)^{\frac{1}{4}}(NMT)^{\frac{3}{4}}\right) + 2\delta NMT.$$

Similarly, with probability 1, we have

$$\mathcal{R}^{I}(T, \pi) \leq (1 - 2\delta)\left((\lambda_T NMT)^{\frac{1}{2}} + (\lambda_T)^{\frac{1}{4}}(NMT)^{\frac{3}{4}}\right) + 2\delta NMT.$$

Let $\delta \in \mathcal{O}((NMT)^{-1})$, $\alpha \in \mathcal{O}((NMT)^{-1})$, $\gamma = 1$, and note that $\beta_T^*(T, T^{-1}, 1) \in \widetilde{\mathcal{O}}\left(\eta^2 \log\left(\mathcal{N}(\mathcal{L}, T^{-1}, \|\cdot\|_2)\right)\right) = \widetilde{\mathcal{O}}\left(\eta^2(N+M)R\right)$ to conclude the proof.

$\square$

C.4 Covering Number for Low-rank Matrices

Lemma C.21. *The covering number of $\mathcal{L}$ given in (3.5) obeys*

$$\log \mathcal{N}(\mathcal{L}, \alpha, \|\cdot\|_F) \leq (N + M + 1)R \log\left(\frac{9\sqrt{NM}}{\alpha}\right). \tag{C.11}$$

Proof. This proof is modified from (Candes and Plan, 2011). Let $\mathcal{S} = \{\boldsymbol{\Theta} \in \mathbb{R}^{N \times M} : \text{rank}(\boldsymbol{\Theta}) \leq R, \|\boldsymbol{\Theta}\|_{\text{F}} \leq 1\}$. We will first show that there exists an ϵ-net $\mathcal{S}^\epsilon$ for the Frobenious norm obeying

$$|\mathcal{S}^\epsilon| \leq (9/\epsilon)^{(N+M+1)R}.$$

For any $\boldsymbol{\Theta} \in \mathcal{S}$, singular value decomposition gives $\boldsymbol{\Theta} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\text{T}}$, where $\|\boldsymbol{\Sigma}\|_{\text{F}} \leq 1$. We will construct an ϵ-net for $\mathcal{S}$ by covering the set of permissible $\mathbf{U}$, $\boldsymbol{\Sigma}$ and $\mathbf{V}$. Let $\mathcal{D}$ be the set of diagonal matrices with nonnegative diagonal entries and Frobenious norm less than or equal to one. We take $\mathcal{D}^{\epsilon/3}$ be an $\epsilon/3$-net for $\mathcal{D}$ with $|\mathcal{D}^{\epsilon/3}| \leq (9/\epsilon)^R$. Next, let $\mathcal{O}_{N,R} = \{\mathbf{U} \in \mathbb{R}^{N \times R} : \mathbf{U}^{\text{T}}\mathbf{U} = \mathbf{I}\}$. To cover $\mathcal{O}_{N,R}$, we use the $\|\cdot\|_{1,2}$ norm defined as

$$\|\mathbf{U}\|_{1,2} = \max_i \|\mathbf{u_i}\|_{\ell_2},$$

where $\mathbf{u_i}$ denotes the ith column of $\mathbf{U}$. Let $\mathcal{Q}_{N,R} = \{\mathbf{U} \in \mathbb{R}^{N,R} : \|\mathbf{U}\|_{1,2} \leq 1\}$. It is easy to see that $\mathcal{O}_{N,R} \subset \mathcal{Q}_{N,R}$ since the columns of an orthogonal matrix are unit normed. We see that there is an $\epsilon/3$-net $\mathcal{O}_{N,R}^{\epsilon/3}$ for $\mathcal{O}_{N,R}$ obeying $|\mathcal{O}_{N,R}^{\epsilon/3}| \leq (9/\epsilon)^{NR}$. Similarly, let $\mathcal{P}_{M,R} = \{\mathbf{V} \in \mathbb{R}^{M \times R} : \mathbf{V}^{\text{T}}\mathbf{V} = \mathbf{I}\}$. By the same argument, there is an $\epsilon/3$-net $\mathcal{P}_{M,R}^{\epsilon/3}$ for $\mathcal{P}_{M,R}$ obeying $|\mathcal{P}_{M,R}^{\epsilon/3}| \leq (9/\epsilon)^{MR}$. We now let $\mathcal{S}^\epsilon = \{\bar{\mathbf{U}}\bar{\boldsymbol{\Sigma}}\bar{\mathbf{V}}^{\text{T}} : \bar{\mathbf{U}} \in \mathcal{O}_{N,R}^{\epsilon/3}, \bar{\mathbf{V}} \in \mathcal{P}_{M,R}^{\epsilon/3}, \bar{\boldsymbol{\Sigma}} \in \mathcal{D}^{\epsilon/3}\}$, and remark $|\mathcal{S}^\epsilon| \leq |\mathcal{O}_{N,R}^{\epsilon/3}||\mathcal{P}_{M,R}^{\epsilon/3}||\mathcal{D}^{\epsilon/3}| \leq (9/\epsilon)^{(N+M+1)R}$. It remains to show that for all $\boldsymbol{\Theta} \in \mathcal{S}$, there exists $\bar{\boldsymbol{\Theta}} \in \mathcal{S}^\epsilon$ with $\|\boldsymbol{\Theta} - \bar{\boldsymbol{\Theta}}\|_{\text{F}} \leq \epsilon$.

Fix $\boldsymbol{\Theta} \in \mathcal{S}$ and decompose it as $\boldsymbol{\Theta} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\text{T}}$. Then, there exists $\bar{\boldsymbol{\Theta}} = \bar{\mathbf{U}}\bar{\boldsymbol{\Sigma}}\bar{\mathbf{V}}^{\text{T}} \in \mathcal{S}^\epsilon$ with $\bar{\mathbf{U}} \in \mathcal{O}_{N,R}^{\epsilon/3}, \bar{\mathbf{V}} \in \mathcal{P}_{M,R}^{\epsilon/3}, \bar{\boldsymbol{\Sigma}} \in \mathcal{D}^{\epsilon/3}$ satisfying $\|\mathbf{U} - \bar{\mathbf{U}}\|_{1,2} \leq \epsilon/3$, $\|\mathbf{V} - \bar{\mathbf{V}}\|_{1,2} \leq \epsilon/3$ and $\|\boldsymbol{\Sigma} - \bar{\boldsymbol{\Sigma}}\|_{\text{F}} \leq \epsilon/3$. This gives

$$\begin{aligned}
\|\boldsymbol{\Theta} - \bar{\boldsymbol{\Theta}}\|_{\text{F}} &= \|\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\text{T}} - \bar{\mathbf{U}}\bar{\boldsymbol{\Sigma}}\bar{\mathbf{V}}^{\text{T}}\|_{\text{F}} \\
&= \|\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\text{T}} - \bar{\mathbf{U}}\boldsymbol{\Sigma}\mathbf{V}^{\text{T}} + \bar{\mathbf{U}}\boldsymbol{\Sigma}\mathbf{V}^{\text{T}} - \bar{\mathbf{U}}\bar{\boldsymbol{\Sigma}}\mathbf{V}^{\text{T}} + \bar{\mathbf{U}}\bar{\boldsymbol{\Sigma}}\mathbf{V}^{\text{T}} - \bar{\mathbf{U}}\bar{\boldsymbol{\Sigma}}\bar{\mathbf{V}}^{\text{T}}\|_{\text{F}} \\
&\leq \|(\mathbf{U} - \bar{\mathbf{U}})\boldsymbol{\Sigma}\mathbf{V}^{\text{T}}\|_{\text{F}} + \|\bar{\mathbf{U}}(\boldsymbol{\Sigma} - \bar{\boldsymbol{\Sigma}})\mathbf{V}^{\text{T}}\|_{\text{F}} + \|\bar{\mathbf{U}}\bar{\boldsymbol{\Sigma}}(\mathbf{V} - \bar{\mathbf{V}})^{\text{T}}\|_{\text{F}}.
\end{aligned}$$

For the first term, since $\mathbf{V}$ is an orthogonal matrix,

$$\begin{aligned}
\|(\mathbf{U} - \bar{\mathbf{U}})\boldsymbol{\Sigma}\mathbf{V}^{\text{T}}\|_{\text{F}}^2 &= \|(\mathbf{U} - \bar{\mathbf{U}})\boldsymbol{\Sigma}\|_{\text{F}}^2 \\
&\leq \|\boldsymbol{\Sigma}\|_{\text{F}}^2\|\mathbf{U} - \bar{\mathbf{U}}\|_{1,2}^2 \leq (\epsilon/3)^2.
\end{aligned}$$

By the same argument, $\|\bar{\mathbf{U}}\bar{\boldsymbol{\Sigma}}(\mathbf{V} - \bar{\mathbf{V}})^{\text{T}}\|_{\text{F}} \leq \epsilon/3$ as well. Lastly, $\|\bar{\mathbf{U}}(\boldsymbol{\Sigma} - \bar{\boldsymbol{\Sigma}})\mathbf{V}^{\text{T}}\|_{\text{F}} = \|\boldsymbol{\Sigma} - \bar{\boldsymbol{\Sigma}}\|_{\text{F}} \leq \epsilon/3$. Therefore, $\|\boldsymbol{\Theta} - \bar{\boldsymbol{\Theta}}\|_{\text{F}} \leq \epsilon$, showing that $\mathcal{S}^\epsilon$ is an ϵ-net for $\mathcal{S}$ with respect to the Frobenious norm.

Next, we will construct an α-net for $\mathcal{L}$ given in equation 3.5. Let $\kappa = \sqrt{NM}$. We start by noting that for all $\boldsymbol{\Theta} \in \mathcal{L}$, the Frobenious norm obeys $\|\boldsymbol{\Theta}\|_{\text{F}} \leq \kappa$. Then, define $\mathbf{X} = \frac{1}{\kappa}\boldsymbol{\Theta} \in \mathcal{S}$ and $\mathcal{L}^\alpha := \{\kappa\bar{\mathbf{X}} : \bar{\mathbf{X}} \in \mathcal{S}^\epsilon\}$. We previously showed that for any $\mathbf{X} \in \mathcal{S}$, there exists $\bar{\mathbf{X}} \in \mathcal{S}^\epsilon$ such that $\|\mathbf{X} - \bar{\mathbf{X}}\|_{\text{F}} \leq \epsilon$. Therefore, for any $\boldsymbol{\Theta} \in \mathcal{L}$, there exists $\bar{\boldsymbol{\Theta}} = \kappa\bar{\mathbf{X}} \in \mathcal{L}^\alpha$ such that $\|\boldsymbol{\Theta} - \bar{\boldsymbol{\Theta}}\|_{\text{F}} \leq \kappa\epsilon$. Setting $\epsilon = \alpha/\kappa$, we obtain that $\mathcal{L}^\alpha$ is an α-net for $\mathcal{L}$ with respect to the Frobenious norm. Finally, the size of $\mathcal{L}^\alpha$ obeys

$$|\mathcal{L}^\alpha| = |\mathcal{S}^{\alpha/\kappa}| \leq (9\kappa/\alpha)^{(N+M+1)R}.$$

This completes the proof. $\square$

D PROOFS FOR CONTEXTUAL ALLOCATION AND PRICING

Let $\boldsymbol{\Theta}^* \in \mathcal{F} := \{\boldsymbol{\Theta} \in \mathbb{R}^{M \times R} : \exists \mathbf{F} \in \mathbb{R}^{N \times R}, \boldsymbol{\Theta} = \mathbf{F}\boldsymbol{\Phi}^{\text{T}}, \|\mathbf{F}\|_{2,\infty} \leq 1\}$ for some known item feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{M \times R}$ that satisfy $\|\boldsymbol{\Phi}\|_{2,\infty} \leq 1$. This condition is the same as having each row $[\boldsymbol{\Theta}^*]_{u,:}$ belonging to the set $\widetilde{\mathcal{F}} := \{\boldsymbol{\theta} : \exists \mathbf{f} \in \mathbb{R}^R, \boldsymbol{\theta} = \boldsymbol{\Phi}\mathbf{f}, \|\mathbf{f}\|_2 \leq 1\}$. In other words,

$$\boldsymbol{\Theta}^* \in \mathcal{F} \iff [\boldsymbol{\Theta}^*]_{u,:} \in \widetilde{\mathcal{F}}, \forall u \in \mathcal{N}.$$

Hence, we can construct separate confidence sets for each row of the matrix $\boldsymbol{\Theta}^*$. The confidence sets that we construct are centered around the regularized least square estimates. For each user u, we let the cumulative squared prediciton error at time t be

$$L_{2,u}^t(\boldsymbol{\theta}) = \sum_{\tau=1}^{t-1} \sum_{i:(u,i)\in\mathcal{A}_\tau} (\theta_i - R_{ui}^\tau)^2, \tag{D.1}$$

and define the regularized least squares estimate at time t as

$$\widehat{\boldsymbol{\theta}}_u^t = \arg\min_{\boldsymbol{\theta} \in \widetilde{\mathcal{F}}} \left\{ L_{2,u}^t(\boldsymbol{\theta}) + \gamma \|\boldsymbol{\theta} - \boldsymbol{\theta}_u^\circ\|_2^2 \right\}, \tag{D.2}$$

where $\boldsymbol{\theta}_u^\circ = [\boldsymbol{\Theta}_\circ]_{u,:}$ is the u-th row of $\boldsymbol{\Theta}_\circ$. Then, the confidence sets take the form $\mathcal{C}_u^t := \{\boldsymbol{\theta} \in \widetilde{\mathcal{F}} : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_u^t\|_{E_{2,u}^t} \leq \sqrt{\rho_t}\}$ where ρ_t is an appropriately chosen confidence parameter, and the regularized empirical L_2-norm $\|\cdot\|_{E_{2,u}^t}$ of user u is defined by

$$\|\boldsymbol{\delta}\|_{E_{2,u}^t}^2 := \sum_{\tau=1}^{t-1} \sum_{i:(u,i)\in\mathcal{A}_\tau} \langle \boldsymbol{\delta}, \mathbf{e}_i \rangle^2 + \gamma \|\boldsymbol{\delta}\|_2^2 = \sum_{i=1}^d (n_{ui}^t + \gamma)(\delta_i)^2.$$

Similar to our analysis in the low-rank setting, we start with providing the following guarantee for the confidence regions $\mathcal{C}_u^t$. First, we let

$$\rho_t^*(\delta, \alpha, \gamma) := 8R\eta^2 \log\left(3N/(\alpha\delta)\right) + 2\alpha t M \left[8B + \sqrt{8\eta^2 \log(4MNt^2/\delta)}\right] + 4\gamma G^2.$$

Lemma D.1. *For any $\delta > 0$, $\alpha > 0$ and $\gamma > 0$, if*

$$\mathcal{C}_u^t = \{\boldsymbol{\theta} \in \widetilde{\mathcal{F}} : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_u^t\|_{E_{2,u}^t} \leq \sqrt{\rho_t^*(\delta, \alpha, \gamma)}\} \tag{D.3}$$

for all $t \in \mathbb{N}$, then

$$\mathbb{P}\left(\boldsymbol{\theta}_u^* \in \mathcal{C}_u^t \quad, \forall t \in \mathbb{N}\right) \geq 1 - 2\delta/N. \tag{D.4}$$

Proof. We start by computing the covering number $\mathcal{N}(\widetilde{\mathcal{F}}, \alpha, \|\cdot\|_2)$. Let $\mathcal{B}^\alpha$ be an α covering of the ball $\mathcal{B} := \{\mathbf{f} : \|\mathbf{f}\|_2 \leq 1\}$ in L_2-norm. Then, for any $\boldsymbol{\theta} \in \widetilde{\mathcal{F}}$, there exists $\mathbf{f} \in \mathcal{B}$ such that $\boldsymbol{\theta} = \boldsymbol{\Phi}^T \mathbf{f}$ and $\mathbf{f}_\alpha \in \mathcal{B}^\alpha$ such that $\|\mathbf{f}_\alpha - \mathbf{f}\|_2 \leq \alpha$. Hence, for $\boldsymbol{\theta}_\alpha = \boldsymbol{\Phi}^T \mathbf{f}_\alpha$, we have $\|\boldsymbol{\theta} - \boldsymbol{\theta}_\alpha\|_2 = \|\boldsymbol{\Phi}^T(\mathbf{f} - \mathbf{f}_\alpha)\|_2 \leq \alpha$. Therefore, the set $\widetilde{\mathcal{F}}^\alpha := \{\boldsymbol{\theta} : \boldsymbol{\theta} = \boldsymbol{\Phi}^T \mathbf{f}, \mathbf{f} \in \mathcal{B}^\alpha\}$ is an α cover of $\widetilde{\mathcal{F}}$ in L_2-norm. Taking the minimum over the size of covers, we obtain $\mathcal{N}(\widetilde{\mathcal{F}}, \alpha, \|\cdot\|_2) \leq (3/\alpha)^R$. Then, the result follows from Lemma C.3 by setting $N = 1$ (since we are only considering a single row) and writing $\mathcal{N}(\widetilde{\mathcal{F}}, \alpha, \|\cdot\|_2) \leq (3/\alpha)^R$.

□

Now, in order to combine the guarantees for confidence intervals of different users in a single expression, let us define empirical $L_{2,\infty}$-norm as

$$\|\boldsymbol{\Delta}\|_{E_{2,\infty}^t} := \max_{u \in \mathcal{N}} \|[\boldsymbol{\Delta}]_{u,:}\|_{E_{2,u}^t}$$

In other words, $\|\boldsymbol{\Delta}\|_{E_{2,\infty}^t}$ is equal to the maximum of emprical L_2-norms for the rows of $\boldsymbol{\Delta}$.

Then, we have the following guarantee

Lemma D.2. *For any $\delta > 0$, $\alpha > 0$ and $\gamma > 0$, if*

$$\mathcal{C}_t = \{\boldsymbol{\Theta} \in \mathcal{F} : \|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}_t\|_{E_{2,\infty}^t} \leq \sqrt{\rho_t^*(\delta, \alpha, \gamma)}\} \tag{D.5}$$

for all $t \in \mathbb{N}$, then

$$\mathbb{P}\left(\boldsymbol{\Theta}^* \in \mathcal{C}_t \quad, \forall t \in \mathbb{N}\right) \geq 1 - 2\delta. \tag{D.6}$$

Proof. We recall the definition of $\widehat{\boldsymbol{\Theta}}_t$:

$$\widehat{\boldsymbol{\Theta}}_t = \arg\min_{\boldsymbol{\Theta} \in \mathcal{F}} \left\{ L_{2,t}(\boldsymbol{\Theta}) + \gamma \|\boldsymbol{\Theta} - \overline{\boldsymbol{\Theta}}\|_F^2 \right\}. \tag{D.7}$$

Furthermore, note that $L_{2,t}(\boldsymbol{\Theta}) = \sum_{u \in \mathcal{N}} L_{2,u}^t([\boldsymbol{\Theta}]_{u,:})$ and $\|\boldsymbol{\Theta} - \overline{\boldsymbol{\Theta}}\|_F^2 = \sum_{u \in \mathcal{N}} \|[\boldsymbol{\Theta}]_{u,:} - \boldsymbol{\theta}_u^\circ\|_2^2$. Therefore, the rows of $\widehat{\boldsymbol{\Theta}}_t$ are equal to $\widehat{\boldsymbol{\theta}}_u^t$. Hence, by a union bound applied to Lemma D.1, the result follows. □

Now, we continue with showing an analog of Lemma C.4 for confidence regions $\mathcal{C}_t$

Lemma D.3. *Let $\mathbf{X}_t \in \mathcal{X}_t$ and $\boldsymbol{\Theta} \in \mathcal{C}_t$. Then,*

$$|\langle \boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}^t, \mathbf{X}_t \rangle| \leq w_t \sqrt{|\mathcal{N}_t|\rho_t} \tag{D.8}$$

where $w_t = \|\mathbf{X}_t\|_{\mathbf{A}_t^{-1}}$ is the "confidence width" of an action $\mathbf{X_t}$ at time t.

Proof. Let $\boldsymbol{\Delta} = \boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}^t$. Then,

$$\begin{aligned}
|\langle \boldsymbol{\Delta}, \mathbf{X}_t \rangle| &= |\sum_{u \in \mathcal{N}_t} \sum_{i:(u,i) \in \mathcal{A}_t} \Delta_{ui}| \\
&\leq \sum_{u \in \mathcal{N}_t} \sum_{i:(u,i) \in \mathcal{A}_t} |\Delta_{ui}| \\
&\leq \sum_{u \in \mathcal{N}_t} \left(\sum_{i:(u,i) \in \mathcal{A}_t} (n_{ui}^t + \gamma)\Delta_{ui}^2 \right)^{1/2} \left(\sum_{i:(u,i) \in \mathcal{A}_t} \frac{1}{(n_{ui}^t + \gamma)} \right)^{1/2} \\
&\leq \sum_{u \in \mathcal{N}_t} \sqrt{\rho_t} \left(\sum_{i:(u,i) \in \mathcal{A}_t} \frac{1}{(n_{ui}^t + \gamma)} \right)^{1/2} \\
&\leq \sqrt{\rho_t} \sum_{u \in \mathcal{N}_t} \left(\sum_{i:(u,i) \in \mathcal{A}_t} \frac{1}{(n_{ui}^t + \gamma)} \right)^{1/2} \\
&\leq \sqrt{\rho_t} \left(|\mathcal{N}_t| \sum_{u \in \mathcal{N}_t} \sum_{i:(u,i) \in \mathcal{A}_t} \frac{1}{(n_{ui}^t + \gamma)} \right)^{1/2} \\
&\leq w_t \sqrt{|\mathcal{N}_t|\rho_t}
\end{aligned}$$

$\square$

Lemma D.4. *Define the regret at time t as $r_t := \langle \mathbf{X}_t^*, \boldsymbol{\Theta}^* \rangle - \langle \mathbf{X}_t, \boldsymbol{\Theta}^* \rangle$. If $\boldsymbol{\Theta}^* \in \mathcal{C}_t$, then*

$$r_t \leq \langle \mathbf{X}_t, \boldsymbol{\Theta}_t - \boldsymbol{\Theta}^* \rangle \leq 2w_t \sqrt{|\mathcal{N}_t|\rho_t}.$$

Proof. The proof is similar to the proof of Lemma C.6 using the results of previous lemma. $\square$

Theorem D.5. *If $\gamma \geq 1$ and $\boldsymbol{\Theta}^* \in \mathcal{C}_t$ for all $t \leq T$, then the T period social welfare regret is bounded by*

$$\mathcal{R}^{SW}(T, \pi) \leq \sqrt{8nNM\rho_T^*(\delta, \alpha, \gamma)T \log \left(1 + \frac{T}{\gamma} \right)}$$

Proof. Recall the definition of instantaneous regret at time t: $r_t = \langle \mathbf{X}_t^*, \boldsymbol{\Theta}^* \rangle - \langle \mathbf{X}_t, \boldsymbol{\Theta}^* \rangle$. Assuming that $\boldsymbol{\Theta}^* \in \mathcal{C}_t$ for all $t \leq T$, we have

$$\begin{aligned}
\mathcal{R}^{SW}(T, \pi) &= \sum_{t=1}^T r_t \\
&\leq \left(T \sum_{t=1}^T r_t^2 \right)^{1/2} \\
&\leq \left(4n\rho_T T \sum_{t=1}^T w_t^2 \right)^{1/2},
\end{aligned}$$

where the last step follows from Lemma D.5. Then, using Lemma C.9, we prove the result.

$\square$

Theorem D.6. *If the offered prices are* $\mathbf{p}_t = \bar{\mathbf{p}}_t$, *and* $\boldsymbol{\Theta}^* \in \mathcal{C}_t$ *for all* $t \leq T$, *then the* T *period instability is bounded as*

$$\mathcal{R}^I(T, \pi) \leq \sqrt{8nNM\rho_T^*(\delta, \alpha, \gamma)T \log\left(1 + \frac{T}{\gamma}\right)}. \tag{D.9}$$

Proof. Using Lemma C.12 and D.4,

$$\mathcal{R}^I(T, \pi) \leq \sum_{t=1}^{T} \langle \mathbf{X}_t, \boldsymbol{\Theta}_t - \boldsymbol{\Theta}^* \rangle$$

$$\leq 2\sqrt{n\rho_T T \sum_{t=1}^{T} w_t^2}$$

$$\leq \sqrt{8n\rho_T NMT \log\left(1 + \frac{T}{\gamma}\right)}.$$

$\square$

Lemma D.7. *For any* $\nu_t \geq 0$, *if* $\boldsymbol{\Theta}^* \in \mathcal{C}_t$, *then*

$$\sum_{(u,i)\in\mathcal{A}_t} \mathbb{1}\left\{\Delta_{ui}^t \geq \nu_t\right\} \leq \frac{2w_t\sqrt{n\rho_t}}{\nu_t}.$$

Proof. Using Lemma D.3,

$$\sum_{(u,i)\in\mathcal{A}_t} \mathbb{1}\left\{\Delta_{ui}^t \geq \nu_t\right\} \leq \frac{1}{\nu_t} \sum_{(u,i)\in\mathcal{A}_t} \left|\Delta_{ui}^t\right|$$

$$\leq \frac{2w_t\sqrt{n\rho_t}}{\nu_t}.$$

$\square$

D.1 Extension to the Model with Acceptance/Rejection Options

Theorem D.8. *In the model with acceptance/rejection options, if the offered prices* $\mathbf{p}_t$ *satisfy* $\mathbf{p}_t = \bar{\mathbf{p}}_t - \nu_t$ *where* $\nu_t = \nu\sqrt{w_t}$ *for some* $\nu \geq 0$, *and* $\boldsymbol{\Theta}^* \in \mathcal{C}_t$ *for all* $t \leq T$, *then the* T *period social welfare regret is bounded as*

$$\mathcal{R}^{SW}(T, \pi) \leq \left(8NMn\rho_T T \log\left(1 + \frac{T}{\gamma}\right)\right)^{1/2} + \frac{2}{\nu}\left(2NMT^3n^2\rho_T^2 \log\left(1 + \frac{T}{\gamma}\right)\right)^{1/4}.$$

Proof. Recall the definition of instantaneous regret at time t in the model with acceptance/rejection options:

$$\widetilde{r}_t := \sum_{(u,i)\in\mathcal{A}_t^*} \Theta_{ui}^* - \sum_{(u,i)\in\mathcal{A}_t} \Theta_{ui}^* \mathbb{1}\{\Theta_{ui}^* \geq p_i^t\}.$$

Recalling the definition of r_t, similar to proof of Lemma C.16, we can write

$$\widetilde{r}_t \leq r_t + \sum_{(u,i)\in\mathcal{A}_t} \mathbb{1}\{\Delta_{ui}^t > \nu_t\}.$$

Then, we have

$$\mathcal{R}^{SW}(T, \pi) = \sum_{t=1}^{T} \widetilde{r}_t$$

$$\leq \sum_{t=1}^{T} r_t + \sum_{t=1}^{T} \sum_{(u,i) \in \mathcal{A}_t} \mathbb{1}\{\Delta_{ui}^t > \nu_t\}$$

$$\leq \left(T \sum_{t=1}^{T} r_t^2 \right)^{1/2} + 2\sqrt{n\rho_T} \sum_{t=1}^{T} \frac{w_t}{\nu_t}$$

$$\leq \left(T \sum_{t=1}^{T} r_t^2 \right)^{1/2} + \frac{2}{\nu} \left(T^3 n^2 \rho_T^2 \sum_{t=1}^{T} w_t^2 \right)^{1/4}$$

$$\leq \left(8NMn\rho_T T \log\left(1 + \frac{T}{\gamma}\right) \right)^{1/2} + \frac{2}{\nu} \left(2NMT^3 n^2 \rho_T^2 \log\left(1 + \frac{T}{\gamma}\right) \right)^{1/4}.$$

$\square$

Theorem D.9. *In the model with acceptance/rejection options, if the offered prices $\mathbf{p}_t$ satisfy $\mathbf{p}_t = \bar{\mathbf{p}}_t - \nu_t$ where $\nu_t = \nu\sqrt{w_t}$ for some $\nu \geq 0$, and $\mathbf{\Theta}^* \in \mathcal{C}_t$ for all $t \leq T$, then the T period instability is bounded as*

$$\mathcal{R}^{I}(T, \pi) \leq \left(8n\rho_T NM \log\left(1 + \frac{T}{\gamma}\right) \right)^{1/2} + \nu nM \left(2NMT^3 \log\left(1 + \frac{T}{\gamma}\right) \right)^{1/4}.$$

Proof. Using Lemma C.17 and D.4,

$$\mathcal{R}^{I}(T, \pi) \leq \sum_{t=1}^{T} \langle \mathbf{X}_t, \mathbf{\Theta}_t - \mathbf{\Theta}^* \rangle + \nu nM \sum_{t=1}^{T} \sqrt{w_t}$$

$$\leq 2\sqrt{n\rho_T} \sum_{t=1}^{T} w_t + \nu nM \sum_{t=1}^{T} \sqrt{w_t}$$

$$\leq \left(8n\rho_T NM \log\left(1 + \frac{T}{\gamma}\right) \right)^{1/2} + \nu nM \left(2NMT^3 \log\left(1 + \frac{T}{\gamma}\right) \right)^{1/4}.$$

$\square$

Theorem D.10. *In the model with acceptance/rejection options, if the offered prices $\mathbf{p}_t$ satisfy $\mathbf{p}_t = \bar{\mathbf{p}}_t - \nu_t$ where $\nu_t = \nu\sqrt{w_t}$ for $\nu = \left(\frac{4\rho_T}{nM^2}\right)^{1/4}$, then, with probability $1 - 2\delta$, the algorithm results in regrets that satisfy*

$$\mathcal{R}^{SW}(T, \pi) \leq \sqrt{\kappa_T nMT} + (\kappa_T)^{\frac{1}{4}}(nMT)^{\frac{3}{4}}$$
$$\mathcal{R}^{I}(T, \pi) \leq \sqrt{\kappa_T nMT} + (\kappa_T)^{\frac{1}{4}}(nMT)^{\frac{3}{4}},$$

where $\kappa_T = 8N\rho_T \log(1 + T/\gamma)$.

Proof. We have

$$\mathcal{R}^{SW}(T, \pi) \leq \left(8NMn\rho_T T \log\left(1 + \frac{T}{\gamma}\right) \right)^{1/2} + \frac{2}{\nu} \left(2NMT^3 n^2 \rho_T^2 \log\left(1 + \frac{T}{\gamma}\right) \right)^{1/4}$$

$$= (\kappa_T nMT)^{\frac{1}{2}} + (\kappa_T)^{\frac{1}{4}}(nMT)^{\frac{3}{4}}.$$

Similarly, we have

$$\mathcal{R}^{I}(T, \pi) \leq \left(8n\rho_T NM \log\left(1 + \frac{T}{\gamma}\right) \right)^{1/2} + \nu nM \left(2NMT^3 \log\left(1 + \frac{T}{\gamma}\right) \right)^{1/4}$$

$$= (\kappa_T nMT)^{\frac{1}{2}} + (\kappa_T)^{\frac{1}{4}}(nMT)^{\frac{3}{4}}.$$

$\square$

Theorem D.11. *In the model with acceptance/rejection options, Letting $\delta = \mathcal{O}((NMT)^{-1})$, $\alpha = \mathcal{O}((NMT)^{-1})$, $\gamma = 1$ and $\nu = \mathcal{O}\left(\left(\frac{\eta^2 nR}{N^2 M^2}\right)^{1/4}\right)$ results in regret bounds that satisfy*

$$\mathcal{R}^{SW}(T, \pi) = \widetilde{\mathcal{O}}\left((\eta^2 nR)^{\frac{1}{4}}(NMT)^{\frac{3}{4}}\right)$$

$$\mathcal{R}^{I}(T, \pi) = \widetilde{\mathcal{O}}\left((\eta^2 nR)^{\frac{1}{4}}(NMT)^{\frac{3}{4}}\right).$$

Proof. By Theorem D.10, with probability 1,

$$\mathcal{R}^{SW}(T, \pi) \leq (1 - 2\delta)\left((\kappa_T NMT)^{\frac{1}{2}} + (\kappa_T)^{\frac{1}{4}}(NMT)^{\frac{3}{4}}\right) + 2\delta NMT.$$

Similarly, with probability 1, we have

$$\mathcal{R}^{I}(T, \pi) \leq (1 - 2\delta)\left((\kappa_T NMT)^{\frac{1}{2}} + (\kappa_T)^{\frac{1}{4}}(NMT)^{\frac{3}{4}}\right) + 2\delta NMT.$$

Letting $\delta = \in \mathcal{O}((NMT)^{-1})$, $\alpha \in \mathcal{O}((NMT)^{-1})$, $\gamma \in \mathcal{O}(1)$, we conclude the proof.

$\square$

E   MARTINGALE EXPONENTIAL INEQUALITIES

Consider a sequence of random variables $(Z_n)_{n \in \mathbb{N}}$ adapted to the filtration $(\mathcal{H}_n)_{n \in \mathbb{N}}$. Assume $\mathbb{E}[\exp(\lambda Z_i)]$ is finite for all λ. Define the conditional mean $\mu_i = \mathbb{E}[Z_i | \mathcal{H}_{i-1}]$, and define the conditional cumulant generating function of the centered random variable $[Z_i - \mu_i]$ by $\psi_i(\lambda) := \log \mathbb{E}[\exp(\lambda[Z_i - \mu_i]) | \mathcal{H}_{i-1}]$. Let

$$M_n(\lambda) = \exp\left\{\sum_{i=1}^{n} \lambda[Z_i - \mu_i] - \psi_i(\lambda)\right\}.$$

Lemma E.1. $(M_n(\lambda))_{n \in \mathbb{N}}$ *is a martingale with respect to the filtration $(\mathcal{H}_n)_{n \in \mathbb{N}}$, and $\mathbb{E}[M_n(\lambda)] = 1$.*

Proof. By definition, we have

$$\mathbb{E}[M_1(\lambda) | \mathcal{H}_0] = \mathbb{E}[\exp\{\lambda[Z_1 - \mu_1] - \psi_1(\lambda)\} | \mathcal{H}_0] = 1.$$

Then, for any $n \geq 2$,

$$\begin{aligned}
\mathbb{E}[M_n(\lambda) | \mathcal{H}_{n-1}] &= \mathbb{E}[M_{n-1}(\lambda) \exp\{\lambda[Z_n - \mu_n] - \psi_n(\lambda)\} | \mathcal{H}_{n-1}] \\
&= M_{n-1}(\lambda)\mathbb{E}[\exp\{\lambda[Z_n - \mu_n] - \psi_n(\lambda)\} | \mathcal{H}_{n-1}] \\
&= M_{n-1}(\lambda),
\end{aligned}$$

since $M_{n-1}(\lambda)$ is a measurable function of the filtration $\mathcal{H}_{n-1}$. $\square$

Lemma E.2. *For all $x \geq 0$ and $\lambda \geq 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} \lambda Z_i \leq x + \sum_{i=1}^{n}[\lambda \mu_i + \psi_i(\lambda)] \quad, \forall t \in \mathbb{N}\right) \geq 1 - e^{-x}.$$

Proof. For any λ, $(M_n(\lambda))_{n \in \mathbb{N}}$ is a martingale with respect to $(\mathcal{H}_n)_{n \in \mathbb{N}}$ and $\mathbb{E}[M_n(\lambda)] = 1$ by Lemma E.1. For arbitrary $x \geq 0$, define $\tau_x = \inf\{n \geq 0 | M_n(\lambda) \geq x\}$ and note that τ_x is a stopping time corresponding to the first time M_n crosses the boundary x. Since τ is a stopping time with respect to $(\mathcal{H}_n)_{n \in \mathbb{N}}$, we have $\mathbb{E}[M_{\tau_x \wedge n}(\lambda)] = 1$. Then, by Markov's inequality

$$x\mathbb{P}(M_{\tau_x \wedge n}(\lambda) \geq x) \leq \mathbb{E}[M_{\tau_x \wedge n}(\lambda)] = 1.$$

Noting that the event $\{M_{\tau_x \wedge n}(\lambda) \geq x\} = \bigcup_{k=1}^{n}\{M_k(\lambda) \geq x\}$, we have

$$\mathbb{P}\left(\bigcup_{k=1}^{n}\{M_k(\lambda) \geq x\}\right) \leq \frac{1}{x}.$$

Taking the limit as $n \to \infty$, and applying monotone convergence theorem shows that $\mathbb{P}\left(\bigcup_{k=1}^{\infty}\{M_k(\lambda) \geq x\}\right) \leq \frac{1}{x}$ or $\mathbb{P}\left(\bigcup_{k=1}^{\infty}\{M_k(\lambda) \geq e^x\}\right) \leq e^{-x}$. Then, by definition of $M_k(\lambda)$, we conclude

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty}\left\{\sum_{i=1}^{n}\lambda[Z_i - \mu_i] - \psi_i(\lambda) \geq x\right\}\right) \leq e^{-x}.$$

$\square$

F   IMPLEMENTATION VIA ALTERNATING OPTIMIZATION

In order to efficiently solve the least squares problem given in line 2 of Algorithm 1 as well as the OFU problem (3.1) in large scales, we take advantage of the matrix factorization model. As a result, we factorize $\boldsymbol{\Theta} = \mathbf{F}\boldsymbol{\Phi}^{\mathsf{T}}$ where $\mathbf{F} \in \mathbb{R}^{N \times R}$ and $\boldsymbol{\Phi} \in \mathbb{R}^{M \times R}$, and solve the problems by optimizing over $\mathbf{F}$ and $\boldsymbol{\Phi}$ rather than directly optimizing over $\boldsymbol{\Theta}$. Even if the least squares problem is not convex in the joint variable $(\mathbf{F}, \boldsymbol{\Phi})$, it is convex in $\mathbf{F}$ for fixed $\boldsymbol{\Phi}$ and it is convex in $\boldsymbol{\Phi}$ for fixed $\mathbf{F}$. Therefore, an alternating minimization algorithm becomes a feasible choice to find a reasonable solution for the least squares problem even in the case with unknown $\boldsymbol{\Phi}$ matrix. Similarly, an alternating minimization approach is also useful to solve the problem (3.1). We can fix an allocation $\mathbf{X}$ and minimize over $\mathbf{F}$ and $\boldsymbol{\Phi}$. Then, for fixed $\mathbf{F}$ and $\boldsymbol{\Phi}$, the allocation $\mathbf{X}$ and prices $\bar{\mathbf{p}}$ are determined through the linear program described in the section 2.1. Note that converged $\widehat{\boldsymbol{\Theta}}_t$ and $\mathbf{X}_t$ are not necessarily the optimum solution in the case with unknown $\boldsymbol{\Phi}$ matrix, since the optimization problems are not convex. However, the alternating optimization algorithm guarantees that the objective value only improves in each iteration of the alternating optimization.

The following algorithm describes an efficient implementation of CX-ILAP algorithm.

Algorithm 2 CX-ILAP with Alternating Optimization

Require: horizon T, initial estimate $\boldsymbol{\Theta}_\circ$ with $\|\boldsymbol{\Theta}_\circ - \boldsymbol{\Theta}^*\|_{\mathrm{F}} \leq G$, parameters $\delta, \alpha, \nu > 0, \gamma \geq 1$.
    for $t = 1, 2, \ldots, T$ **do**
        randomly initialize $\widehat{\mathbf{F}}$ and $\widehat{\boldsymbol{\Phi}}$
        $\widehat{\mathbf{F}} \leftarrow \arg\min_{\mathbf{F} \in \mathbb{R}^{N \times R}} \left\{\sum_{\tau=1}^{t-1}\sum_{(u,i) \in \mathcal{A}_\tau}(\mathbf{f}_u^{\mathsf{T}}\phi_i - R_{ui}^\tau)^2 + \gamma\|\mathbf{F}\boldsymbol{\Phi}^{\mathsf{T}} - \boldsymbol{\Theta}_\circ\|_{\mathrm{F}}^2\right\}$
        $\widehat{\boldsymbol{\Theta}}_t \leftarrow \widehat{\mathbf{F}}\boldsymbol{\Phi}^{\mathsf{T}}$
        $\mathbf{X} \leftarrow \mathbb{1}_{N \times M}, \mathbf{F} \leftarrow \widehat{\mathbf{F}}, \boldsymbol{\Phi} \leftarrow \widehat{\boldsymbol{\Phi}}$
        while convergence criterion not satisfied **do**
            $\mathbf{F} \leftarrow \arg\max_{\mathbf{F} \in \mathbb{R}^{N \times R}} \langle \mathbf{X}, \mathbf{F}\boldsymbol{\Phi}^{\mathsf{T}} \rangle$ s.t. $\|\mathbf{F}\boldsymbol{\Phi}^{\mathsf{T}} - \widehat{\boldsymbol{\Theta}}_t\|_{E_{2,\infty}^t} \leq \sqrt{\rho_t^*(\delta, \alpha, \gamma)}$
            $\boldsymbol{\Theta} \leftarrow \mathbf{F}\boldsymbol{\Phi}^{\mathsf{T}}$
            $(\mathbf{X}, \bar{\mathbf{p}}) \leftarrow$ optimal primal-dual variables for the problem $\max_{\mathbf{X} \in \mathcal{X}_t} \langle \mathbf{X}, \boldsymbol{\Theta} \rangle$
        end while
        $(\mathbf{X}_t, \bar{\mathbf{p}}_t) \leftarrow (\mathbf{X}, \bar{\mathbf{p}})$
        $\mathbf{p}_t = \bar{\mathbf{p}}_t - \nu\sqrt{w_t}$ for w_t defined in (3.4)
        Offer the allocation $\mathbf{X}_t$ together with prices $\mathbf{p}_t$
        Observe R_{ui}^t for all $(u, i) \in \mathcal{A}_t$
    end for

The following algorithm describes an efficient implementation of LR-ILAP algorithm using matrix factorization.

Algorithm 3 LR-ILAP with Alternating Optimization using Matrix Factorization

---

Require: horizon T, initial estimate $\mathbf{\Theta}_\circ$ with $\|\mathbf{\Theta}_\circ - \mathbf{\Theta}^*\|_F \leq G$, parameters $\delta, \alpha, \nu > 0$, $\gamma \geq 1$.

  for $t = 1, 2, \ldots, T$ **do**

    randomly initialize $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{\Phi}}$

    while convergence criterion not satisfied **do**

$$\widehat{\mathbf{F}} \leftarrow \arg\min_{\mathbf{F} \in \mathbb{R}^{N \times R}} \left\{ \sum_{\tau=1}^{t-1} \sum_{(u,i) \in \mathcal{A}_\tau} (\mathbf{f}_u^{\mathrm{T}} \phi_i - R_{ui}^\tau)^2 + \gamma \|\mathbf{F}\widehat{\mathbf{\Phi}}^{\mathrm{T}} - \mathbf{\Theta}_\circ\|_F^2 \right\}$$

$$\widehat{\mathbf{\Phi}} \leftarrow \arg\min_{\mathbf{\Phi} \in \mathbb{R}^{M \times R}} \left\{ \sum_{\tau=1}^{t-1} \sum_{(u,i) \in \mathcal{A}_\tau} (\mathbf{f}_u^{\mathrm{T}} \phi_i - R_{ui}^\tau)^2 + \gamma \|\widehat{\mathbf{F}}\mathbf{\Phi}^{\mathrm{T}} - \mathbf{\Theta}_\circ\|_F^2 \right\}$$

    end while

    $\widehat{\mathbf{\Theta}}_t \leftarrow \widehat{\mathbf{F}}\widehat{\mathbf{\Phi}}^{\mathrm{T}}$

    $\mathbf{X} \leftarrow \mathbb{1}_{N \times M}$, $\mathbf{F} \leftarrow \widehat{\mathbf{F}}$, $\mathbf{\Phi} \leftarrow \widehat{\mathbf{\Phi}}$

    while convergence criterion not satisfied **do**

      while convergence criterion not satisfied **do**

$$\mathbf{F} \leftarrow \arg\max_{\mathbf{F} \in \mathbb{R}^{N \times R}} \langle \mathbf{X}, \mathbf{F}\mathbf{\Phi}^{\mathrm{T}} \rangle \text{ s.t. } \|\mathbf{F}\mathbf{\Phi}^{\mathrm{T}} - \widehat{\mathbf{\Theta}}_t\|_{E_2^t} \leq \sqrt{\beta_t^*(\delta, \alpha, \gamma)}$$

$$\mathbf{\Phi} \leftarrow \arg\max_{\mathbf{\Phi} \in \mathbb{R}^{M \times R}} \langle \mathbf{X}, \mathbf{F}\mathbf{\Phi}^{\mathrm{T}} \rangle \text{ s.t. } \|\mathbf{F}\mathbf{\Phi}^{\mathrm{T}} - \widehat{\mathbf{\Theta}}_t\|_{E_2^t} \leq \sqrt{\beta_t^*(\delta, \alpha, \gamma)}$$

      end while

      $\mathbf{\Theta} \leftarrow \mathbf{F}\mathbf{\Phi}^{\mathrm{T}}$

      $(\mathbf{X}, \bar{\mathbf{p}}) \leftarrow$ optimal primal-dual variables for the problem $\max_{\mathbf{X} \in \mathcal{X}_t} \langle \mathbf{X}, \mathbf{\Theta} \rangle$

    end while

    $(\mathbf{X}_t, \bar{\mathbf{p}}_t) \leftarrow (\mathbf{X}, \bar{\mathbf{p}})$

    $\mathbf{p}_t = \bar{\mathbf{p}}_t - \nu\sqrt{w_t}$ for w_t defined in (3.4)

    Offer the allocation $\mathbf{X}_t$ together with prices $\mathbf{p}_t$

    Observe R_{ui}^t for all $(u,i) \in \mathcal{A}_t$

  end for

G   ADDITIONAL EXPERIMENTAL RESULTS AND DETAILS

All experiments are implemented in Python and carried out on a cluster with Lenovo NeXtScale nx360m5 nodes equipped with two Intel Xeon 12-core Haswell processors (24 cores per node) with core frequency 2.3 GHz and 64 GB of memory. We solve the allocation program (2.1) using large-scale mixed integer programming (MIP) solver packages to have efficient computations.

Parameter setup:

- In all data sets, $\mathbf{\Theta}^*$ is scaled such that $\|\mathbf{\Theta}^*\|_\infty \leq 1$.

- Standard deviation of the rewards: $\eta = 0.2$.

- In the static setting, $d_u^t = 1$ for all $u \in [N]$.

- In the dynamic setting, $d_u^t = 1$ with probability 0.2, 0 otherwise, independently for each $u \in [N]$.

- $C_{\max} = \text{ceil}\left(\frac{1}{M} \sum_{u=1}^N d_u^t\right)$.

- $c_{t,i}$ are uniformly sampled over $\{1, \ldots, C_{\max}\}$ independently for each $t \in [T]$ and $i \in [M]$.

In Figures 3, 4, 5 and 6, we provide detailed results for different experimental settings described in Section 4.
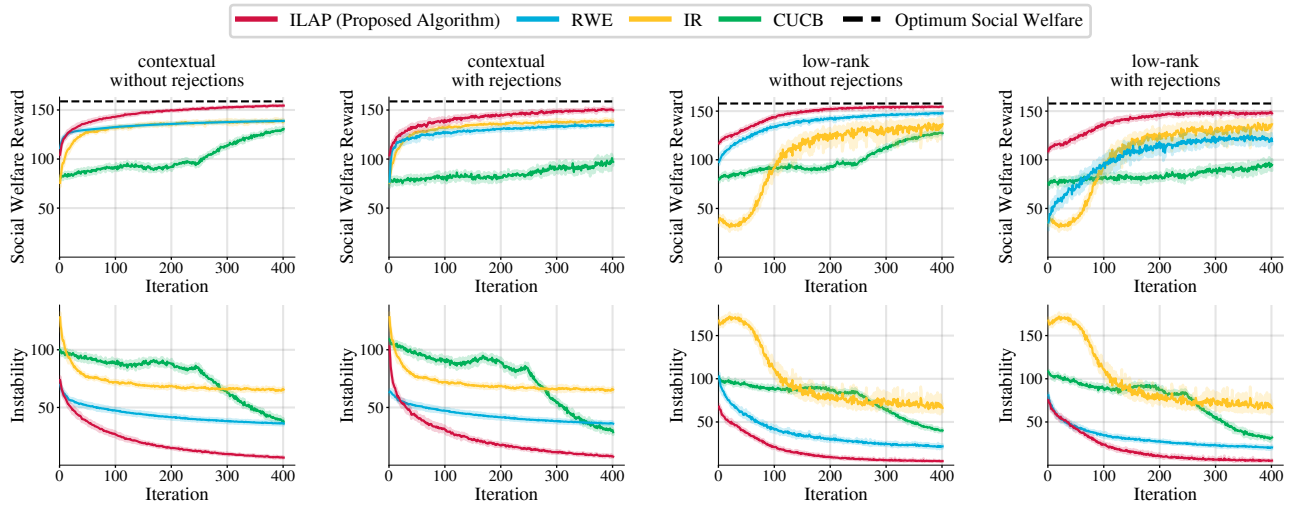
Figure 3: Experimental results for synthetic data in a static setting with $N = 250$, $M = 200$, $R = 20$. The experiments are run on 20 problem instances and means are reported together with error regions that indicate one standard deviation of uncertainty.
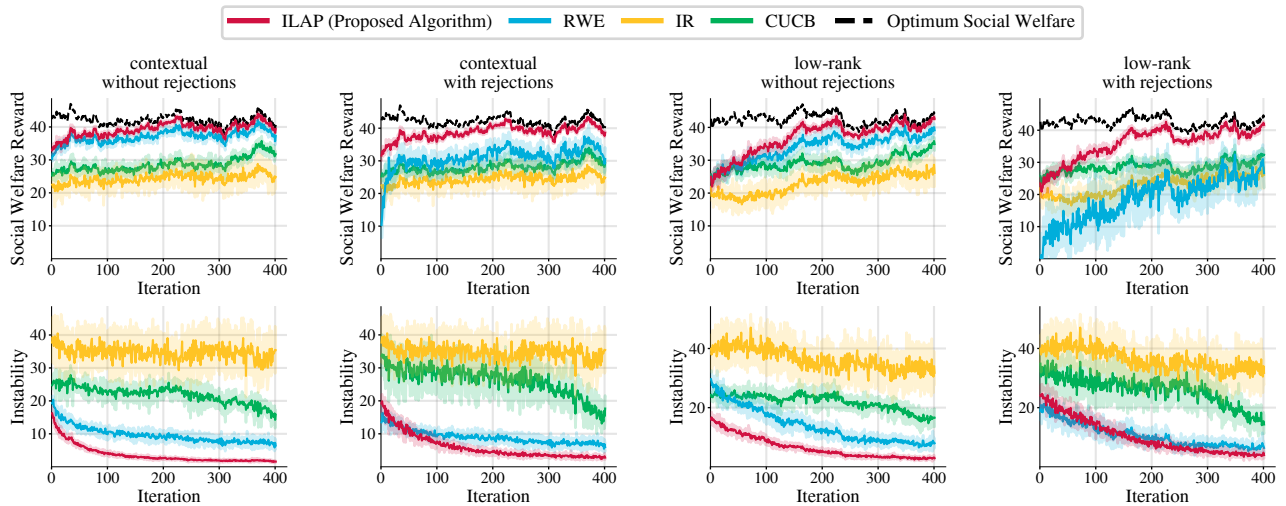

Figure 4: Experimental results for synthetic data in a dynamic setting with $N = 350$, $M = 50$, $R = 10$, probability of activity 0.2. The experiments are run on 20 problem instances and means are reported together with error regions that indicate one standard deviation of uncertainty.
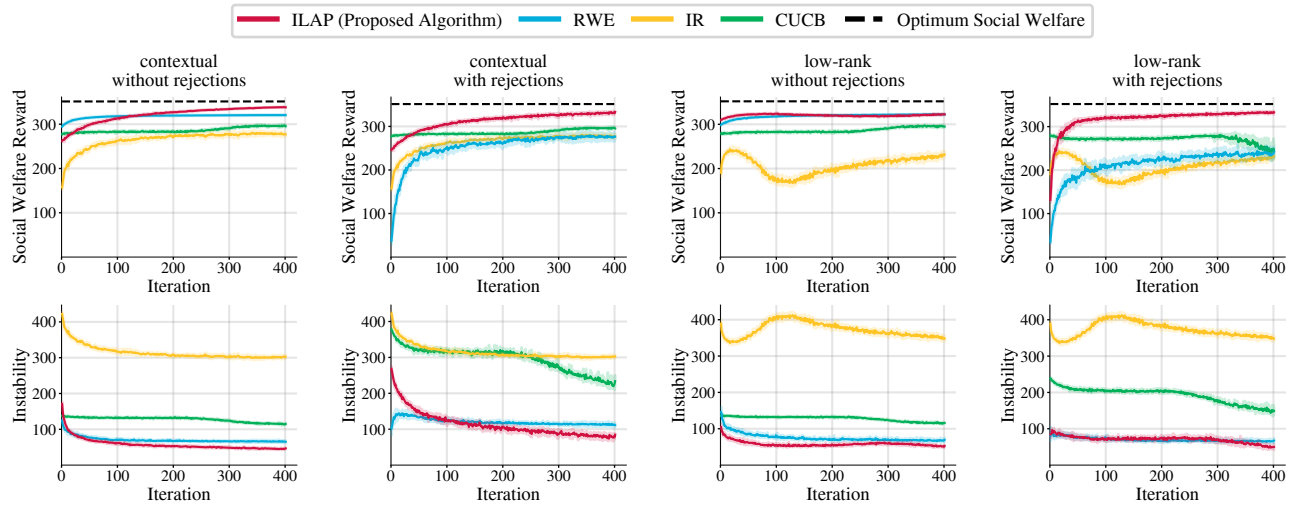
Figure 5: Experimental results for MovieLens 100k data in a static setting. The experiments are run on 20 problem instances and means are reported together with error regions that indicate one standard deviation of uncertainty.
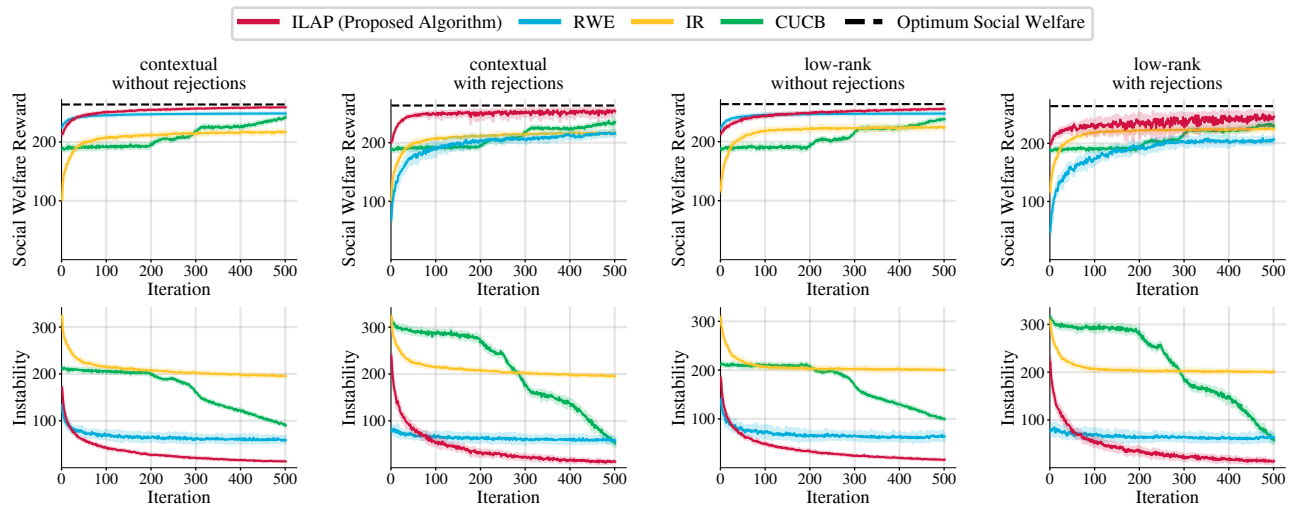

Figure 6: Experimental results for Yelp data in a static setting. The experiments are run on 20 problem instances and means are reported together with error regions that indicate one standard deviation of uncertainty.