LTP-MMF: Toward Long-Term Provider Max-Min Fairness under Recommendation Feedback Loops

CHEN XU, XIAOPENG YE, JUN XU, XIAO ZHANG, WEIRAN SHEN, and JI-RONG WEN,
Renmin University of China, Beijing, China

Multi-stakeholder recommender systems involve various roles, such as users and providers. Previous work pointed out that max-min fairness (MMF) is a better metric to support weak providers. However, when considering MMF, the features or parameters of these roles vary over time, and how to ensure long-term provider MMF has become a significant challenge. We observed that recommendation feedback loops (RFL) will influence the provider MMF greatly in the long term. RFL means that recommender systems can only receive feedback on exposed items from users and update recommender models incrementally based on this feedback. When utilizing the feedback, the recommender model will regard the unexposed items as negative. In this way, the tail provider will not get the opportunity to be exposed, and its items will always be considered negative samples. Such phenomena will become more and more serious in RFL. To alleviate the problem, this article proposes an online ranking model named Long-Term Provider Max-min Fairness (LTP-MMF). Theoretical analysis shows that the long-term regret of LTP-MMF enjoys a sub-linear bound. Experimental results on three public recommendation benchmarks demonstrated that LTP-MMF can outperform the baselines in the long term.

CCS Concepts: • **Information systems** → **Information retrieval**;

Additional Key Words and Phrases: Max-min Fairness, Provider Fairness, Recommender System

ACM Reference format:
Chen Xu, Xiaopeng Ye, Jun Xu, Xiao Zhang, Weiran Shen, and Ji-Rong Wen. 2024. LTP-MMF: Toward Long-Term Provider Max-Min Fairness under Recommendation Feedback Loops. *ACM Trans. Inf. Syst.* 43, 1, Article 11 (November 2024), 29 pages.
https://doi.org/10.1145/3695867

This work was funded by the National Key R&D Program of China (2023YFA1008704), the National Natural Science Foundation of China (No. 62376275 & No. 62106273 & No. 72192805), Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China, fund for building world-class universities (disciplines) of Renmin University of China. Supported by the Outstanding Innovative Talents Cultivation Funded Programs 2024 of Renmin University of China.
Authors' Contact Information: Chen Xu, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China; e-mail: xc_chen@ruc.edu.cn; Xiaopeng Ye, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China; e-mail: xpye@ruc.edu.cn; Jun Xu (corresponding author), Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China; e-mail: junxu@ruc.edu.cn; Xiao Zhang, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China; e-mail: zhangx89@ruc.edu.cn; Weiran Shen, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China; e-mail: shenweiran@ruc.edu.cn; Ji-Rong Wen, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China; e-mail: jrwen@ruc.edu.cn.

ACM 1558-2868/2024/11-ART11
https://doi.org/10.1145/3695867

1 Introduction

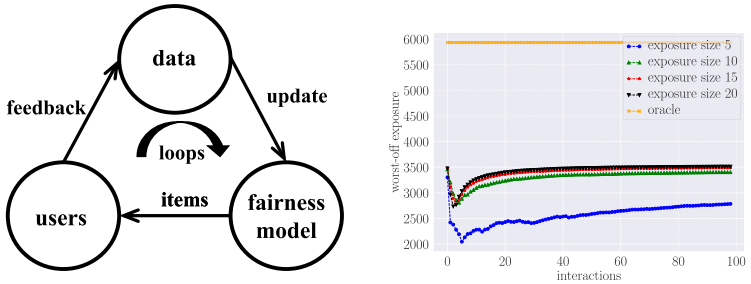
In multi-stakeholder **recommender systems (RS)**, different roles are involved, including users, providers, recommender models, etc. [2]. In recent years, provider fairness has received increasing attention, including provider demographic parity fairness [4, 22, 39, 48], proportion fairness [40, 57] and **max-min fairness (MMF)** [19, 58]. MMF aims to support worst-off providers, and it is proposed based on distributive justice concept [31, 58]. Worse-off providers, who occupy the majority of the platform, cannot survive with the necessary support. Supporting the weak providers will increase the stability of the recommender market and make a good ecosystem [20, 43]. In this article, we focused on amortized MMF [5, 9, 19, 58], where fairness accumulated across a series of rankings.

In real scenarios, the components of RS are not static. When conducting provider fairness, how to improve long-term performance under a time-changing environment is receiving increasing attention [21, 22, 38]. Previous work [21, 22, 38] aimed to conduct long-term fairness under time-vary content providers, item popularity, and provider features, respectively. Different from these studies, we observed that **recommendation feedback loops (RFL)** [14] will influence the provider MMF greatly in the long term. As shown in Figure 1(a), the users and the RS are in RFL, where the model recommends a list of items to one user in each loop. Then, the user gives feedback, and the model will be updated incrementally according to the feedback. When updating the recommendation model, items without exposures are often viewed as negative samples (known as exposure/self-selection bias [14]). Next, we will conduct an empirical study to describe how the RFL influences MMF performance.

Figure 1(b) illustrates a simulation of how the unfairness increases in the repeated interactions between RS and users. We simulated 100 interactions between 512 users and an optimal fairness model, which can always obtain the optimal objective of the tradeoff between predicting user preference and the MMF objective. The preference estimation is updated using the users' feedback at the end of each interaction. Every point in Figure 1(b) represents the long-term lowest exposures among all providers (abbreviated as Lowest Exposures) during the entire interaction process, and different lines represent the simulation performance under different exposure sizes (i.e., ranking size). The larger the exposure size becomes, the smaller providers will get more exposure opportunities [41]. From Figure 1(b), one can easily observe that the magnitude of the exposure size (i.e., the impact of exposure bias) will significantly affect the long-term performance of the MMF performance. The reason is that the provider who has the lowest exposures will not get the opportunity to be exposed in such loop because his/her items will always be considered as negative samples. In RFL, providers with few exposures will get fewer and fewer opportunities to be exposed, intensifying the unfairness over time.

Although existing studies [27, 49, 50] propose to utilize reinforcement learning frameworks to alleviate the exposure bias amplification, how to optimize the long-term objective under RFL subject to fairness constraints is still a challenging problem. In this article, we propose an online ranking model named Optimizing **Long-Term Provider Max-Min Fairness (LTP-MMF)** to address the issue. Intuitively, LTP-MMF exploits the objective while exploring the feedback of unexposed items.

Specifically, LTP-MMF formulates the provider fair recommendation problem as a repeated resource allocation problem under batched bandit settings. Each item is considered a bandit's arm. At each round, when a user comes to access the RS, LTP-MMF will choose k items (i.e., arms) to generate a recommendation ranking list. For each arm, LTP-MMF first utilizes the **Matrix Factorization (MF)** model [26] to generate the predicted accuracy rewards from the side of users. Then, LTP-MMF will get provider exposures as the fairness reward in terms of the ranking list. Finally, LTP-MMF adds an exploration term toward the accuracy and fairness rewards by utilizing the **Upper Confidence Bound (UCB)** algorithm. Such exploration terms under fairness constraints



(a) Feedback loops. (b) Simulation of feedback loops.

Fig. 1. (a) The feedback loops of the interaction between the fairness model and users. (b) Simulations of the long-term lowest exposures among all provider (abbreviated as Lowest Exposures).

help the model to access the feedback of unexposed items of small providers, improving the fairness performance in the long term.

As for the parameters of predicting rewards on the user's side, it has a serious problem of long update time. Therefore, we first collect enough users' feedback and then update them incrementally in a batched style. Theoretical analysis shows that the regret of LTP-MMF enjoys a sub-linear bound, and the batch size is a tradeoff coefficient between accuracy and fairness in the long term.

We summarize the major contributions of this article as follows:

(1) In this article, we analyze the importance of the LTP-MMF when considering the RFL.
(2) We formulate the long-term provider fair recommendation problem as a repeated resource allocation problem under a batched bandit setting, and a ranking model called LTP-MMF is proposed. Theoretical analysis shows that the regret of LTP-MMF can be bounded.
(3) Extensive experiments on three public datasets demonstrated that LTP-MMF steadily outperforms the baselines in the long term. Moreover, we verified that LTP-MMF is computationally efficient in the online inference phase.

2 Related Work

The fairness problem in multi-stakeholder RS has become a hot research topic [2, 3]. According to different stakeholders, the fairness problem can be divided into customer fairness (C-fairness) and provider fairness (P-fairness) [13, 40]. In a recent publication, the stream of provider fairness is defined as ensuring that the item exposures of each provider are relatively similar to each other [18, 23, 29, 40, 43, 44, 57, 58, 60]. In generative RS, provider fairness can also entail generating non-discriminatory item content for providers based on large language models [16, 34]. In this article, our primary focus lies on mainstream provider fairness, which emphasizes maintaining as much equality as possible in the item exposures across different providers.

When dealing with provider fairness, most methods are conducted under re-ranking scenarios [19, 23, 29, 40, 43, 57, 58, 60]. For example, FairRec [43] and its extension FairRec+ [12] proposed an offline recommender model to guarantee equal frequency for all items in a series of ranking lists. Welf [19] proposed the Frank-Welf algorithm to solve the provider fair problem. Some work [7, 23, 29, 40, 57] proposed a **Linear Programming (LP)**-based method to ensure the group fairness. P-MMF [58] and FairSync [60] proposed an online mirror gradient descent to improve the worst-off provider's exposures in the dual space. In this article, our main objective is to mitigate the influence of feedback loops in mirror gradient descent approaches [58, 60], thereby achieving long-term provider fairness.

Another line of research proposed that the recommendation system is dynamic; the attributes, features, and parameters can change as time goes on [14, 21, 22, 38]. Previous studies aimed to improve the long-term performance of fairness. Ge et al. [22] focused on the time-vary recommendation attributes and formulated the long-term problem as a Constrained Markov Decision Process. Mladenov et al. [38] proposed that content providers may leave if they cannot obtain enough support and formulated the problem as a constrained matching problem. Akpinar et al. [4] studied the provider population distributions that will change in the long term and proposed an intervention adjustment method. There are also some models [10, 37] proposed to utilize reinforcement learning techniques to optimize long-term fairness. However, these long-term-based fairness methods did not consider the impact of RFL (i.e., exposure bias) in the long term.

In recent years, many debiasing approaches [35, 62] have been proposed to break the feedback loop. For example, [62] proposed to utilize the influence function to remedy the bias from the loop, and Liu et al. [35] proposed to utilize uniform data to debias from the loop. At the same time, there are also some reinforcement learning methods to break the loop. Exploitation and exploration in bandit settings were proposed to solve the problem. Li et al. [32] and Wang et al. [52, 55] all utilized UCB to explore the unexposed items. Other models [54, 63] also utilized policy gradient methods to break the feedback loop. However, they do not consider provider MMF constraints in the feedback loop.

In the causal inference literature [24, 33, 47, 59], some methods try to alleviate exposure bias in RS. For example, some work [33, 47] suggested utilizing user-social networks to mitigate exposure bias, as users who are closely connected are more likely to share the same items with each other. Some work [24] proposed to utilize propensity scores to mitigate exposure bias. DEPS [59] considers the user-social network and item side propensity score together. However, these methods do not account for feedback loop scenarios and cannot be applied to fairness-aware RS.

3 Formulation

In this section, we first define the notations in multi-stakeholder RS. Then, we give the formal definition of amortized MMF accordingly. Finally, we formulate the interactions between users and the recommender model as a bandit problem.

3.1 Multi-Stakeholders RS

A multi-stakeholder recommender system consists of different participants, including users, item providers, etc. Let $\mathcal{U}, \mathcal{I}$, and $\mathcal{P}$ be the set of users, items, and providers, respectively. Each item $i \in \mathcal{I}$ is associated with a unique provider $p \in \mathcal{P}$. The set of items associated with a specific provider p is denoted by $\mathcal{I}_p$.

In the ranking phase, when a specific user $u \in \mathcal{U}$ accesses the recommender system, for each user-item pair (u, i), the click rate $s_{u,i} = P(c_{u,i} = 1)$ of the user u on the item i is estimated through a ranking model based on the user's historical click feedback. We denote such feedback by $c_{u,i} \in \{0, 1\}$, which represents whether or not the user clicked the item. The vector $\boldsymbol{s}_u = [s_{u,1}, s_{u,2}, \cdots, s_{u,|\mathcal{I}|}]$ is the click rates of u to all items. In the ranking phase, the items are ranked according to the estimated click rates.

In provider-fair recommendation, we should consider the user-side utilities and provider-side fairness together. Formally, we define the true user-side utility of exposing item list $L_K(u)$ to u as the summation of the preference scores in the list, denoted by $g(L_K(u)) = \sum_{i \in L_K(u)} s_{u,i}$, where $L_K(u)$ contains the K items exposed to u. Following the literature convention [40, 43, 58], we define the fairness vector as $\mathbf{e}$, for a specific provider p, $\mathbf{e}_p \in \mathbb{R}^+$ denotes the exposure of provider p. In this article, we mainly focus on provider MMF [8, 19, 58], which aims to improve the exposure

opportunities of worst-off providers. Formally, $r(\boldsymbol{e}) = \min_{p\in|\mathcal{P}|}(\boldsymbol{e}/\gamma)$, where $\boldsymbol{\gamma}$ is the weighting vector of providers.

In this article, we aim to compute a fair list $L_K^F(u) \in \mathcal{I}^K$, which well balances the user utilities $g(\cdot)$ and provider fairness metric $r(\boldsymbol{e})$ under RFL in the long-term.

3.2 Amortized Provider MMF

In real-world applications, the users arrive at the recommender system sequentially. Assume that at time t user u_t arrives. The recommender system needs to consider long-term provider exposure during the entire time horizon from $t = 0$ to T. Our task can be formulated as a resource allocation problem with amortized fairness [5, 58]. Specifically, the optimal utility of the recommender system can be defined as an amortized fairness function [5, 9, 58], which is the accumulated exposures over periods from 0 to T. In this case, $\mathbf{e}_p$ can be seen as the total number of exposed items of provider p, accumulated over the period 0 to T.

Formally, when trading off the user utilities and provider fairness, we have the following mathematical program:

$$
\begin{aligned}
R_{OPT} = \max_{\mathbf{x}_t} \quad & \frac{1}{T}\sum_{t=1}^{T} g(\boldsymbol{x}_t) + \lambda r(\boldsymbol{e}) \\
= \max_{\mathbf{x}_t} \quad & \frac{1}{T}\sum_{t=1}^{T}\sum_{i\in\mathcal{I}} \mathbf{x}_{ti} s_{u_t,i} + \lambda \min_{p\in|\mathcal{P}|}\left(\sum_{t=1}^{T} \boldsymbol{e}_t/\boldsymbol{\gamma}\right) \\
\text{s.t.} \quad & \sum_{i\in\mathcal{I}} \mathbf{x}_{ti} = K, \quad \forall t \in [1,2,\ldots,T] \\
& \mathbf{e}_{t,p} = \sum_{i\in\mathcal{I}_p} x_{ti}, \quad \forall p \in \mathcal{P},\ \forall t \in [1,2,\ldots,T] \\
& \sum_{t=1}^{T} \boldsymbol{e}_t \le \boldsymbol{\gamma}, \quad \forall t \in [1,2,\ldots,T] \\
& \mathbf{x}_{ti} \in \{0,1\}, \quad \forall i \in \mathcal{I},\ \forall t \in [1,2,\ldots,T]
\end{aligned}
\tag{1}
$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{|\mathcal{P}|}$ denotes the weights of different providers, i.e., weighted MMF [8, 58]. In amortized fairness, $\boldsymbol{e} \in \mathbb{R}^{T\times|\mathcal{P}|}$ is the exposure vector, i.e., $\boldsymbol{e}_{t,p}$ denotes the utility of provider p at time t. $\mathbf{x}_t \in \{0,1\}^{|\mathcal{I}|}$ is the decision vector for user u_t. Specifically, for each item i, $\mathbf{x}_{ti} = 1$ if it is in the fair ranking list $L_K^F(u_t)$, otherwise, $\mathbf{x}_{ti} = 0$. The MMF regularizer $\min(\cdot)$ in the objective function suggests that we should improve the exposure opportunities of worst-off providers. The first constraint in Equation (1) ensures that the recommended lists are of size K. The second constraint in Equation (1) suggests that the exposures of each provider p are the accumulated exposures of the corresponding items over all periods. In general, we think time-separable fairness would be preferred under scenarios with weak timeliness. For example, recommending items with long service life (e.g., games and clothes).

Although here we have already given an LP solution Equation (1) to the problem, it can only solve small-scale problems in an offline way. In online recommendation systems, for each user u_t access, the model needs to generate a fair ranking list $L_K^F(u_t)$ from large-scale item corpus immediately. This means we have no idea about the information after t. Next, we will discuss how to use MMF in the online recommendation problem.

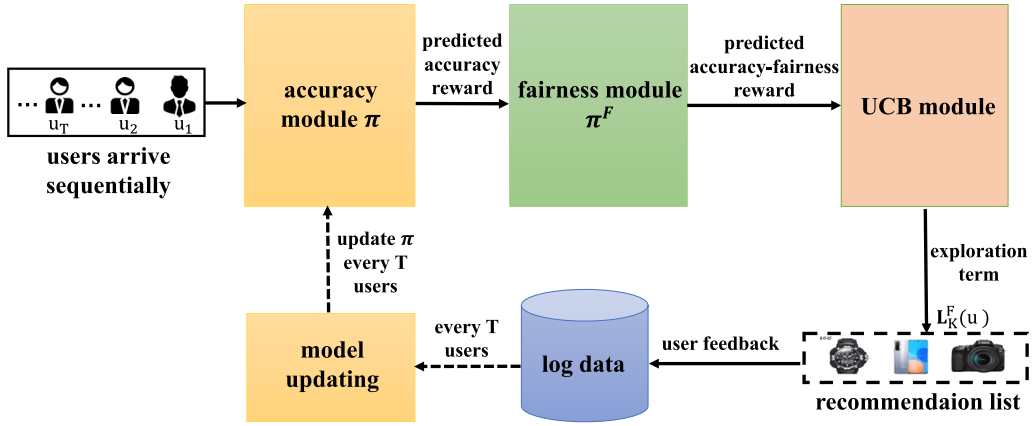


Fig. 2. Sequential item ranking process of LTP-MMF.

3.3 Bandit with Provider Fairness

We first define some notations for the problem. For any symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{M\times M}$ and vector $\boldsymbol{x} \in \mathbb{R}^M$, let $\boldsymbol{x}_i$ denote the ith element of the vector and $\boldsymbol{A}_i$ denote the ith column of the matrix $\boldsymbol{A}$. Define $\|\boldsymbol{x}\|_A = \sqrt{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}$. We also define the weighted ℓ_2 norm to be $\|\boldsymbol{x}\|^2_{\boldsymbol{y}^2} = \sum_{i=1}^M \boldsymbol{x}_i^2 \boldsymbol{y}_i^2$.

To help the fairness model break the feedback loop, we formulate the provider-fair problem as a context bandit process [52]. As shown in Figure 2, at each iteration, a batch of users $\{u_t\}_{t=1}^T$ arrive sequentially. For each user u_t and each item i, the accuracy module π takes user features and item context as input and predicts the accuracy reward $\hat{s}_{u_t,i}, \forall i \in \mathcal{I}$. Then the fairness module π^F takes the accuracy reward as input and generates a predicted accuracy-fairness reward according to the provider exposures. Finally, the exploration module gives the exploration term according to previous recommendations to explore unexposed items.

After getting the reward of each bandit arm (i.e., item), LTP-MMF chooses K items with the highest rewards to generate an item list $L_K^F(u_t)$. Then the user gives their feedback $\{c_{u_t,i}, \forall i \in L_K^F(u_t)\}$, which is stored in the log. When a batch of users finishes their access, the recommendation results and rewards are collected.

In this article, we propose to formulate such a process as a batched bandit which can be represented by a 7-tuple $< \mathcal{S}, \mathcal{L}, \pi, \pi^F, R, N, T >$:

Context space $\mathcal{S}$: denotes a hidden context space that summarizes the embedding space of both users and items.

Action space $\mathcal{L}$: denotes a given action space, and each action corresponds to selecting K items (arms): $L_K^F(u) \in \mathcal{L}$.

Accuracy module π: takes user and item context as input and predicts the accuracy reward if the user is recommended with this item, i.e., $\hat{s}_{u,i} = \pi(u, i)$.

Fairness module π^F: An online algorithm π^F produces a real-time decision vector $\mathbf{x}_t \in \{0,1\}^{|\mathcal{I}|}$ based on the current user u_t and the history $\mathcal{H}_{t-1} = \{u_s, \mathbf{x}_s\}_{s=1}^{t-1}$. π^F takes the estimated user-item preference score and item-provider relations as input and outputs an accuracy-fairness reward if the user is recommended with this item, i.e. $\hat{f}_{u,i} = \pi^F(u, \hat{s}_{u,i}, \mathcal{H}_{t-1})$.

Reward R: is defined as a linear combination of the two types of feedback defined in Equation (1): $R = \frac{1}{T} g(\boldsymbol{x}_t) + \lambda r(\boldsymbol{e})$.

Total data number N: the process of a batched fairness-aware bandit is partitioned into $\lfloor N/T \rfloor$ episodes. Within each episode, the platform first updates the ranking policy π using the collected

user feedback and then applies the re-ranking policy π^F with users for time horizon T using the updated ranking policy π.

Batch size T: is the number of steps (time horizons) in each episode. That is, in each episode, the platform makes recommendations for each user and collects log data $\mathcal{B} = \{\boldsymbol{x}_t, c_{u_t,i}, i \in L_K^F(u_t)\}, \boldsymbol{e}_t\}_{t=1}^T$. Finally, a new accuracy predicted π is trained on $\mathcal{B}$ at the beginning of the next episode.

Within each batch, the estimated reward is defined as $\hat{R} = \frac{1}{T}\hat{g}(\boldsymbol{x}_t) + \lambda r(\boldsymbol{e})$, where $\hat{g}(\cdot), r(\cdot)$ are based on the estimated user-item score $\hat{s}$. The *long-term regret* of the LTP-MMF is defined as the expectation over all log data $\mathcal{B}$:

$$\text{Regret}\left(\pi, \pi^F\right) = \mathbb{E}_{\mathcal{B}}\left[R_{OPT} - \hat{R}\right]. \tag{2}$$

4 Our Approach: LTP-MMF

In this section, we propose a novel recommender model which we call LTP-MMF. The model aims to optimize the long-term accuracy-fairness tradeoff in recommendation loops. The whole procedure can be formulated as a batched bandit process in Figure 2.

4.1 Overall Architectures

In this section, we will introduce the overall relations between the bandit settings discussed in Section 3.3. In the bandit problem, the overall reward R will be divided into three parts. In the following three sections, we employ three modules—the Accuracy Module, the Fairness Module, and the UCB module—to compute three distinct parts of the reward, respectively.

4.2 Accuracy Module: MF

For each (u, i), we can define their hidden embeddings as $\boldsymbol{v}_u, \boldsymbol{v}_i \in \mathbb{R}^d$, where d is the pre-defined dimension. The MF model takes two embeddings to determine the estimated preference score $\hat{s}_{u,i}$:

$$\hat{s}_{u,i} = \pi(u, i) = \boldsymbol{v}_u^\top \boldsymbol{v}_i + \epsilon, \tag{3}$$

where the random noise ϵ is drawn from a zero-mean Gaussian distribution $\mathcal{N}(0, \sigma^2)$.

The MF [30] model uses a coordinate descent algorithm built upon the ridge regression to estimate the unknown parameter $\boldsymbol{v}_u$ for each user and the unknown hidden feature $\boldsymbol{v}_i$ for each item. Specifically, the objective function of the ridge regression can be written as follows

$$\min_{\boldsymbol{v}_u, \boldsymbol{v}_i} \frac{1}{2}\mathbb{E}_{i,u}\left[\left(\boldsymbol{v}_u^\top \boldsymbol{v}_i - c_{u,i}\right)^2 + \frac{\lambda_u}{2}\|\boldsymbol{v}_u\|_2 + \frac{\lambda_i}{2}\|\boldsymbol{v}_i\|_2\right], \tag{4}$$

where λ_u and λ_i are the tradeoff parameters for the ℓ_2 regularization.

In real-world applications, the users arrive at the recommender system sequentially. Assume that at time t user u_t arrives, the recommender model recommendes list $L_K(u_t)$ to u_t based on estimated scores and receives the user click feedback $\{c_{u_t,i}, i \in L_K^t(u)\}$. The closed-form estimation of $\boldsymbol{v}_i, \boldsymbol{v}_u$ at time t can be derived as $\hat{\boldsymbol{v}}_{u,t} = (\boldsymbol{A}_{u,t})^{-1}\boldsymbol{b}_{u,t}$ and $\hat{\boldsymbol{v}}_{i,t} = (\boldsymbol{C}_{i,t})^{-1}\boldsymbol{d}_{i,t}$, where

$$\begin{aligned} \boldsymbol{A}_{u,t} &= \lambda_i \boldsymbol{I} + \sum_{j=1}^{t} \hat{\boldsymbol{v}}_{i,j}(\hat{\boldsymbol{v}}_{i,j})^\top, \quad \boldsymbol{b}_{u,t} = \sum_{j=1}^{t} \hat{\boldsymbol{v}}_{i,j} c_{u,i}; \\ \boldsymbol{C}_{i,t} &= \lambda_u \boldsymbol{I} + \sum_{j=1}^{t} \hat{\boldsymbol{v}}_{u,j}(\hat{\boldsymbol{v}}_{u,j})^\top, \quad \boldsymbol{d}_{i,t} = \sum_{j=1}^{t} \hat{\boldsymbol{v}}_{u,j} c_{u,i}. \end{aligned} \tag{5}$$

4.3 Fairness Module

In this section, we will take the most state-of-the-art dual gradient descent method P-MMF [58] as our fairness module. It is important to note that our methods can be readily applied to various score-based re-ranking methods [5, 40, 60], as our model LTP-MMF can easily substitute the fairness reward of the fairness module with other fairness rewards. We aim to describe how to tradeoff user accuracy and provider fairness in an online fashion. Given the ranking scores from accuracy reward $s_{u,i}$, we can consider its dual problem:

Theorem 1 (Dual Problem). *The dual problem of Equation (1) can be written as:*

$$W_{OPT} \leq W_{Dual} = \min_{\boldsymbol{\mu} \in \mathcal{D}} \left[g^*(\boldsymbol{M}\boldsymbol{\mu}) + \lambda r^*(-\boldsymbol{\mu}) \right], \tag{6}$$

where $\boldsymbol{M} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{P}|}$ is the item-provider adjacency matrix with $M_{ip} = 1$ indicating item $i \in \mathcal{I}_p$, $g^*(\cdot), r^*(\cdot)$ the conjugate functions:

$$g^*(c) = \max_{\mathbf{x}_t \in \mathcal{X}} \sum_{t=1}^{T} \left[g(\mathbf{x}_t)/T - \mathbf{c}^\top \mathbf{x}_t \right], r^*(-\boldsymbol{\mu}) = \max_{\mathbf{e} \leq \boldsymbol{\gamma}} \left[r(\mathbf{e}) + \boldsymbol{\mu}^\top \mathbf{e}/\lambda \right] \tag{7}$$

with $\mathcal{X} = \{\mathbf{x}_t \mid \mathbf{x}_t \in \{0,1\} \wedge \sum_{i \in \mathcal{I}} \mathbf{x}_{ti} = K\}$, and $\mathcal{D} = \{\boldsymbol{\mu} | r^*(-\boldsymbol{\mu}) < \infty\}$ is the feasible region of dual variable $\boldsymbol{\mu}$.

Moreover, the feasible region of the dual problem is

$$\mathcal{D}_\mu = \left\{ \boldsymbol{\mu} \middle| \sum_{p \in \mathcal{S}} \boldsymbol{\gamma}_p \boldsymbol{\mu}_p \geq -\lambda, \forall \mathcal{S} \in \mathcal{P}_s \right\}, \tag{8}$$

where $\mathcal{P}_s$ is the power set of $\mathcal{P}$, i.e., the set of all the subsets of $\mathcal{P}$.

Lemma 1. *The conjugate function $r^*(\cdot)$ has a closed form:*

$$\max_{\boldsymbol{\mu} \leq \boldsymbol{\gamma}} r^*(-\boldsymbol{\mu}) = \boldsymbol{\gamma}^T \boldsymbol{\mu}/\lambda + 1,$$

and the optimal dual variable is:

$$\arg\max_{\boldsymbol{\mu} \leq \boldsymbol{\gamma}} r^*(-\boldsymbol{\mu}) = \boldsymbol{\gamma}/\lambda.$$

Remark 1. Following the practice in [58], we also try to optimize the integral LP from the dual space. It involves a vast variable $\boldsymbol{x}$'s space of size $T \times |\mathcal{I}|$ to dual variable μ's size $|\mathcal{P}| \ll T \times |\mathcal{I}|$, and thanks to the sparsity of $\mathbf{M}$, the computation of $\mathbf{M}\boldsymbol{\mu}$ is highly efficient. Note that any other dual transformation way [5, 60] can be also applied in LTP-MMF.

The proofs of Theorem 1 and Lemma 1 are provided in Appendix A.1. From Theorem 1, we can have a new non-integral decision variable $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{P}|}$. In practice, we usually have $|\mathcal{P}| \ll |\mathcal{I}|$. Besides, due to $\boldsymbol{M}$'s sparsity, it is very efficient to compute $\boldsymbol{M}\boldsymbol{\mu}$, which aims to project the variable $\boldsymbol{\mu}$ from the provider space onto the item spaces. According to Theorem 1, the accuracy-fairness reward of a user-item pair is obtained through the conjunction function $g^*(\boldsymbol{M}\boldsymbol{\mu})$: $\hat{f}_{u,i} = \hat{s}_{u,i} - \boldsymbol{M}_i^\top \boldsymbol{\mu}$.

For each time step t in a batch, we utilize the mirror momentum gradient descent [5, 45] to learn the dual variable $\boldsymbol{\mu}_t$: first, we evaluate the conjugate function of the max-min regularizer $r^*(-\boldsymbol{\mu}_t)$ according to Lemma 1. Then, we can get the subgradient of the estimated dual function as

$$-\mathbf{M}^\top \mathbf{x}_t + \mathbf{e}_t \in \partial \left(g^*(\boldsymbol{M}\boldsymbol{\mu}_t) + \lambda r^*(-\boldsymbol{\mu}_t) \right).$$

Finally, we utilize $\mathbf{g}_t = \boldsymbol{M}\mathbf{x}_t + \boldsymbol{e}_t$ to update the dual variable by performing the online descent. Therefore, the dual variable will move toward the directions of the providers with fewer exposures, and the primal variable $\mathbf{x}_t$ will move to a better solution.

Note that the projection step can be efficiently solved using QP solvers [5] since $\mathcal{D}$ is coordinate-wisely symmetric.

$$\begin{aligned}\boldsymbol{\mu}_{t+1} &= \arg\min_{\boldsymbol{\mu}} \sum_{p\in\mathcal{P}} (\boldsymbol{\mu}_p \boldsymbol{\gamma}_p - \widetilde{\boldsymbol{\mu}}_{t,p}\boldsymbol{\gamma}_p)^2 \\ &\text{s.t.} \sum_{j=1}^{m} \boldsymbol{\gamma}_j \widetilde{\boldsymbol{\mu}}_j + \lambda \geq 0, \forall m = 1, 2, \ldots, |\mathcal{P}|, \end{aligned} \tag{9}$$

where $\widetilde{\boldsymbol{\mu}}$ satisfies:

$$\boldsymbol{\gamma}_1\widetilde{\boldsymbol{\mu}}_1 \leq \boldsymbol{\gamma}_2\widetilde{\boldsymbol{\mu}}_2 \leq \cdots, \leq \boldsymbol{\gamma}_{|\mathcal{P}|}\widetilde{\boldsymbol{\mu}}_{|\mathcal{P}|}.$$

4.4 UCB for Accuracy-Fairness Reward

UCB [52, 53] has proven to be an effective strategy to estimate the confidence of predicted rewards during exploration. In fairness-aware recommendations, the system should not only exploit the accuracy-fairness objective in the ranking process but also should explore less exposed items to avoid giving too few exposures to some providers.

Next, we will bound the corresponding UCB of the predicted accuracy-fairness reward $\hat{f}_{u,i}$.

Theorem 2 (Confidence Radius of Ranking). *The parameter $\boldsymbol{v}_u, \boldsymbol{v}_i$ is q-linear to the optimizer. For any $\sigma > 0$, with probability at least $1-\sigma$, the confidence radius $\triangle f^t_{u,i}$ of user-item preference score at time t satisfies:*

$$f_{u,i} - \hat{f}_{u_t,i} \leq \triangle f^t_{u,i} = \alpha_t(\|\hat{\boldsymbol{v}}_{i,t}\|_{\boldsymbol{A}^{-1}_{u,t}} + C_t/2) + \beta_t(\|\hat{\boldsymbol{v}}_{u,t}\|_{C^{-1}_{i,t}} + C_t/2). \tag{10}$$

We can decompose the confidence radius into three terms:

- *$\boldsymbol{v}_u$ and $\boldsymbol{v}_i$ bias terms $\|\hat{\boldsymbol{v}}_{u,t} - v^*_u\|_{\boldsymbol{A}_{u,t}}, \|\hat{\boldsymbol{v}}_{i,t} - v^*_i\|_{C_{i,t}}$ and their upper bound at time t are denoted by $\alpha_t, \beta_{i,t}$, respectively.*
- *$\boldsymbol{v}_u$ and $\boldsymbol{v}_i$ variance terms $\|\hat{\boldsymbol{v}}_{u,t}\|_{\boldsymbol{A}^{-1}_{u,t}}, \|\hat{\boldsymbol{v}}_{i,t}\|_{C^{-1}_{i,t}}$.*
- *collaborative variance terms $\|\hat{\boldsymbol{v}}_{u,t} - v^*_u\|_{\boldsymbol{A}^{-1}_{u,t}}, \|\hat{\boldsymbol{v}}_{i,t} - v^*_i\|_{C^{-1}_{i,t}}$ and their upper bound at time t are defined as $C_t = (q+\epsilon_q)^t$, where for any $\epsilon_q > 0$ and $\boldsymbol{v}_u, \boldsymbol{v}_i$ are q−linear ($0 < q < 1$ to the optimizer.*

The bias bound (i.e., exploration weight) α_u, α_i are defined as

$$\begin{aligned} &\|\hat{\boldsymbol{v}}_{u,t} - \boldsymbol{v}^*_u\|_{\boldsymbol{A}_{u,t}} \leq \alpha_t, \quad \|\hat{\boldsymbol{v}}_{i,t} - \boldsymbol{v}^*_i\|_{C_{i,t}} \leq \beta t, \\ &\alpha_t = \sqrt{\lambda_u} + \frac{2(q+\epsilon_q)(1-(q+\epsilon_q)^t)}{1-q-\epsilon_q} + \sqrt{d\ln\frac{\lambda_u d + t}{\lambda_u d\sigma}}, \\ &\beta_t = \sqrt{\lambda_i} + \frac{2(q+\epsilon_q)(1-(q+\epsilon_q)^t)}{1-q-\epsilon_q} + \sqrt{d\ln\frac{\lambda_i d + t}{\lambda_i d\sigma}}. \end{aligned} \tag{11}$$

The confidence radius can be bounded by $O\left(\frac{[1-(q+\epsilon_r)^t]\sqrt{\ln(t)}}{\sqrt{t}}\right)$, which is a sub-linear decreasing function of t when t becomes large. $O((1-(q+\epsilon_r)^t)\sqrt{\ln(t)})$ is the bias growth rate and $O(1/\sqrt{t})$ is the variance converge rate .

Remark 2. From Theorem 2, we can see the exploration term is $\triangle f_{u,i}^t = \alpha_t(\|\hat{\boldsymbol{v}}_{i,t}\|_{A_{u,t}^{-1}} + C_t/2) + \beta_t(\|\hat{\boldsymbol{v}}_{u,t}\|_{C_{i,t}^{-1}} + C_t/2)$. The exploration term has three parts:

- —Bias terms $\|\hat{\boldsymbol{v}}_{u,t} - v_u^*\|_{A_{u,t}}, \|\hat{\boldsymbol{v}}_{i,t} - v_i^*\|_{C_{i,t}}$ is bounded by α_t, β_t, which captures the user preference shift with fairness. Intuitively, the presence of more shifts indicates that previous estimations may not be accurate, thus resulting in a larger degree of exploration.
- —Variance terms $\|\hat{\boldsymbol{v}}_{u,t}\|_{A_{u,t}^{-1}}, \|\hat{\boldsymbol{v}}_{i,t}\|_{C_{i,t}^{-1}}$ captures the variance of user and item embedding estimation. Intuitively, the presence of more variance indicates that previous estimations may not be accurate, thus resulting in a larger degree of exploration.
- —Similarly, the collaborate variance term C_t captures the variance of user-item interaction. Intuitively, the presence of more variance indicates that previous estimations may not be accurate, thus resulting in a larger degree of exploration.

Those three parts together form the exploration term to balance exploration in fairness constraints with accuracy.

The proof of Theorem 2 is deferred to Appendix A.3. Equation (10) measures the estimation uncertainty of parameter $\boldsymbol{v}_u, \boldsymbol{v}_i$ in ranking module. The exploration and exploitation tradeoff is balanced by the prediction confidence bound of both the user side and the item side. From Theorem 2, we can also observe that the confidence radius is a tradeoff between the bias term and the variance term. With more observations, the bias will increase but the variance will decrease more rapidly so that the radius bound will become smaller (i.e., the bound is a sub-linear decreasing function of t when t becomes large).

4.5 Our Algorithm: LTP-MMF

In each iteration of the algorithm, we first conduct online recommendations of user accuracy and provider MMF with fixed ranking model for a batch of users $\{u_t\}_{t=1}^T$. Then we will collect these users' feedback on the recommended items. Finally, the accuracy module parameters are updated using the collected users' feedback.

Algorithm 1 illustrates our main algorithm.

In lines 6–11, the algorithm describes the process of fair-aware RS model giving recommendation results to users (fairness model->user edge in Figure 1). Specifically, when user u comes to the recommender system, in line 6, LTP-MMF first estimates the accuracy score (accuracy Module). Then in line 7, the LTP-MMF computes the UCB bound (UCB module) in line 10, we will also add the final score with the fairness term (fairness module).

These three modules together form the total reward: (1) For the error from the ranking model, we modify the ranking score according to its UCB. (2) For the MMF objective, we apply the primal-dual theory to adjust the score with the dual variable $\boldsymbol{\mu}_t$. Intuitively, for $\boldsymbol{\mu}_t$, when the values of dual variables are higher, the algorithm naturally recommends fewer items related to the corresponding provider. (3) $\boldsymbol{m}$ ensures that the algorithm only recommends items from providers with remaining resources. Note that in line 9, the formulation is linear with respect to $\mathbf{x}_t$. Therefore, it is efficient to compute $\mathbf{x}_t$ through a top-K sort algorithm in constant time. Finally, in line 12, we will adapt the top-K sorting operation to get the final output variable $\boldsymbol{x}$ according to the reward combined with these three components.

In line 14, the algorithm outlines the procedure by which users generate data for model training, specifically focusing on the user->data edge as depicted in Figure 1.

In lines 15–28, the algorithm outlines the process of model updating (data->fairness model edge in Figure 1). In lines 17–21, we aim to update the parameters of the fairness module utilizing the dual mirror descent algorithm. In lines 23–29, we utilize the user behavior data to update the parameters of the accuracy module.

Algorithm 1: Online Learning of LTP-MMF

Input: User arriving order $\{u_i\}_{i=1}^N$, batch size T, ranking size K. **Ranking parameters:** User embedding $\{\boldsymbol{v}_u \in \mathbb{R}^d, \forall u \in \mathcal{U}\}$, item embedding $\{\boldsymbol{v}_i \in \mathbb{R}^d, \forall i \in \mathcal{I}\}$. Confidence level σ, and semi-positive matrix $\boldsymbol{A}_u = \lambda_u \boldsymbol{I}, \boldsymbol{b}_u = \boldsymbol{0}, \boldsymbol{C}_i = \lambda_i \boldsymbol{I}, \boldsymbol{d}_i = \boldsymbol{0}$.
Re-ranking parameters: Item-provider adjacent matrix $\mathbf{M}$, maximum resources $\boldsymbol{\gamma}$ and the coefficient λ.

Output: The decision variables $\{\mathbf{x}_i, i = 1, 2, \ldots, N\}$.

for $n = 1, \cdots, \lfloor N/T \rfloor$ **do**
  Initialize dual solution $\boldsymbol{\mu}_1 = 0$, remaining resources $\boldsymbol{\beta}_1 = \boldsymbol{\gamma}$, and momentum gradient $\mathbf{g}_0 = 0$, training buffer $\mathcal{B} = \emptyset$.
  for $t = 1, \cdots, T$ **do**
    User u_{nT+t} arrives, we abbreviate u_{nT+t} as u.
    fairness model->users edge in Figure 1(a): lines 6-11.
    Compute estimated score $s_{u,i} = \boldsymbol{v}_u^\top \boldsymbol{v}_i, \forall i \in \mathcal{I}$.
    Compute the UCB $\triangle \boldsymbol{s}$ following Equation (10): $\triangle f_{u,i}^n = \alpha_n(\|v_i\|_{A_u^{-1}} + C/2) + \beta_n(\|v_u\|_{C_i^{-1}} + C/2)$.
    $\mathbf{m}_p = (1 - \boldsymbol{I}(\boldsymbol{\beta}_{tp} > 0) * 1000.0)$
    //Rewards for exploit-explore trade-off:
    $r_{u,i} = \hat{s}_{u,i}/T - \mathbf{M}_i^\top(\boldsymbol{\mu}_t + \mathbf{m}) + \triangle f_{u,i}^n$
    yields output variable $\boldsymbol{x}$
    $\mathbf{x}_t = \arg\max_{\mathbf{x}_t \in \mathcal{X}} \left[\boldsymbol{r}_u^\top \boldsymbol{x}_t\right]$
    users->data edge in Figure 1(a): lines 13-20.
    $\mathcal{B} = \mathcal{B} \cup \{(u, i, c_{u,i})$ //Receive user's feedback:
    //Update the remaining resources:
    $\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \boldsymbol{M}^\top \mathbf{x}_t, \mathbf{e}_t = \arg\max_{\mathbf{e}_t \le \boldsymbol{\beta}_t} r^*(-\boldsymbol{\mu}_t)$
    $\widetilde{\mathbf{g}}_t = -\mathbf{M}^\top \mathbf{x}_t + \mathbf{e}_t,$
    $\mathbf{g}_t = \alpha \widetilde{\mathbf{g}}_t + (1 - \alpha)\mathbf{g}_{t-1}$
    //Update the dual variable by gradient descent
    $\boldsymbol{\mu}_{t+1} = \arg\min_{\boldsymbol{\mu} \in \mathcal{D}} \left[\langle \mathbf{g}_t, \boldsymbol{\mu} \rangle + \eta \|\boldsymbol{\mu} - \boldsymbol{\mu}_t\|_{\boldsymbol{\gamma}^2}^2\right]$
  end for
  data->fairness model edge in Figure 1(a) lines 22-27.
  for $(u, i, c_{u,i}) \in \mathcal{B}$ **do**
    //Update accuracy module:
    $\boldsymbol{A}_u = \boldsymbol{A}_u + \boldsymbol{v}_i \boldsymbol{v}_i^\top, \boldsymbol{C}_i = \boldsymbol{C}_i + \boldsymbol{v}_u \boldsymbol{v}_u^\top$
    $\boldsymbol{v}_u = \boldsymbol{A}_u^{-1} \boldsymbol{b}_u, \boldsymbol{v}_u = \boldsymbol{v}_u / \|\boldsymbol{v}_u\|_2$
    $\boldsymbol{b}_u = \boldsymbol{b}_u + \boldsymbol{v}_i c_{u,i}, \boldsymbol{d}_i = \boldsymbol{d}_i + \boldsymbol{v}_u c_{u,i}$
    $\boldsymbol{v}_i = \boldsymbol{C}_i^{-1} \boldsymbol{d}_i, \boldsymbol{v}_i = \boldsymbol{v}_i / \|\boldsymbol{v}_i\|_2$
  end for
end for

Theorem 3 (Regret Bound). *Assume that the function $\|\cdot\|_{\boldsymbol{\gamma}^2}^2$ is σ-strong convex and there exists a constant $G, L \in \mathbb{R}^+$ such that $\|\widetilde{g}_t\| < G, \|\boldsymbol{\mu}\|_\gamma \le L$. Then, the regret can be bounded as follows:*

$$Regret(\pi, \pi^F) \le Regret(\pi^F) + Regret(\pi)L, \tag{12}$$

where the regret bound of fairness module $Regret(\pi^F)$ is the regret of the dual fairness model, which can be bounded as

$$Regret(\pi^F) \le \frac{N}{T}\left[\frac{K(1 + \lambda\bar{r} + \bar{r})}{\min_p \boldsymbol{\gamma}_p} + \frac{L^2}{\eta} + \frac{G^2\eta(T-1)}{(1-\alpha)\sigma} + \frac{G^2}{2(1-\alpha)^2\sigma\eta}\right], \tag{13}$$

where $\bar{r}$ is the upper bound of MMF regularzier, and in practice, $\bar{r} \leq 1$. The regret bound of accuracy module Regret(π) can be bounded as

$$Regret(\pi) \leq KT[\alpha_{\lfloor N/T \rfloor}(\rho_u + \kappa) + \beta_{\lfloor N/T \rfloor}(\rho_i + \kappa)], \tag{14}$$

where

$$\kappa = \frac{(q+\epsilon_q)(1-(q+\epsilon_q)^{\lfloor N/T \rfloor})}{1-q-\epsilon_q},$$

$$\rho_l = \sqrt{2d\frac{N}{T}\ln\left(1+\frac{N}{T\lambda_l d}\right)}, l = u, i.$$

Remark 3 (Accuracy-Fairness Regret Trade-off). Setting the learning rate as $\eta = O(T^{-1/2})$, we can obtain a fairness regret Regret(π^F) upper bound of order $O(N/\sqrt{T})$. The accuracy regret Regret(π) is $O(\sqrt{NT\ln(\frac{N}{T})})$. The larger batch size T, the less error raised of accuracy in the long term, but the more bias caused by the fairness module π^F.

Remark 4 (Sublinear Long-Term Regret). Overall, the long-term regret of LTP-MMF can be obtained from order $O(N \ln N)$. Moreover, we can observe that the more we tend to consider fairness, the larger L will become, leading to larger regret in the long term.

4.6 Discussion

We will summarize the key contribution of our method as follows. Firstly, our chosen accuracy module is the widely represented two-tower model of RS [61], which utilizes the dot product between two learned user item embeddings. However, it can be readily replaced with any other two-tower recommender system models [25, 46, 51] by simply substituting the learned embeddings for the user-item embeddings v_i and v_u. Secondly, as mentioned in the Fairness Module, the fairness module can also be replaced by any score-based re-ranking methods [5, 40, 61]. Secondly, Our UCB exploration term exhibits wide adaptability, and our theoretical analysis also provides guidance for the application of other two-tower and score-based fairness models. Our method uniquely integrates the UCB technique to effectively balance accuracy estimation and fairness optimization within the feedback loop scenarios of RS, offering the literature a notable example of addressing the accuracy-fairness tradeoff in such long-term contexts.

5 Experiment

We conducted experiments to show the effectiveness of the proposed LTP-MMF in the long term. The source code and experiments have been shared at GitHub https://github.com/XuChen0427/LTP-MMF.

5.1 Experimental Settings

5.1.1 Datasets. The experiments were conducted on four large-scale, publicly available recommendation datasets, including:

Yelp:[1] a large-scale businesses recommendation dataset. We only utilized the clicked data, which is simulated as the 4–5 star rating samples.

[1] https://www.yelp.com/dataset

Table 1. Statistics of the Datasets

Dataset	#User	#Item	#Provider	#Interaction	Sparsity	Provider Removing Ratio
Yelp	17,034	11,821	23	154,543	99.92%	0.0%
Amazon-Beauty	9,625	2,756	104	49,217	99.81%	50.55%
Amazon-Baby	11,680	2,687	112	59,836	99.80%	60.94%
Steam	5,902	591	81	29,530	99.15%	25.06%

Amazon-Beauty/Amazon-Baby: Two subsets (beauty and digital music domains) of Amazon Product dataset.[2] We only utilized the clicked data, which is simulated as the 4–5 star rating samples. Also, the brands are considered as providers.

Steam[3] [28]: We used the data for games played for more than 10 hours in our experiments. The publishers of games are considered as providers.

As a pre-processing step, the users, items, and providers who interacted with less than five items/users were removed from all datasets to avoid the extremely sparse cases. We removed providers associated with fewer than five items, which also included providers lacking a brand name, resulting in the removal of their associated items as well. We will give a more detailed provider-removing ratio in Table 1. Table 1 lists some statistics of the four datasets.

5.1.2 Evaluation. We sorted all the interactions according to the time and used the first 80% of the interactions as the training data to train the base model (i.e., BPR [46]). The remaining 20% of interactions were used as the test data for evaluation. Based on the trained base model, we can obtain a preference score $s_{u,i}$ for each user-item pair (u, i). The chronological interactions in the test data were split into interaction sequences where the horizon length was set to T. We calculated the metrics separately for each sequence, and the averaged results are reported as the final performances.

As for the evaluation metrics, the performances of the models were evaluated from three aspects: user-side preference, provider-side fairness, and the tradeoff between them. As for the user-side accuracy, following the practices in [57], we utilized the CTR@K, which measures the averaged click rate in the long term:

$$\text{CTR@K} = \frac{1}{N}\sum_{t=1}^{N}\sum_{i\in \mathbf{L}_K^F(u_t)} s_{u_t,i}/K. \tag{15}$$

As for provider fairness, we directly utilized the definition of MMF in Section 3 as the metric:

$$\text{MMF@}K = \frac{T}{N}\sum_{i=1}^{\lfloor N/T\rfloor}\min_{p\in\mathcal{P}}\left\{\sum_{t=1}^{T}\sum_{i\in \mathbf{L}_K^F(u_t)} \mathbb{I}(i\in\mathcal{I}_p)/\boldsymbol{\gamma}_p\right\}, \tag{16}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

As for the tradeoff performance, we used the online objective tradeoff to measure the fairness:

$$r_\lambda@K = \text{CTR@K} + \lambda\cdot\text{MMF@}K, \tag{17}$$

where $\lambda \geq 0$ is the tradeoff coefficient.

[2] http://jmcauley.ucsd.edu/data/amazon/
[3] http://cseweb.ucsd.edu/wckang/Steam_games.json.gz

5.1.3 Baselines. The following representative provider fair re-ranking models were chosen as the baselines: FairRec [43] and FairRec+ [12] aimed to guarantee at least Maximin Share of the provider exposures. CPFair [40] formulated the tradeoff problem as a knapsack problem and proposed a greedy solution.

We also chose the following MMF models: Welf [19] use the Frank-Wolfe algorithm to maximize the Welfare functions of worst-off items. However, it is developed under off-line settings; RAOP [5] is a state-of-the-art online resource allocation method. We applied it to the recommendation by regarding the items as the resources and users as the demanders. Fairco [39]: added a regularizer that measures the exposure gaps between the target provider and the worst providers.

We also compared the proposed with one heuristic MMF baseline: K-neighbor: at each time step t, only the items associated to the top-K providers with the least cumulative exposure are recommended. At the same time, we also compared three bandit baselines, which aim to remedy the bias from the feedback loops of RS. Note that, to make a fair comparison, all the bandit algorithms are updated based on the batched log. hLinUCB [52]: it utilized UCB for learning hidden features for both users and items. EXP3 [11] aimed to choose actions according to a distribution constructed by the exponential weights. BLTS [17] will sample a reward based on a normal distribution with the estimated reward variance. However, these bandit baselines did not consider provider fairness in the long term.

Meanwhile, we also compare the baseline Dyn [4] of optimizing long-term fairness in feedback loops. However, Dyn addresses long-term fairness concerns by considering the dynamic nature of user social networks, where changes in social connections may lead users to belong to different social groups over time. Therefore, we apply their re-ranking methods to our framework by treating specific user groups as providers and users within those groups as items belonging to the respective provider. To ensure a fair comparison, we substitute their dynamic environment, which involved changing social networks, with dynamic user-item score estimation.

5.1.4 Implementation Details. As for the hyper-parameters in all models, the learning rate was tuned among $[1e-2, 1e-3]/T^{1/2}$, and the momentum coefficient α was tuned among $[0.2, 0.5]$. For the maximum resources (i.e., the weights) $\boldsymbol{\gamma}$, following the practices in [57, 58], we set $\boldsymbol{\gamma}$ based on the number of items provided by the providers:

$$\boldsymbol{\gamma}_p = KT\eta|\mathcal{I}_p|\big/|\mathcal{I}|, \tag{18}$$

where η is the factor controlling the richness of resources. In all the experiments, we set $\eta = 1+1/|\mathcal{P}|$. We implemented LTP-MMF with PyTorch [42]. The experiments were conducted with a single NVIDIA GeForce RTX 3090.

5.2 Main Experiments

In this section, we conducted the experiments on four large-scale datasets with the LTP-MMF and other baselines.

Due to variations in dataset characteristics, we will provide a comprehensive list of critical parameters related to fairness performance ranges for our model LTP-MMF and other baselines in Table 2. To make a fair comparison, the other important variables such as γ for evaluation and the learning rate are all listed in Section 5.1.4 in the revised version.

5.2.1 Overall Performance. Table 3 reports the experimental results of LTP-MMF and the baselines on all three datasets to investigate the long-term performance. Underlined numbers mean the best-performed baseline and $*$ means the improvements over the best baseline are statistically significant (t-test, p-value< 0.05). To make fair comparisons, all the baselines were tuned and used

Table 2. List of Critical Parameters Related to Fairness Is in Table 3

Models	Fairness-aware Parameters
k-neighbor	$k \in [1, 5]$
P-MMF	$\lambda = 0.5$
CPFair	$\lambda = 0.5$
Fairco	$\lambda = 0.5$
FairRec/FairRec+	$\alpha \in [0.5, 1]$
Welf	$\lambda = 0.5$
Dyn	$\lambda = 0.5$
LTP-MMF (ours)	$\lambda = 0.5, q = 0.8, \lambda_u = \lambda_i = 1$

To ensure a fair comparison, if the baseline includes a linear tradeoff coefficient, we uniformly set $\lambda = 0.5$. For other parameters, we tuned them within the following ranges for each dataset.

$r@K$ as the evaluation metric. Note that similar experiment phenomena have also been observed on other λ values.

From the reported results, we found that LTP-MMF outperformed all of the bandit baselines, including hLinUCB, EXP3, and BLTS, which verified that LTP-MMF can serve the provider fairness well in the long term. We also observed that LTP-MMF outperformed all the MMF-based baselines, indicating that LTP-MMF can better remedy the bias in the RFL compared to other provider-fair baselines. The reason is that LTP-MMF can explore the feedback of unexposed items, avoiding always treating unexposed items as negative ones.

Due to exposure bias, larger exposure sizes tend to decrease bias, as they offer more items to be exposed to users. However, in modern RS, such as YouTube [15], the ranking size exposed to users is typically limited to single-digit items. Moreover, in the example depicted in Figure 1, we can observe that as the ranking size increases, the severity of exposure bias diminishes. From the empirical studies presented in Table 2, we can observe that the average improvements over the best baselines are 5.5%, 2.7%, and 1.8% for ranking sizes of 5, 10, and 20, respectively. This means that the smaller the magnitude of the exposure size (i.e., the larger the exposure bias in RFL), the more improvements LTP-MMF increases. It indicates the effectiveness of LTP-MMF especially when the ranking size is small.

Finally, it is evident that our model LTP-MMF significantly outperforms the long-term fairness baseline Dyn, underscoring the effectiveness of our methods in optimizing long-term fairness amidst the changing user preference estimations influenced by the feedback loops of RS.

5.2.2 Pareto Frontier. Figure 3 shows the Pareto frontiers [36] of CTR@K and MMF@K. The Pareto frontiers were drawn by tuning the accuracy-fairness tradeoff coefficient $\lambda \in [1e-3, 1]$ with best (CTR@K, MMF@K) long-term performances. In the experiment, we selected the baselines of P-MMF, CPFair, fairco, Welf, FairRec+, and BLTS, which achieved relatively good performances among fairness baselines and bandit baselines.

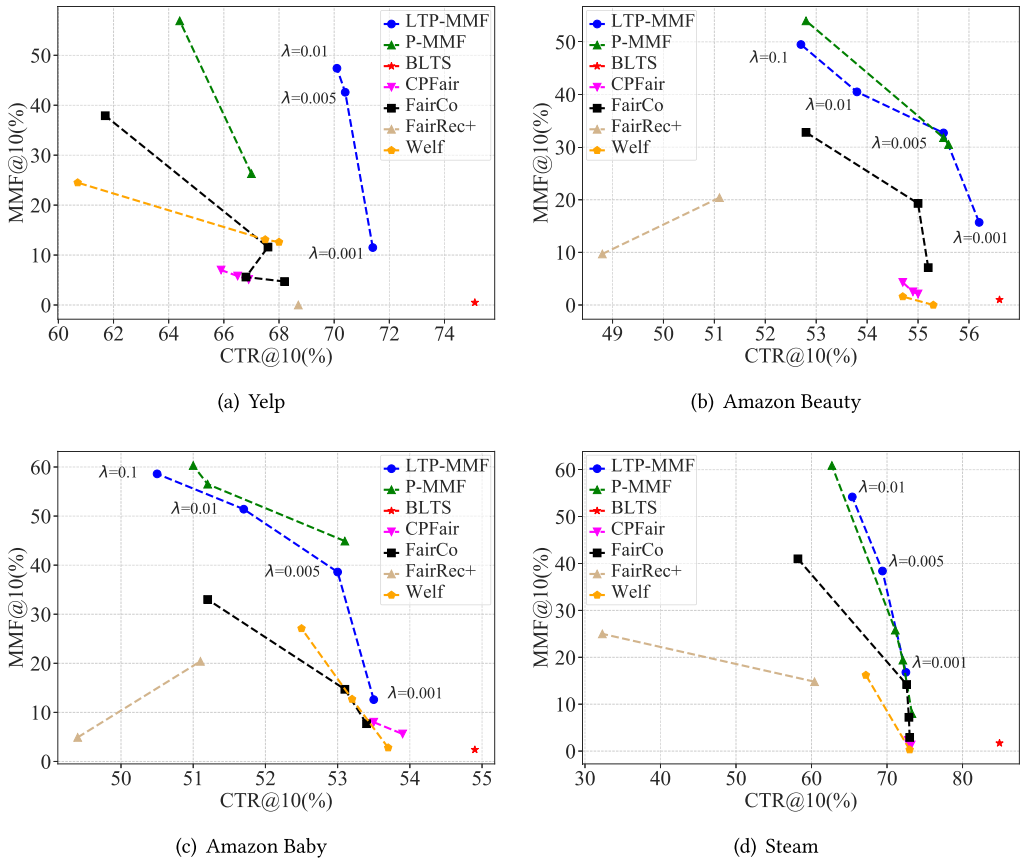
From the Pareto frontiers, we can see that the proposed LTP-MMF Pareto dominated all the baselines (i.e., the LTP-MMF curves are at the upper right corner) expect P-MMF [58] on the Amazon-Baby and Amazon-Beauty, indicating that LTP-MMF can achieve better user accuracy (i.e., CTR@K) with the same provider fairness (MMF@K) level. The results demonstrate that LTP-MMF can better tradeoff accuracy and fairness in the long term.

Table 3. Performance Comparisons between LTP-MMF and the Baselines

Model	User Accuracy(CTR@K)			Provider Fairness(MMF@K)			**Tradeoff Performance(r@K)***		
	CTR@5	*CTR*@10	*CTR*@20	*MMF*@5	*MMF*@10	*MMF*@20	*r*@5	*r*@10	*r*@20
Yelp									
hLinUCB	0.685	0.654	0.661	0.000	0.000	0.000	0.685	0.654	0.661
EXP3	0.463	0.465	0.466	0.014	0.022	0.028	0.470	0.466	0.480
BLTS	0.703	0.751	0.770	0.002	0.005	0.005	0.703	0.753	0.773
FairRec	0.710	0.687	0.762	0.000	0.000	0.000	0.710	0.687	0.762
FairRec+	0.717	0.802	0.762	0.000	0.000	0.000	0.717	0.802	0.762
CPFair	0.625	0.659	0.709	0.071	0.070	0.109	0.661	0.694	0.764
k-neighbor	0.441	0.444	0.465	0.145	0.145	0.143	0.514	0.517	0.537
Welf	0.556	0.577	0.607	0.158	0.160	0.182	0.635	0.657	0.698
Fairco	0.569	0.587	0.606	0.271	0.278	0.283	0.705	0.726	0.748
P-MMF	0.581	0.611	0.648	0.225	0.314	0.390	0.694	0.768	0.843
Dyn	0.507	0.504	0.464	0.252	0.134	0.001	0.663	0.571	0.4645
LTP-MMF (Ours)	**0.621**	**0.665**	**0.704**	**0.386**	**0.363**	**0.327**	**0.814***	**0.847***	**0.868***
Improv.							13.5%	5.6%	3.0%
Amazon-Beauty									
hLinUCB	0.588	0.601	0.565	0.000	0.000	0.000	0.588	0.601	0.565
EXP3	0.455	0.455	0.456	0.004	0.076	0.135	0.457	0.493	0.524
BLTS	0.559	0.566	0.580	0.000	0.010	0.026	0.559	0.571	0.593
FairRec	0.556	0.575	0.557	0.000	0.000	0.110	0.556	0.575	0.612
FairRec+	0.556	0.575	0.559	0.000	0.000	0.111	0.556	0.575	0.615
CPFair	0.541	0.549	0.558	0.006	0.013	0.095	0.544	0.556	0.606
k-neighbor	0.453	0.469	0.499	0.125	0.133	0.156	0.516	0.536	0.577
Welf	0.536	0.554	0.556	0.000	0.000	0.167	0.536	0.554	0.640
Fairco	0.502	0.507	0.518	0.227	0.218	0.229	0.616	0.616	0.633
P-MMF	0.505	0.526	0.543	0.452	0.535	0.583	0.731	0.794	0.835
Dyn	0.439	0.442	0.444	0.003	0.002	0.001	0.4405	0.443	0.4445
LTP-MMF (Ours)	**0.505**	**0.531**	**0.553**	**0.468**	**0.545**	**0.572**	**0.739***	**0.804***	**0.840***
Improv.							1.1%	1.3%	0.6%
Amazon-Baby									
hLinUCB	0.570	0.511	0.514	0.000	0.000	0.000	0.570	0.511	0.514
EXP3	0.463	0.463	0.463	0.040	0.094	0.138	0.483	0.510	0.532
BLTS	0.546	0.549	0.551	0.006	0.024	0.055	0.549	0.561	0.578
FairRec	0.519	0.531	0.546	0.000	0.000	0.087	0.519	0.531	0.590
FairRec+	0.519	0.531	0.549	0.000	0.000	0.086	0.519	0.531	0.592
CPFair	0.524	0.536	0.536	0.017	0.046	0.104	0.533	0.559	0.588
k-neighbor	0.464	0.472	0.487	0.144	0.150	0.173	0.536	0.547	0.574
Welf	0.514	0.525	0.529	0.199	0.271	0.246	0.614	0.661	0.652
Fairco	0.496	0.504	0.507	0.253	0.252	0.250	0.623	0.630	0.632
P-MMF	0.493	0.512	0.521	0.533	0.560	0.580	0.760	0.792	0.811
Dyn	0.455	0.452	0.451	0.032	0.016	0.006	0.471	0.46	0.454
LTP-MMF (Ours)	**0.498**	**0.515**	**0.526**	**0.534***	**0.571***	**0.595***	**0.765***	**0.801***	**0.824***
Improv.							0.7%	1.1%	1.6%
Steam									
hLinUCB	0.820	0.574	0.707	0.000	0.000	0.000	0.820	0.574	0.707
EXP3	0.313	0.313	0.315	0.158	0.205	0.224	0.392	0.415	0.427
BLTS	0.829	0.849	0.018	0.011	0.017	0.835	0.834	0.857	0.844
FairRec	0.573	0.603	0.612	0.141	0.140	0.138	0.644	0.673	0.681
FairRec+	0.583	0.604	0.614	0.153	0.148	0.139	0.660	0.678	0.684
CPFair	0.713	0.734	0.756	0.013	0.016	0.046	0.720	0.742	0.779
k-neighbor	0.364	0.408	0.597	0.127	0.179	0.220	0.428	0.498	0.707
Welf	0.680	0.672	0.702	0.000	0.162	0.177	0.680	0.753	0.791
Fairco	0.538	0.534	0.547	0.264	0.265	0.254	0.670	0.676	0.674
P-MMF	0.573	0.593	0.624	0.547	0.602	0.608	0.847	0.894	0.928
Dyn	0.442	0.258	0.198	0.159	0.039	0.015	0.521	0.277	0.211
LTP-MMF (Ours)	**0.602**	**0.611**	**0.633**	**0.603**	**0.617**	**0.628**	**0.904***	**0.920***	**0.947***
Improv.							6.7%	2.9%	2.0%

The experimental settings are listed in Table 2. The bold numbers denote the performance of our models, and the underlined numbers denote the best-performing baselines. Improv., improvements.

$*$: Improvements over the best baseline are statistically significant (t-test, p-value < 0.05).



(a) Yelp (b) Amazon Beauty

(c) Amazon Baby (d) Steam

Fig. 3. Pareto frontier in four dataset.

Moreover, we can observe the changes in user accuracy and provider fairness with tradeoff co-efficient $\lambda = [0.001, 0.01]$. From Figure 3(d), we can observe that when λ increases from 0.001 to 0.01, the user accuracy (i.e., CTR@K) only decreases by 0.09%, but provider fairness (i.e., MMF@K) increases 220%. The result indicated that we can achieve better MMF in the long term while sacrificing little accuracy in LTP-MMF.

5.3 Experiment Analysis

We also conducted experiments to analyze LTP-MMF on the Steam dataset under top-10 settings.

5.3.1 Ablation Study on Exploration Term. To investigate the impact of exploration term (i.e., the upper-confidence-bound $\triangle f_{u,i}$ in Equation (10)), we conduct an ablation study shown in Figure 5. Specifically, we showed the CTR@10 and MMF@10 of two LTP-MMF variations, including the variations without exploration term (denoted as "LTP-MMF w/o exploration term"), and the complete LTP-MMF. From Figure 5, we found that in the first 25 interactions, the exploration term of LTP-MMF is dominant, leading to relatively poor performance. However, in the rest of the periods, LTP-MMF explores more trustful feedback of items, leading to better estimation of user preference. Therefore, LTP-MMF has better performance in terms of both user accuracy and provider fairness in the long term compared to "LTP-MMF w/o exploration term". The experiment also verified the importance of giving the exposures fairly to the providers in the feedback loop.

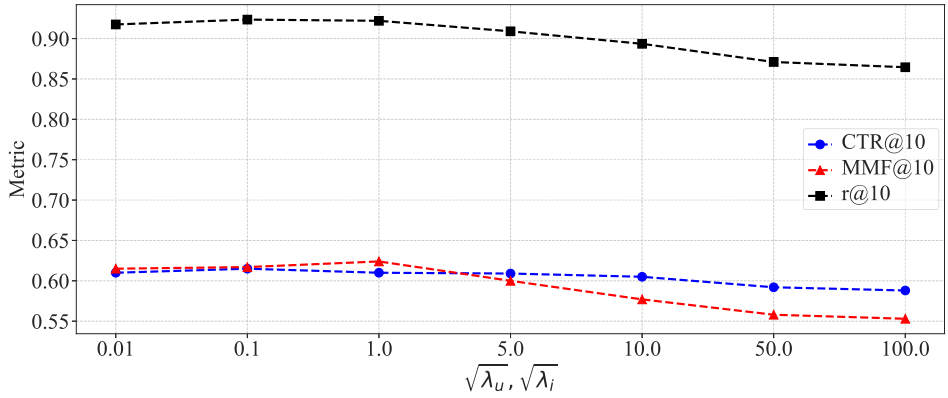
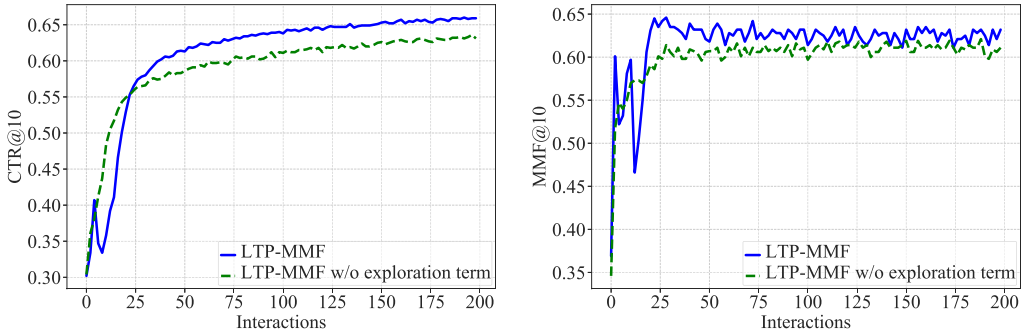


Fig. 4. Exploration weight.



(a) CTR (b) MMF

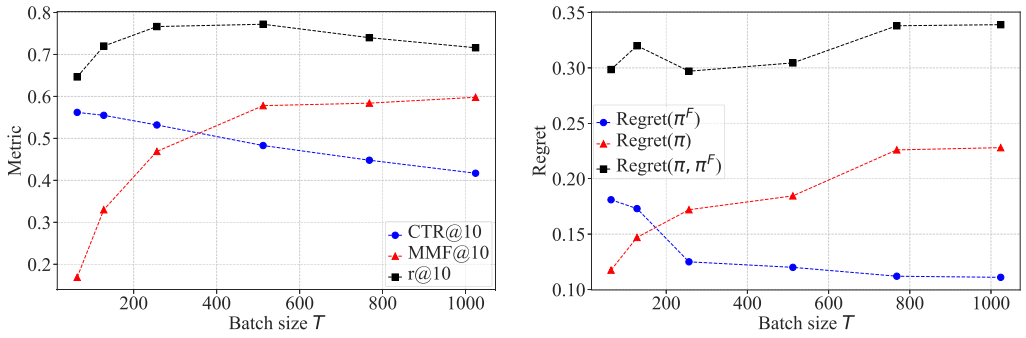
Fig. 5. Ablation study for the exploration of LTP-MMF.

At the same time, we also conduct the experiment to investigate the long-term impact of different exploration weights (i.e. $\sqrt{\lambda_i}$, $\sqrt{\lambda_u}$ in Equation (11)) shown in Figure 4. From the curve, we find that the performance improved when $\sqrt{\lambda_i}, \sqrt{\lambda_u} \in [0.01, 1]$ and then they dropped between $[1, 100]$, respectively. The reason is that the more exploration, the less bias will be in the long term. However, too much exploration will inevitably hurt the exploit reward, hurting the long-term performance. The opposite also holds. Therefore, we need to balance the exploration-exploit in the long term.

5.3.2 Ablation Study of Different Fair-Aware Modules. To investigate the impact of different fair-aware modules (fairness module and UCB module), we conduct an ablation study shown in Figure 9. Specifically, we showed the r@K of two LTP-MMF variations, including the variations without fairness module (denoted as "w/o fairness module"), without UCB module (denoted as "w/o UCB module"), and the complete LTP-MMF (denoted as "full modules").

From Figure 9, We can observe that when dropping out of the fairness and UCB modules, the overall performance decreases under different ranking sizes, indicating the effectiveness of the different fairness-aware modules.

5.3.3 Impact of the Batch Size T. According to Theorem 3, the batch size T balances the regret of user accuracy and provider fairness. In this experiment, Figure 6(a) studied how CTR@K, MMF@K, and r@K changed when the batch size T was set to different values from $[64, 1024]$. Figure 6(b)



(a) Performance (b) Regret

Fig. 6. (a) The long-term performance. (b) The long-term regret w.r.t. batch size T.

studied how regret of accuracy (Regret(π)), regret of fairness (Regret(π^F)) and regret of LTP-MMF (Regret(π, π^F)) changed when the batch size T was set from $[64, 1024]$.

From the curves shown in Figure 6(a), we found the accuracy performance CTR@K improved while the fairness performance MMF@K dropped when the batch size became smaller. Similarly, in Figure 6(b), the regret of accuracy Regret(π) dropped while the regret of fairness Regret(π^F) dropped when the batch size became smaller. The results verified the theoretical analysis that small T (e.g., $T = 64$) results in more bias in fairness estimation while large T (e.g., $T = 1024$) results in large bias in accuracy estimation in the long term.

Moreover, The overall performance $r@K$ improved and regret of LTP-MMF (Regret(π, π^F)) dropped when $T \in [64, 512]$ and then $r@K$, Regret(π, π^F) dropped and improved between $[512, 1024]$, respectively. Therefore, it is important to balance the accuracy and fairness in real applications through batch size T.

5.3.4 Periodical Performance. Figure 7 reports the experimental results of LTP-MMF and performing baselines to investigate how the LTP-MMF performs for each period. Note that similar experiment phenomena have also been observed on other λ and top-k values. To make fair comparisons, all the baselines were tuned $\lambda = 1$ as the evaluation metric.

From the $r@10$ curve, we found that although the performance gaps between LTP-MMF and other baselines are not obvious in the beginning (interaction 0), the LTP-MMF can well explore the feedback of unexposed items, leading to a huge improvement over other baselines in the long term (interaction 200). Moreover, in most periods, LTP-MMF can steadily outperform other baselines, verifying the effectiveness of LTP-MMF.

5.3.5 Online Inference Time. We experimented with investigating the online inference time of LTP-MMF. Figure 8 reports the curves of inference time (ms) of the accuracy module, fairness module, and exploration term computation per user access w.r.t. item size. We can see that LTP-MMF with GPU versions needs only within 6ms to calculate user-item scores $s_{u,i}$ and exploration term $\triangle f_{u,i}$. The reason is that this operation only needs matrix multiplication, and the inverse of the matrix can be stored in the training (offline) phase, leading to low latency in the online phase. Moreover, the inference time for the fairness module can also be maintained as 14–16 ms even when the item size becomes larger. We conclude that LTP-MMF can be adapted to online scenarios efficiently because of its low latency, even when the item size grows rapidly.

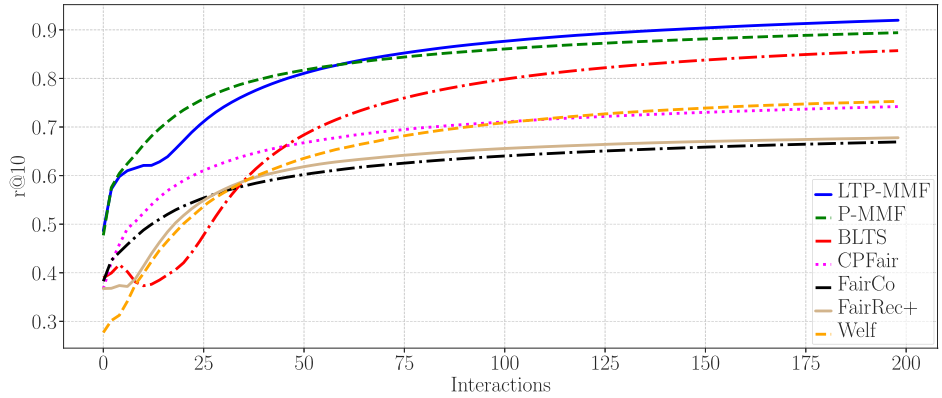


Fig. 7. The periodical performance of LTP-MMF and other top baselines.

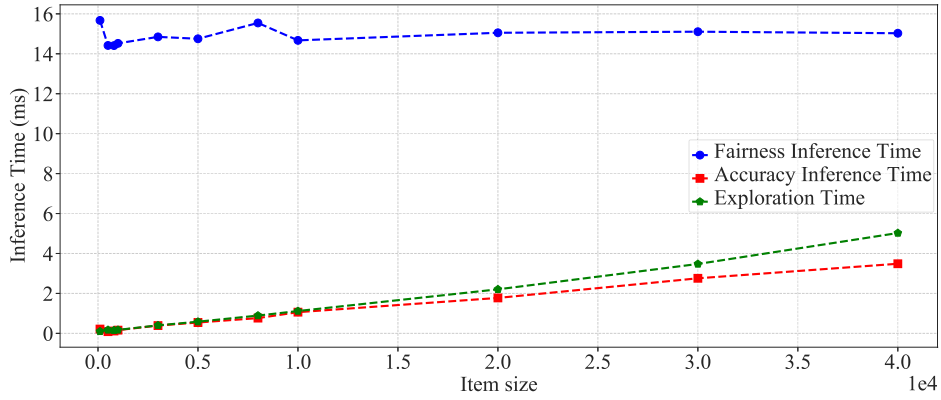


Fig. 8. The online inference time per user w.r.t. the number of total items.

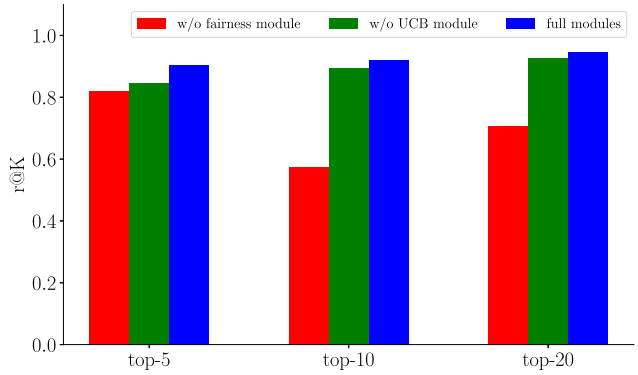


Fig. 9. Ablation studies of the individual impact of each component. We report the overall performance $r@K$ on Steam datasets under different ranking sizes K.

6 Conclusion

We proposed a novel ranking model called LTP-MMF that aims to consider provider MMF in the long term. Firstly, we formulated the provider fair recommendation as a repeated resource

allocation problem under a batched bandit setting. LTP-MMF applies the exploration term to break the loop while exploiting the fairness-aware rewards. Our theoretical analysis showed that the regret of LTP-MMF can be bounded. Experiments on four available datasets demonstrated that LTP-MMF can conduct ranking in an effective and efficient way.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*.

[2] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction* 30, 1 (2020), 127–158.

[3] Himan Abdollahpouri and Robin Burke. 2019. Multi-stakeholder recommendation and its connection to multi-sided fairness. arXiv:1907.13158. Retrieved from https://arxiv.org/abs/1907.13158

[4] Nil-Jana Akpinar, Cyrus DiCiccio, Preetam Nandy, and Kinjal Basu. 2022. Long-term dynamics of fairness intervention in connection recommender systems. arXiv:2203.16432. Retrieved from https://arxiv.org/abs/2203.16432

[5] Santiago Balseiro, Haihao Lu, and Vahab Mirrokni. 2021. Regularized online allocation problems: Fairness and beyond. In *Proceedings of the International Conference on Machine Learning*. PMLR, 630–639.

[6] Santiago R. Balseiro, Haihao Lu, Vahab Mirrokni, and Balasubramanian Sivan. 2022. From online optimization to PID controllers: Mirror descent with momentum. arXiv:2202.06152. Retrieved from https://arxiv.org/abs/2202.06152

[7] Omer Ben-Porat and Rotem Torkan. 2023. Learning with exposure constraints in recommendation systems. In *Proceedings of the ACM Web Conference*, 3456–3466.

[8] Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. 2011. The price of fairness. *Operations Research* 59, 1 (2011), 17–31.

[9] Asia J. Biega, Krishna .P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 405–414.

[10] Ilai Bistritz, Tavor Baharav, Amir Leshem, and Nicholas Bambos. 2020. My fair bandit: Distributed learning of max-min fairness with multi-player bandits. In *Proceedings of the International Conference on Machine Learning*. PMLR, 930–940.

[11] Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. 2019. Online EXP3 learning in adversarial bandits with delayed feedback. In *Proceedings of the 33rd International Conference onNeural Information Processing Systems*, 11349–11358.

[12] Arpita Biswas, Gourab K. Patro, Niloy Ganguly, Krishna P. Gummadi, and Abhijnan Chakraborty. 2021. Toward fair recommendation in two-sided platforms. *ACM Transactions on the Web* 16, 2 (2021), 1–34.

[13] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. PMLR, 202–214.

[14] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and debias in recommender system: A survey and future directions. arXiv:2010.03240. Retrieved from https://arxiv.org/abs/2010.03240

[15] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 191–198.

[16] Yashar Deldjoo. 2024. Understanding biases in ChatGPT-based recommender systems: Provider fairness, temporal stability, and recency. arXiv:2401.10545.

[17] Maria Dimakopoulou, Zhengyuan Zhou, Susan Athey, and Guido Imbens. 2019. Balanced linear contextual bandits. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 3445–3453.

[18] Karlijn Dinnissen and Christine Bauer. 2023. Amplifying artists' voices: Item provider perspectives on influence and fairness of music streaming platforms. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23)*. ACM, New York, 238–249. DOI: https://doi.org/10.1145/3565472.3592960

[19] Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. 2021. Two-sided fairness in rankings via Lorenz dominance. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 8596–8608.

[20] KJ Erickson. 2018. In Their Own Words: Why Sellers Are Fed Up With Amazon. Retrieved from https://medium.com/public-market/in-their-own-words-why-sellers-are-fed-up-with-amazon-e97da44f7f18

[21] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. Break the loop: Gender imbalance in music recommenders. In *Proceedings of the Conference on Human Information Interaction and Retrieval*, 249–254.

[22] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. 2021. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 445–453.

[23] Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. 2022. Provider fairness across continents in collaborative recommender systems. *Information Processing & Management* 59, 1 (2022), 102719.

[24] Xiangnan He, Yang Zhang, Fuli Feng, Chonggang Song, Lingling Yi, Guohui Ling, and Yongdong Zhang. 2023. Addressing confounding feature issue for causal recommendation. *ACM Transactions on Information Systems* 41, 3 (2023), 1–23.

[25] Balázs Hidasi, Massimo Quadrana, Alexandros Karatzoglou, and Domonkos Tikk. 2016. Parallel recurrent neural network architectures for feature-rich session-based recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 241–248.

[26] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining*. IEEE, 263–272.

[27] Rolf Jagerman, Ilya Markov, and Maarten de Rijke. 2019. When people change their mind: Off-policy evaluation in non-stationary recommendation environments. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*, 447–455.

[28] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.

[29] Saeedeh Karimi, Hossein A. Rahmani, Mohammadmehdi Naghiaei, and Leila Safari. 2023. Provider fairness and beyond-accuracy trade-offs in recommender systems. arXiv:2309.04250. Retrieved from https://arxiv.org/abs/2309.04250

[30] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[31] Julian Lamont and Christi Favor. 2017. Distributive Justice. In *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.). Edward N. Zalta (Ed.), Metaphysics Research Lab, Stanford University.

[32] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, 661–670.

[33] Qian Li, Xiangmeng Wang, Zhichao Wang, and Guandong Xu. 2023. Be causal: De-biasing social network confounding in recommendation. *ACM Transactions on Knowledge Discovery from Data* 17, 1 (Feb 2023), Article 14, 23 pages. DOI: https://doi.org/10.1145/3533725

[34] Xinyi Li, Yongfeng Zhang, and Edward C. Malthouse. 2023. A preliminary study of chatgpt on news recommendation: Personalization, provider fairness, fake news. arXiv:2306.10702. Retrieved from https://arxiv.org/abs/2306.10702

[35] Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2020. A general knowledge distillation framework for counterfactual recommendation via uniform data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 831–840.

[36] Alexander V. Lotov and Kaisa Miettinen. 2008. Visualizing the Pareto frontier. In *Multiobjective Optimization*. Springer, 213–243.

[37] Debmalya Mandal and Jiarui Gan. 2022. Socially fair reinforcement learning. arXiv:2208.12584. Retrieved from https://arxiv.org/abs/2208.12584

[38] Martin Mladenov, Elliot Creager, Omer Ben-Porat, Kevin Swersky, Richard Zemel, and Craig Boutilier. 2020. Optimizing long-term social welfare in recommender systems: A constrained matching approach. In *Proceedings of the International Conference on Machine Learning*. PMLR, 6987–6998.

[39] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. ACM, New York, NY, 429–438. DOI: https://doi.org/10.1145/3397271.3401100

[40] Mohammadmehdi Naghiaei, Hossein A Rahmani, and Yashar Deldjoo. 2022. Cpfair: Personalized consumer and producer fairness re-ranking for recommender systems. arXiv:2204.08085. Retrieved from https://arxiv.org/abs/2204.08085

[41] Harrie Oosterhuis and Maarten de Rijke. 2020. Policy-aware unbiased learning to rank for top-k rankings. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 489–498.

[42] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS '17)*.

[43] Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. 2020. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of the Web Conference*, 1194–1204.

[44] Tao Qi, Fangzhao Wu, Chuhan Wu, Peijie Sun, Le Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2022. ProFairRec: Provider fairness-aware news recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1164–1173.

[45] Ning Qian. 1999. On the momentum term in gradient descent learning algorithms. *Neural Networks* 12, 1 (1999), 145–151.

[46] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. arXiv:1205.2618.

[47] Paras Sheth, Ruocheng Guo, Kaize Ding, Lu Cheng, K. Selçuk Candan, and Huan Liu. 2022. Causal disentanglement with network information for debiased recommendations. In *International Conference on Similarity Search and Applications*. Springer, 265–273.
[48] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2219–2228.
[49] Wenlong Sun, Sami Khenissi, Olfa Nasraoui, and Patrick Shafto. 2019. Debiasing the human-recommender system feedback loop in collaborative filtering. In *Companion Proceedings of the World Wide Web Conference*, 645–651.
[50] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. 2017. Off-policy evaluation for slate recommendation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.
[51] Yong Kiam Tan, Xinxing Xu, and Yong Liu. 2016. Improved recurrent neural networks for session-based recommendations. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, 17–22.
[52] Huazheng Wang, Qingyun Wu, and Hongning Wang. 2016. Learning hidden features for contextual bandits. In *Proceedings of the 25th ACM international on Conference on Information and Knowledge Management*, 1633–1642.
[53] Huazheng Wang, Qingyun Wu, and Hongning Wang. 2017. Factorization bandits for interactive recommendation. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.
[54] Xiting Wang, Yiru Chen, Jie Yang, Le Wu, Zhengtao Wu, and Xing Xie. 2018. A reinforcement learning framework for explainable recommendation. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. IEEE, 587–596.
[55] Xin Wang, Steven CH Hoi, Chenghao Liu, and Martin Ester. 2017. Interactive social recommendation. In *Proceedings of the ACM on Conference on Information and Knowledge Management*, 357–366.
[56] David Williams. 1991. *Probability with Martingales*. Cambridge University Press.
[57] Yao Wu, Jian Cao, Guandong Xu, and Yudong Tan. 2021. Tfrom: A two-sided fairness-aware recommendation model for both customers and providers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1013–1022.
[58] Chen Xu, Sirui Chen, Jun Xu, Weiran Shen, Xiao Zhang, Gang Wang, and Zhenhua Dong. 2023. P-MMF: Provider Max-min Fairness re-ranking in recommender system. In *Proceedings of the ACM Web Conference*, 3701–3711.
[59] Chen Xu, Jun Xu, Xu Chen, Zhenghua Dong, and Ji-Rong Wen. 2022. Dually enhanced propensity score estimation in sequential recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*. ACM, New York, NY, 2260–2269. DOI: https://doi.org/10.1145/3511808.3557299
[60] Chen Xu, Jun Xu, Yiming Ding, Xiao Zhang, and Qi Qi. 2024. FairSync: Ensuring amortized group exposure in distributed recommendation retrieval. arXiv:2402.10628. Retrieved from https://arxiv.org/abs/2402.10628
[61] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2017. Deep matrix factorization models for recommender systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, Vol. 17, 3203–3209.
[62] Jiangxing Yu, Hong Zhu, Chih-Yao Chang, Xinhua Feng, Bowen Yuan, Xiuqiang He, and Zhenhua Dong. 2020. Influence function for unbiased recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1929–1932.
[63] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang. 2018. Deep reinforcement learning for page-wise recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 95–103.

A Appendix

A.1 Proof of Theorem 1

Proof. For max-min fairness, we have the regularizer as $r(\mathbf{e}) = \min_{p\in\mathcal{P}} \left(\mathbf{e}_p/\boldsymbol{\gamma}_p\right)$, we can easily proof that the exposure vector $\mathbf{e}$ can be represented as the dot-product between decision varible $\mathbf{x}_t$ and the item-provider adjacent matrix $\mathbf{A}$: $\mathbf{e} = \sum_{t=1}^{T} \left(\mathbf{A}^\top \mathbf{x}_t\right)$. Then, we treat the $\mathbf{e}$ as the auxiliary variable, and the ideal objective can be written as:

$$W_{OPT} = \max_{\mathbf{x}_t\in\mathcal{X}, \mathbf{e}\le\boldsymbol{\gamma}} \left[\sum_{t=1}^{T} g(\mathbf{x}_t)/T + \lambda r(\mathbf{e})\right]$$
$$s.t.\ \mathbf{e} = \sum_{t=1}^{T} \left(\mathbf{A}^\top \mathbf{x}_t\right),$$

where $\mathcal{X} = \{\mathbf{x}_t | \mathbf{x}_t \in 0, 1 \wedge \sum_{i \in I} \mathbf{x}_{ti} = K\}$. Then, we move the constraints to the objective using a vector of Lagrange multipliers $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{P}|}$:

$$\begin{aligned} W_{OPT} &= \max_{\mathbf{x}_t \in \mathcal{X}, \mathbf{e} \leq \boldsymbol{\gamma}} \min_{\boldsymbol{\mu} \in \mathcal{D}} \left[\sum_{t=1}^{T} g(\mathbf{x}_t)/T + \lambda r(\mathbf{e}) - \boldsymbol{\mu}^\top \left(-\mathbf{e} + \sum_{t=1}^{T} \mathbf{A}^\top \mathbf{x}_t \right) \right] \\ &\leq \min_{\boldsymbol{\mu} \in \mathcal{D}} \left[\max_{\mathbf{x}_t \in \mathcal{X}} \left[\sum_{t=1}^{T} g(\mathbf{x}_t)/T - \boldsymbol{\mu}^\top \sum_{t=1}^{T} \mathbf{A}^\top \mathbf{x}_t \right] + \max_{\mathbf{e} \leq \boldsymbol{\gamma}} \left(\lambda r(\mathbf{e}) - \boldsymbol{\mu}^\top \mathbf{e} \right) \right] \\ &= \min_{\boldsymbol{\mu} \in \mathcal{D}} \left[f^*(\mathbf{A}\boldsymbol{\mu}) + \lambda r^*(-\boldsymbol{\mu}) \right] = W_{Dual}, \end{aligned}$$

where $\mathcal{D} = \boldsymbol{\mu} | r^*(-\boldsymbol{\mu}) < \infty\}$ is the feasible region of dual variable $\boldsymbol{\mu}$. According to the Lemma 1 in the Balseiro et al. [5], we have $\mathcal{D}$ is convex and positive orthant is inside the recession cone of $\mathcal{D}$.

We let the variable $\mathbf{z}_p = (\mathbf{e}_p / \boldsymbol{\gamma}_p - 1)$, we have:

$$\begin{aligned} r^*(\boldsymbol{\mu}) &= \max_{\mathbf{e} \leq \boldsymbol{\gamma}} \left[\min \left(\mathbf{e}_p / \boldsymbol{\gamma}_p \right) + \boldsymbol{\mu}^\top \mathbf{e} / \lambda \right] \\ &= \boldsymbol{\mu}^\top \boldsymbol{\gamma} / \lambda + 1 + \max_{\mathbf{z}_p \leq 0} \left[\min(\mathbf{z}_p) + 1/\lambda \sum_{p \in \mathcal{P}} \boldsymbol{\mu}_p \boldsymbol{\gamma}_p \mathbf{z}_p \right] \end{aligned}$$

Let $s(\mathbf{z}) = \min_p \mathbf{z}_p$ and $\mathbf{v} = (\boldsymbol{\mu} \odot \boldsymbol{\gamma})/\lambda$, $\odot$ is the hadamard product. Then we define $s^*(\mathbf{v}) = \max_{\mathbf{z} \leq 0} \left(s(\mathbf{z}) + \mathbf{z}^T \mathbf{v} \right)$. We firstly show that if $\sum_{p \in \mathcal{S}} \mathbf{v}_p \geq -1, \forall \mathcal{S} \in \mathcal{P}_s$, then $s^*(\mathbf{v}) = 0$ and $\mathbf{z} = 0$ is the optimal solution, otherwise $s^*(\mathbf{v}) = \infty$.

We can equivalently write $\mathcal{D} = \{\mathbf{v} | \sum_{p \in \mathcal{S}} \mathbf{v}_p \geq -1, \forall \mathcal{S} \in \mathcal{P}_s\}$. We firstly show that $s^*(\mathbf{v}) = \infty$ for $\mathbf{v} \notin \mathcal{D}$. Suppose that there exists a subset $\mathcal{S} \in \mathcal{P}_s$ such that $\sum_{p \in \mathcal{S}} \mathbf{v}_p < -1$. For any $b > 1$, we can get a feasible solution:

$$\mathbf{v}_p = \begin{cases} -b, & p \in \mathcal{S} \\ 0, & otherwise. \end{cases}$$

Then, because such solution is feasible and $s(\mathbf{z}) = -b$, we obtain that

$$s^*(\mathbf{v}) \geq s(\mathbf{z}) - b \left(\sum_{p \in \mathcal{S}} \mathbf{v}_p \right) = -b \left(\sum_{p \in \mathcal{S}} \mathbf{v}_p + 1 \right).$$

Let $b \to \infty$, we have $s^*(\mathbf{v}) \to \infty$.

Then, we show that $s^*(\boldsymbol{\mu}) = 0$ for $\mathbf{v} \in \mathcal{D}$. Note that $\mathbf{z} = 0$ is feasible. Therefore, we have

$$s^*(\mathbf{v}) \geq s^*(0) = 0.$$

Then we have $\mathbf{z} \leq 0$ and without loss of generality, that the vector $\mathbf{z}$ is sorted in increasing order, i.e., $\mathbf{z}_1 \leq \mathbf{z}_2, \cdots, \leq \mathbf{z}_{|\mathcal{P}|}$. The objective value is

$$\begin{aligned} s^*(\mathbf{v}) &= \mathbf{z}_1 + \sum_{j \in |\mathcal{P}|} \mathbf{z}_j \mathbf{v}_j \\ &= \sum_{j=1}^{|\mathcal{P}|} \left(\mathbf{z}_j - \mathbf{z}_{j+1} \right) \left(1 + \sum_{i=1}^{j} \mathbf{v}_j \right) \leq 0. \end{aligned}$$

□

A.2 Proof of Lemma 1

Proof. From the proof of Theorem 1, we have

$$s^*(\boldsymbol{\mu}) = \max_{\boldsymbol{\mu} \le 0}(\min_p \boldsymbol{z}_p + \boldsymbol{z}^\top \boldsymbol{\mu}),$$

and $s^*(\boldsymbol{\mu}) = 0$ for $\mathbf{v} \in \mathcal{D}$. Therefore, we have

$$r^*(-\boldsymbol{\mu}) = \boldsymbol{\mu}^\top \boldsymbol{\gamma}/\lambda + 1.$$

□

A.3 Proof of Theorem 2

Proof. Our proof will take the following four steps:

Bias Term Bound. Firstly, we will bound the bias term $\|\hat{\boldsymbol{v}}_{u,t} - \boldsymbol{v}_u^*\|_{\boldsymbol{A}_{u,t}}, \|\hat{\boldsymbol{v}}_{i,t} - \boldsymbol{v}_i^*\|_{C_{i,t}}$:

We define the bias term upper bound as α_t, β_t respectively. We will bound the bias term as follows: by taking the gradient of the objective function respect to v_u, v_i, we have,

$$\boldsymbol{A}_{u,t}(\hat{\boldsymbol{v}}_{u,t} - \boldsymbol{v}_u^*) = \sum_{j=1}^{t} \hat{\boldsymbol{v}}_{i,j}(\boldsymbol{v}_i^* - \hat{\boldsymbol{v}}_{i,j})^\top \boldsymbol{v}_u^* + \sum_{j=1}^{t} (\hat{\boldsymbol{v}}_{i,j})\epsilon_j - \lambda_u \boldsymbol{v}_u^*, \tag{A1}$$

where ϵ_j is the Gaussian noise at time j. Without loss of generality, in ranking task, we can always scale the user-item score $s_{u,i}$, so that the l2-norm of v_u, v_i can be bounded by a constant factor:

$$\|v_u\|_2 \le 1, \quad \|v_i\|_2 \le 1$$

Therefore, we can bound the function norm of the above equation as

$$\begin{aligned} \|\hat{\boldsymbol{v}}_{u,t} - v_u^*\|_{\boldsymbol{A}_{u,t}} &= \| \sum_{j=1}^{t} \hat{\boldsymbol{v}}_{i,j}(v_i^* - \hat{\boldsymbol{v}}_{i,j})^\top \boldsymbol{v}_u^* + \sum_{j=1}^{t} (\hat{\boldsymbol{v}}_{i,j})\epsilon_j - \lambda_u \boldsymbol{v}_u^* \|_{\boldsymbol{A}_{u,t}} \\ &\le \| \sum_{j=1}^{t} \hat{\boldsymbol{v}}_{i,j}\epsilon_j \|_{\boldsymbol{A}_{u,t}} + \frac{1}{\sqrt{\lambda_u}} \sum_{j=1}^{t} \|\boldsymbol{v}_i^* - \hat{\boldsymbol{v}}_{i,j}\|_2 + \sqrt{\lambda_u} \end{aligned} \tag{A2}$$

Since the q-linearly convergent to the optimizer of parameter $\boldsymbol{v}_u, \boldsymbol{v}_i$ in [52], we have for every $\epsilon_q > 0$ and $0 < q < 1$, we have

$$\|\boldsymbol{v}_i^* - \hat{\boldsymbol{v}}_{i,t+1}\|_2 \le (q + \epsilon_q)\|\boldsymbol{v}_i^* - \hat{\boldsymbol{v}}_{i,t}\|_2 \tag{A3}$$

Therefore, by applying the self-normalized vector-valued martingales [1], we have for any $\sigma > 0$, with probability at least $1 - \sigma$,

$$\|\hat{\boldsymbol{v}}_{u,t} - v_u^*\|_{\boldsymbol{A}_{u,t}} \le \alpha_t,$$

$$\alpha_t = \sqrt{\lambda_u} + \frac{2(q + \epsilon_q)(1 - (q + \epsilon_q)^t)}{1 - q - \epsilon_q} + \sqrt{d \ln \frac{\lambda_u d + t}{\lambda_u d \sigma}}.$$

Similarly, we have

$$\|\hat{\boldsymbol{v}}_{i,t} - v_i^*\|_{C_{i,t}} \le \beta_t,$$

$$\beta_t = \sqrt{\lambda_i} + \frac{2(q + \epsilon_q)(1 - (q + \epsilon_q)^t)}{1 - q - \epsilon_q} + \sqrt{d \ln \frac{\lambda_i d + t}{\lambda_i d \sigma}}.$$

Therefore, the bias term α_t, β_t are comparable with $O((1 - (q + \epsilon_r)^t)\sqrt{\ln t})$ for every $\epsilon_r > 0$.

Variance Term Bound. Next, we will bound the variance term. Then we prove the converge rate of variance term: $\|\hat{\boldsymbol{v}}_{u,t}\|^2_{\boldsymbol{A}_{u,t}}, \|\hat{\boldsymbol{v}}_{i,t}\|^2_{\boldsymbol{C}_{i,t}}$ as follows:

$$\begin{aligned}\|\hat{\boldsymbol{v}}_{u,t}\|^2_{\boldsymbol{A}_{u,t}} &= \hat{\boldsymbol{v}}^\top_{u,t}\boldsymbol{A}_{u,t}\hat{\boldsymbol{v}}_{u,t} \\ &= \hat{\boldsymbol{v}}^\top_{u,t}\left(\lambda_u \boldsymbol{I} + \sum_{j=1}^{t}\hat{\boldsymbol{v}}_{i,t}\hat{\boldsymbol{v}}^\top_{i,t}\right)\hat{\boldsymbol{v}}_{u,t} \\ &\leq \|\hat{\boldsymbol{u}}_{i,t}\|_2^2 + \sum_{j=1}^{T}\|\hat{\boldsymbol{v}}^\top_{i,j}\hat{\boldsymbol{v}}_{u,t}\|_2^2 \leq (t+1) \sim O(t),\end{aligned} \tag{A4}$$

since we can always scale the l2-norm of v_u, v_i by any rate in ranking tasks (we only care the relative score). Thus we can easily obtain that

$$\|\hat{\boldsymbol{v}}_{u,t}\|_{\boldsymbol{A}^{-1}_{u,t}} \sim (1/\sqrt{t}).$$

Collaborative Variance Term Bound. Next, we will bound the variance term of collaborative term $\|\hat{\boldsymbol{v}}_{u,t} - \boldsymbol{v}^*_u\|^2_{\boldsymbol{A}^{-1}_{u,t}}, \|\hat{\boldsymbol{v}}_{i,t} - \boldsymbol{v}^*_i\|^2_{\boldsymbol{C}^{-1}_{i,t}}$ as follows:

$$\|\hat{\boldsymbol{v}}_{u,t} - \boldsymbol{v}^*_u\|_{\boldsymbol{A}^{-1}_{u,t}} \leq \|\hat{\boldsymbol{v}}_{u,t} - \boldsymbol{v}^*_u\|_2 \leq (q+\epsilon_q)^t,$$

where q, ϵ_q follows Equation (A3). Similarly, we have

$$\|\hat{\boldsymbol{v}}_{i,t} - \boldsymbol{v}^*_i\|_{\boldsymbol{C}^{-1}_{i,t}} \leq \|\hat{\boldsymbol{v}}_{i,t} - \boldsymbol{v}^*_i\|_2 \leq (q+\epsilon_q)^t.$$

Let's abbreviate the upper bound of collaborative error term $(q+\epsilon_q)^t$ as C

UCB. Finally, the UCB of user-item score can have easily by putting the aforementioned term together:

$$\begin{aligned}s^*_{u,i} - \hat{s}_{u_t,i} &= (\boldsymbol{v}^*_u)^\top \boldsymbol{v}^*_i - \hat{v}^\top_{u,t}\hat{v}_{i,t} \\ &= 1/2\left[(\boldsymbol{v}^*_u + \hat{v}_{u,t})^\top(\boldsymbol{v}^*_i - \hat{v}_{i,t}) + (\boldsymbol{v}^*_u - \hat{v}_{u,t})^\top(\boldsymbol{v}^*_i + \hat{v}_{i,t})\right] \\ &\leq \alpha_t(C/2 + \|\hat{v}_{i,t}\|_{A^{-1}_{u,t}}) + \beta_t(C/2 + \|\hat{v}_{u,t}\|_{C^{-1}_{i,t}}).\end{aligned} \tag{A5}$$

We can easily have the upper bound of user-item score have the converge term

$$O\left(\frac{(1-(q+\epsilon_q)^t)\sqrt{\ln t}}{\sqrt{t}}\right), \tag{A6}$$

which is a decrease function of t when the t becomes large. □

A.4 Proof of Theorem 3

Proof. Firstly, in practice, we normalize the user-item preference score $s_{u,i}$ to $[0,1]$. Therefore, $\sum_{t=1}^{T} g(\mathbf{x}_t)/T \leq K$. In max-min regularizer $r(\mathbf{e})$. Let's abbreviate its upper bound to $\bar{r}$. In practice, $\bar{r} \leq 1$ We have

$$W_{OPT} \leq K + \lambda\bar{r}. \tag{A7}$$

We consider the stopping time τ of Algorithm 1 as the first time the provider will have the maximum exposures, i.e.

$$\sum_{t=1}^{\tau}\mathbf{M}^\top\mathbf{x}_t \geq \boldsymbol{\gamma}.$$

Note that is τ a random variable.

Similarly, following the proven idea of Balseiro et al. [5], first, we analyze the primal performance of the objective function. Second, we bound the complementary slackness term by the momentum gradient descent. Finally, We conclude by putting it to achieve the final regret bound.

Primal Performance Proof: Consider a time $t < \tau$, the recommender action will not violate the resource constraint. Therefore, we have:

$$g(\mathbf{x}_t)/T = g^*(\boldsymbol{M}\boldsymbol{\mu}_t) + \lambda \boldsymbol{\mu}_t{}^T \boldsymbol{M}^\top \mathbf{x}_t,$$

and we have $\mathbf{e}_t = \arg\max_{\mathbf{e} \le \boldsymbol{\gamma}} \{r(\mathbf{e}) + \boldsymbol{\mu}^\top \mathbf{e}/\lambda\}$

$$r(\mathbf{e}_t) = r^*(-\boldsymbol{\mu}) - \boldsymbol{\mu}_\mathbf{t}{}^T \mathbf{e}_t/\lambda.$$

We make the expectations for the current time step t for the primal functions:

$$\begin{aligned}\mathbb{E}\left[g(\mathbf{x}_t)/T + \lambda r(\mathbf{e}_t)\right] &= \mathbb{E}\left[g^*(\boldsymbol{M}\boldsymbol{\mu}_t) + \boldsymbol{\mu}_\mathbf{t}{}^T \boldsymbol{M}^\top \mathbf{x}_t + \lambda r^*(-\boldsymbol{\mu}) - \boldsymbol{\mu}_\mathbf{t}{}^T \mathbf{e}_t\right] \\ &= W_{Dual}(\boldsymbol{\mu}_t) - \mathbf{E}\left[\boldsymbol{\mu}_t^T(-\boldsymbol{M}^\top \mathbf{x}_t + \mathbf{e}_t)\right].\end{aligned}$$

Consider the process $Z_t = \sum_{j=1}^T \boldsymbol{\mu}_j^t(-\boldsymbol{M}^\top \mathbf{x}_t + \mathbf{e}_t) - \mathbf{E}\left[\boldsymbol{\mu}_t^T(-\boldsymbol{M}^\top \mathbf{x}_t + \mathbf{e}_t)\right]$ is a martingale process. The Optional Stopping Theorem in martingale process [56] implies that $\mathbb{E}[Z_\tau] = 0$. Consider the variable $w_t(\boldsymbol{\mu}_t) = \boldsymbol{\mu}_t^T(-\mathbf{A}^\top \mathbf{x}_t + \mathbf{e}_t)$, we have

$$\mathbb{E}\left[\sum_{t=1}^{\tau} w_t(\boldsymbol{\mu}_t)\right] = \mathbb{E}\left[\sum_{t=1}^{\tau} \mathbb{E}\left[w_t(\boldsymbol{\mu}_t)\right]\right]$$

Moreover, in MMF, the dual function W_{Dual} is convex proofed in Theorem 1, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{\tau} g(\mathbf{x}_t)/T + \lambda r(\mathbf{e}_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{\tau} W_{Dual}(\boldsymbol{\mu}_t)\right] - \mathbb{E}\left[\sum_{t=1}^{\tau} w_t(\boldsymbol{\mu}_t)\right] \\ &\le \mathbb{E}\left[\tau W_{Dual}(\widetilde{\boldsymbol{\mu}_\tau})\right] - \mathbb{E}\left[\sum_{t=1}^{\tau} w_t(\boldsymbol{\mu}_t)\right], \end{aligned} \tag{A8}$$

where $\widetilde{\boldsymbol{\mu}_\tau} = \sum_{t=1}^{\tau} \boldsymbol{\mu}_t/\tau$.

Next, we will bound the bias of the primal performance due to the estimation error of the ranking model.

From the proof of Theorem 1, we can bound the $W_{Dual}(\boldsymbol{\mu}_t)$ as follows: at iteration n, for any $\sigma > 0$, with probability at least $1 - \sigma$,

$$\begin{aligned}W_{Dual}(\boldsymbol{\mu}_t) - \hat{W}_{Dual}(\boldsymbol{\mu}_t) &= g^*(\boldsymbol{M}\boldsymbol{\mu}_t) + \lambda r^*(-\mu) - \hat{g}^*(\boldsymbol{M}\boldsymbol{\mu}_t) - \lambda r^*(-\mu) \\ &= \boldsymbol{s}_u^\top \boldsymbol{x}_t - (\boldsymbol{M}\boldsymbol{\mu}_t)^\top \boldsymbol{x}_t - \boldsymbol{s}_u^\top \boldsymbol{x}_t - (\boldsymbol{M}\boldsymbol{\mu}_t)^\top \boldsymbol{x}_t \le \sum_{i, \boldsymbol{x}_{ti}=1} \triangle f_{u_t,i}^n,\end{aligned}$$

Complementary Slackness Proof. Then we aim to proof the complementary slackness $\sum_{t=1}^T w_t(\boldsymbol{\mu}_t) - w_t(\boldsymbol{\mu})$ is bounded. Suppose there exists $G, s.t.$ the gradient norm is bounded $\|\widetilde{\mathbf{g}}_t\| \le G$. Then we have:

$$\sum_{t=1}^{\tau} w_t(\boldsymbol{\mu}_t) - w_t(\boldsymbol{\mu}) \le \frac{L^2}{\eta} + \frac{G^2}{(1-\alpha)\sigma}\eta(\tau - 1) + \frac{G^2}{2(1-\alpha)^2\sigma\eta}, \tag{A9}$$

where the project function $\|\boldsymbol{\mu} - \boldsymbol{\mu}_t\|_\gamma^2$ is $\sigma-$strongly convex.

Next, we prove the inequality in Equation. According to the Theorem 1 in [6], we have

$$\|\mathbf{g}_t\|_2^2 = \|(1-\alpha)\sum_{s=1}^{t} \alpha^{t-s}(\widetilde{\mathbf{g}}_s)\|_2^2 \le G^2,$$

and

$$\sum_{t=1}^{\tau} w_t(\boldsymbol{\mu}_t) - w_t(\boldsymbol{\mu}) \leq \frac{\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_0\|^2_{\boldsymbol{\gamma}^2}}{\eta} + \frac{G^2}{(1-\alpha)\sigma}\eta(\tau-1) + \frac{G^2}{2(1-\alpha)^2\sigma\eta}, \forall \boldsymbol{\mu}.$$

We have $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_0\|^2_{\boldsymbol{\gamma}^2} \leq L^2$ according to the Cauchy-Schwarz' inequality. The results follows. Let $M = \frac{L^2}{\eta} + \frac{G^2}{(1-\alpha)\sigma}\eta(T-1) + \frac{G^2}{2(1-\alpha)^2\sigma\eta}$. We now choose a proper $\boldsymbol{\mu}$, s.t. the complementary stackness can be further bounded.

For $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}} + \theta$, where $\theta \in \mathbb{R}^{|P|}$ is non-negative to be determined later and $\hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} - \boldsymbol{\mu}^\top(\sum_{i=1}^{T} \mathbf{A}^\top \mathbf{x}_t)/\lambda$. According to the constraint $\mathbf{e} = \sum_{i=1}^{T} \mathbf{A}^\top \mathbf{x}_t$, we have that

$$\sum_{t=1}^{T}(r(\mathbf{e}_t) + \boldsymbol{\mu}^\top \mathbf{e}_t \lambda) \leq r^*(-\hat{\boldsymbol{\mu}}) = r\left(\sum_{i=1}^{T} \mathbf{A}^\top \mathbf{x}_t\right) + \hat{\boldsymbol{\mu}}^T\left(\sum_{i=1}^{T} \mathbf{A}^\top \mathbf{x}_t\right)/\lambda.$$

Note that in proof of Theorem 1, the feasible region $\mathcal{D}$ is recession cone, therefore, $\boldsymbol{\mu} \in \mathcal{D}$.

Therefore, we have

$$\sum_{t=1}^{\tau} w_t(\boldsymbol{\mu}_t) = \sum_{t=1}^{T} w_t(\hat{\boldsymbol{\mu}}) - \sum_{t=\tau+1}^{T} w_t(\hat{\boldsymbol{\mu}}) + \sum_{t=1}^{\tau} w_t(\theta) + M. \tag{A10}$$

For each iteration n, we have

$$w_t \boldsymbol{\mu}_t - \hat{w}_t \boldsymbol{\mu}_t = \boldsymbol{\mu}_t((\boldsymbol{M} - \boldsymbol{M})^\top \boldsymbol{x}_t) \leq \sum_{i, \boldsymbol{x}_{ti}=1} \triangle f^n_{u_t,i} L$$

Put Them Together: For each batch T at iteration n, we obtain that

$$\begin{aligned} W_{OPT} &= \frac{\tau}{T} W_{OPT} + \frac{T-\tau}{T} W_{OPT} \\ &\leq \tau W_{Dual}(\widetilde{\boldsymbol{\mu}_\tau}) + (T-\tau)(K + \lambda \bar{r}) \\ &\leq \tau \hat{W}_{Dual}(\widetilde{\boldsymbol{\mu}_\tau}) + \sum_{t=1}^{\tau} \sum_{i, \boldsymbol{x}_{ti}} \triangle f^n_{u_t,i} + (T-\tau)(K + \lambda \bar{r}) \end{aligned} \tag{A11}$$

Let's abbreviate $\triangle r(n) = \sum_{t=1}^{\tau} \sum_{i, \boldsymbol{x}_{ti}} \triangle f^n_{u_t,i} L$. Therefore, combining Equations (10), (12), and (13) the regret $\text{Regret}(n)$ of iteration n, can be bounded as:

$$\begin{aligned} \text{Regret}(n) &= \mathbb{E}\left[W_{OPT} - \hat{W}\right] \\ &\leq \mathbb{E}\left[W_{OPT} - \sum_{t=1}^{\tau}\left(\hat{g}(\mathbf{x}_t)/T - \lambda r(\boldsymbol{M}^\top \mathbf{x}_t/\boldsymbol{\gamma})\right)\right] \\ &\leq \mathbb{E}\left[W_{OPT} - \tau \hat{W}_{Dual}(\widetilde{\boldsymbol{\mu}_t}) + \sum_{t=1}^{\tau} \hat{w}_t(\boldsymbol{\mu}_t) + \sum_{t=1}^{T}(\mathbf{e}_t - \boldsymbol{M}^\top \mathbf{x}_t)\right] \\ &\leq \mathbb{E}\left[(T-\tau)(K + \lambda \bar{r}) + \sum_{t=1}^{T} w_t(\hat{\boldsymbol{\mu}}) + \sum_{t=1}^{\tau} w_t(\theta)\right] + M + \triangle r(n) \\ &\leq (T-\tau)(K + \lambda \bar{r} + \lambda K) + \sum_{t=1}^{\tau} w_t(\theta) + M + \triangle r(n) \\ &= T\text{Regret}(\pi^F)/N + \triangle r(n) \end{aligned} \tag{A12}$$

Let $C = K + \lambda\bar{r} + \lambda K$, then setting the $\theta = C\min_p \boldsymbol{\gamma}_p \mathbf{u}_p$, where $\mathbf{u}_p$ is the pth unit vector. We have

$$\sum_{t=1}^{\tau} w_t(\theta) = C/(\min_p \boldsymbol{\gamma}_p) - C(T - \tau).$$

Then the $\text{Regret}(n) \leq M + C/(\min_p \boldsymbol{\gamma}_p)$, when we set $\eta = O(T^{-1/2})$, the $\text{Regret}(\pi^F)$ is comparable with $O(T^{-1/2})$.

According to our algorithm in Algorithm of LTP-MMF, the user will interact with the system in N/T iterations. Following the Lemma 11 in Abbasi-Yadkori et al. [1], the error $\text{Regret}(\pi)$ raised of accuracy module is bounded as

$$\text{Regret}(\pi) = \sum_{n=1}^{N/T} \triangle r(n) \leq KT[\alpha_{N/T}(\rho_u + \kappa) + \beta_{N/T}(\rho_i + \kappa)],$$

where

$$\kappa = \frac{(q + \epsilon_q)(1 - (q + \epsilon_q)^{N/T})}{1 - q - \epsilon_q},$$

and

$$\rho_u = \sqrt{2d\frac{N}{T}\ln\left(1 + \frac{N}{T\lambda_u d}\right)}, \rho_i = \sqrt{2d\frac{N}{T}\ln\left(1 + \frac{N}{T\lambda_i d}\right)}.$$

From the conclusion, we can see that the regret of accuracy module is comparable with $O(\sqrt{NT\ln\frac{N}{T}})$.

Finally, the total regret can be bounded as

$$\text{Regret}(\pi, \pi^F) = \sum_{n=1}^{N/T} \text{Regret}(n) = \text{Regret}(\pi^F) + \text{Regret}(\pi)L \tag{A13}$$

Setting the learning rate as $\eta = O(T^{-1/2})$, we can obtain a fairness regret $\text{Regret}(\pi^F)$ upper bound of order $O(\frac{N}{\sqrt{T}})$. Overall, the long-term regret of LTP-MMF can be obtained of order $O(N\ln N)$. □

Received 10 August 2023; revised 23 March 2024; accepted 24 August 2024