Information Sciences 645 (2023) 119367

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Semi-parametric contextual bandits with graph-Laplacian regularization

Young-Geun Choi [a], Gi-Soo Kim [b,*], Seunghoon Paik [c], Myunghee Cho Paik [d,e]

[a] Department of Mathematics Education, Sungkyunkwan University, Seoul, 03063, Republic of Korea
[b] Department of Industrial Engineering and Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology, Ulsan, 44919, Republic of Korea
[c] Department of Statistics, University of California Berkeley, Berkeley, 94720, CA, United States
[d] Department of Statistics, Seoul National University, Seoul, 08826, Republic of Korea
[e] Shepherd23 Inc., Seoul, 08826, Republic of Korea

A R T I C L E   I N F O

Keywords:
Contextual multi-armed bandit
Graph Laplacian
Semi-parametric reward model
Thompson sampling

A B S T R A C T

Non-stationarity is ubiquitous in human behavior and addressing it in the contextual bandits is challenging. Several works have addressed the problem by investigating semi-parametric contextual bandits and warned that ignoring non-stationarity could harm performances. Another prevalent human behavior is social interaction which has become available in a form of a social network or graph structure. As a result, graph-based contextual bandits have received much attention. In this paper, we propose SemiGraphTS, a novel contextual Thompson-sampling algorithm for a graph-based semi-parametric reward model. Our algorithm is the first to be proposed in this setting. We derive an upper bound of the cumulative regret that can be expressed as a multiple of a factor depending on the graph structure and the order for the semi-parametric model without a graph. We evaluate the proposed and existing algorithms via simulation and real data example.

1. Introduction

In contextual multi-armed bandits (MAB), a learning agent sequentially chooses actions while balancing to maximize the reward (exploitation) and to learn the reward mechanism as a function of contexts with higher precision (exploration). Algorithms for contextual MAB problems have demonstrated their usefulness in many applications including recommendations of news articles, advertisements, or behavioral interventions [16,22,23]. Thompson sampling (TS)-based algorithms randomly choose an action from repeatedly updated posterior, and have been widely used among other bandit algorithms [21,12,3].

The semi-parametric contextual bandit [11,14,13] models the mean of the reward by a linear function of the contexts and a time-varying intercept. The algorithms for semi-parametric models allow the reward distribution to change over time in a non-stationary manner. For example, behavior may change over time depending on the user's circumstances or preference for a shopping item may change according to a time trend. These may not be captured in the context vectors. In the single-user setting, the semi-parametric bandits have demonstrated success in accommodating non-stationarity in mobile health and product recommendation [11,13,19,18].

* Corresponding author.
E-mail address: gisookim@unist.ac.kr (G.-S. Kim).

https://doi.org/10.1016/j.ins.2023.119367
Received 28 October 2022; Received in revised form 11 April 2023; Accepted 25 June 2023

Available online 29 June 2023

In many real-life settings, there are multiple users and the relationships among the users in a social network are often available as side information. Such graph information has been utilized in recommendation [16,9,20]. Several graph-based contextual MAB algorithms have been proposed to take the graph information into account under the ordinary linear reward assumption [6,10,24,17,26]. The aforementioned graph-based methods have shown to take advantages of a graph structure and perform well, but may be restrictive in real-life settings when the rewards tend to change over time.

Our goal is to construct a semi-parametric bandit algorithm that accommodates multiple users equipped with a network, with practically feasible computational cost. To the best of our knowledge, our algorithm is the first algorithm proposed in this setting. The main contributions of the work presented in this paper are as follows.

- We propose SemiGraphTS (semi-parametric-graph-Thompson-sampling), a novel TS algorithm for a setting in which each user's reward follows the semi-parametric model and user-specific parameters are regularized by the given graph.
- We derive an upper bound of the cumulative regret for SemiGraphTS, which be expressed as a multiple of a factor depending on the graph structure and the bound from the semi-parametric model without a graph.
- We propose a novel scalable estimator for the user-specific parameter that incorporates the estimators from the neighbors defined by the graph structure while conditioning out time-dependent coefficients. This plays a crucial role in building the SemiGraphTS algorithm. We establish a high-probability upper bound for its estimation error.

2. Model and problem setting

We study the semi-parametric contextual bandit problem for multiple users equipped with a user network. Suppose that there are n users, say $j \in V = \{1, \dots, n\}$. For each time step $t = 1, \dots, T$, the learning agent is instructed which user to serve, say j_t. The agent is supposed to recommend an item or pull an arm for the target user based on the previous action history and the contexts describing the items. Suppose that there are N candidate arms, say $i = 1, \dots, N$, and that a context vector $b_i(t) \in \mathbb{R}^d$ represents the feature of the i-th item at time t. We denote by $a(t)$ the selected arm to recommend to the target user. We let $r_{i,j}(t)$ be the reward for arm i, user j at time t. Upon the action, the user returns a user-specific reward for the chosen arm, say $r_{a(t),j_t}(t)$. The information given to the learner at time t is formally described as filtration $\mathcal{F}_{t-1} = \{j_t, \{b_i(t)\}_{i=1}^N\} \bigcup \left(\bigcup_{\tau=1}^{t-1} \{j_\tau, \{b_i(\tau)\}_{i=1}^N, a(\tau), r_{a(\tau),j_\tau}(\tau)\} \right)$.

The multiple-user semi-parametric reward model is described as below:

$$r_{i,j}(t) = \nu_j(t) + b_i(t)^T \mu_j + \eta_{i,j}(t), \tag{1}$$

for $i = 1, \dots, N$, $j = 1, \dots, n$, and $t = 1, \dots, T$. Here, $\mu_j \in \mathbb{R}^d$ denotes the unknown user-specific parameter that represents the preference of the j-th user for a given context. The intercept $\nu_j(t)$ indicates the baseline reward for user j at time t. We do not impose any parametric assumption on the functional form of $\nu_j(t)$; we allow the baseline to arbitrarily change over time and users, whatever gradually and abruptly. When $\nu_j(t) = 0$ for all j, (1) is reduced to the standard linear reward model. Without loss of generality, we assume a uniform boundedness of the contexts and true parameters, i.e., $|\nu_j(t)| \le 1$, $\|b_i(t)\| \le 1$ and $\|\mu_j\| \le 1$ for all i, j and t, where $\|\cdot\|$ denotes the vector ℓ_2 norm. This assumption can be satisfied by rescaling the data. We assume that the random error $\eta_{i,j}(t)$ satisfies $\mathbb{E}(\eta_{i,j}(t)|\mathcal{F}_{t-1}) = 0$. If $n = 1$, (1) coincides with the single-user semi-parametric bandit problem. In addition, we assume that $\eta_{i,j}(t)$ given $\mathcal{F}_{t-1}$ is R-sub-Gaussian, that is, for every $c \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\{c\eta_{i,j}(t)\}|\mathcal{F}_{t-1}\right] \le \exp(c^2R^2/2), \tag{2}$$

for all i, j, t, which is a common assumption in the literature for theoretical derivations.

The optimal arm $a^*(t)$ is defined as the arm that maximizes the expected reward for the j_t-th user given the history, that is, $a^*(t) = \operatorname{argmax}_i \mathbb{E}(r_{i,j_t}(t)|\mathcal{F}_{t-1}) = \operatorname{argmax}_i\{\nu_{j_t}(t) + b_i(t)^T\mu_{j_t}\} = \operatorname{argmax}_i\{b_i(t)^T\mu_{j_t}\}$. Although $a^*(t)$ may be different across users, in each round t, only one user enters, and we omit the subscript. Regret at time t is defined by the difference between the expected rewards from the optimal arm and the chosen arm,

$$regret(t) = \mathbb{E}(r_{a^*(t),j_t}(t)|\mathcal{F}_{t-1}) - \mathbb{E}(r_{a(t),j_t}(t)|\mathcal{F}_{t-1}) = b_{a^*(t)}(t)^T\mu_{j_t} - b_{a(t)}(t)^T\mu_{j_t}.$$

The goal of the agent is to minimize the cumulative regret, $R(T) = \sum_{t=1}^T regret(t)$.

In graph-based bandit settings, the user network is given a priori as the side information. Without any information on the user network, the problem reduces to learning n independent instances. In order to clarify the concept, we initially assume that the given network is undirected and simple. Nonetheless, our approach can readily be extended to accommodate weighted and/or directed networks. This will be discussed further in the final portion of Section 4. Let $\mathcal{G} = (V, E)$ be an undirected simple graph, where a node $j \in V = \{1, \dots, n\}$ corresponds to a user and an edge $\{j,k\} \in E$ represents the link between users. There are several ways to uniquely represent $\mathcal{G}$ as a Laplacian matrix $L = (l_{jk}) \in \mathbb{R}^{n\times n}$. We employ the random-walk normalized Laplacian defined by

$$l_{jj} = 1, \quad l_{jk} = \begin{cases} -1/\deg(j) & \text{if } \{j,k\} \in E, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

for $j, k = 1, \dots, n$ with $j \ne k$. The choice of random-walk normalized Laplacian is particularly useful in the regret analysis and discussed after the proof sketch. In addition, let $\Delta_j = \sum_{k=1}^n l_{jk}\mu_k = \mu_j - \sum_{k:\{j,k\}\in E}\mu_k/\deg(j)$. Δ_j is the difference between the parameter associated with user j and the average of her neighbors' parameters. Clearly, $\|\Delta_j\|$ is small if μ_j and its neighboring μ_k's are similar.

In an extreme case, we have $\Delta_j = 0$ if $\mu_j = \mu_k$ for all neighbors k. In Section 4, we will see that the regret bound and the parameter convergence rate depends on $\max_{1\le j\le n} \|\Delta_j\|$, the maximum "dissimilarities" between user-specific parameters. Our working assumption is that $\|\Delta_j\|$ is small for all j, i.e., the user-specific parameters are smooth on the graph. Without loss of generality, we assume that $\mathcal{G}$ is connected. If not, each connected component of users do not share any information of parameters and it suffices to learn each connected component separately.

In addition, let $\|x\|_A = \sqrt{x^T A x}$ for $x \in \mathbb{R}^d$ and a positive semi-definite $A \in \mathbb{R}^{d\times d}$. A matrix-valued inequality $A \ge B$ ($A > B$) denotes that $A - B$ is positive semi-definite (positive definite).

2.1. Related work

Since linear contextual MAB problems for single users were investigated [2,3], there has been a rich line of works on contextual bandits in recent years. For conciseness, we focus on works that consider either the semi-parametric model for single user or the linear model for multiple user equipped with graph.

Semi-parametric contextual MABs for single user The semi-parametric reward model for a single user [11,14,13] assumes, say

$$r_i(t) = \nu(t) + b_i(t)^T \mu + \eta_i(t), \tag{4}$$

which is a special case of our model (1) with $n = 1$. Greenewald et al. [11] first proposed (4). A novel challenge in the semi-parametric bandit problem is to mitigate the confounding effect from the baseline reward. Greenewald et al. [11] considered a two-stage TS algorithm that fixes a random base action and contrasts the base and other actions. Krishnamurthy, Wu, and Syrgkanis [14] proposed another TS algorithm that contrasts every pair of actions repeatedly. Kim and Paik [13] proposed a single-step TS algorithm and arguably the state-of-the-art in this setting. Specifically, for each time t, they estimate μ in (4) by $\hat{\mu}(t) = B(t)^{-1}\sum_{\tau=1}^{t-1} 2X_\tau r_{a(\tau),k}(\tau)$, where $X_\tau = b_{a(\tau)}(\tau) - \mathbb{E}(b_{a(\tau)}(\tau)|\mathcal{F}_{\tau-1})$ and $B(t) = \hat{\Sigma}_t + \Sigma_t + I_d$ where $\hat{\Sigma}_t = \sum_{\tau=1}^{t-1} X_\tau X_\tau^T$, and $\Sigma_t = \sum_{\tau=1}^{t-1} \mathbb{E}(X_\tau X_\tau^T|\mathcal{F}_{\tau-1})$. Compared with Agrawal and Goyal [3], a TS algorithm under the standard linear reward model, the context vector and covariance part were centered by $\mathbb{E}(b_{a(\tau)}(\tau)|\mathcal{F}_{\tau-1})$, which is crucial for ruling out the confounding effect of $\nu(t)$. The regret bound derived in Kim and Paik [13] has the same order with that in Agrawal and Goyal [3].

Linear graph-based bandit algorithms for multiple users Algorithms for graph-based linear contextual bandits have been proposed under the following model [6,10,24,17,26,15]:

$$r_{i,j}(t) = b_i(t)^T \mu_j + \eta_{i,j}(t), \tag{5}$$

which coincides with a special case of (1) when $\nu_j(t) = 0$. Gentile, Li, and Zappella [10] proposed an algorithm utilizing the given graph for clustering users, where those in the same cluster are represented by the same parameter. Li et al. [17] generalized Gentile, Li, and Zappella [10]'s algorithm to address non-uniform user frequencies. Li, Wu, and Wang [15] proposed another clustering-based algorithm that allows each μ_j to change abruptly over time. The regret bound proposed in this work depends on the number of abrupt shifts and can be linear in T if the shifts occur proportionally to T. On the other hand, Casa-bianchi et al. [6] and Vaswani, Schmidt, and Lakshmanan [24] proposed UCB- and TS-based algorithms with regret bound $\tilde{O}(dn\sqrt{T})$, where the entire parameters for all users are estimated under regularization by a graph Laplacian. However, this led to scalability issues as a result of solving an equation involving nd by nd matrix. Yang, Toni, and Dong [26] proposed a local version of the Casa-bianchi et al. [6] with an improved regret bound $\tilde{O}(\Phi d\sqrt{nT})$, where $\Phi \in (0,1)$ depends on $\mathcal{G}$. It updates only the parameter associated with the user to serve at each round. Specifically, Yang, Toni, and Dong [26] first calculates the ordinary least squares estimator $\bar{\mu}_k(t)$ for each user k as if running n bandits independently. Then, μ_{j_t} is estimated by adjusting $\bar{\mu}_{j_t}$ for $\bar{\mu}_k(t)$ weighted by the Laplacian, particularly $\hat{\mu}_{j_t}(t) = \bar{\mu}_{j_t}(t) - \lambda C_{j_t}(t)^{-1}\sum_{k=1}^n l_{j_t k}\bar{\mu}_k(t)$, where λ is a tunable parameter and C_{j_t} is the gram matrix of the selected arm features for user j_t up to time t.

3. Proposed algorithm

A challenge is that we observe rewards that are correlated with neighbors defined from the given graph structure and yet whose conditional mean changes over time. Existing algorithms designed for the semi-parametric reward model with single user settings have only two choices: (i) to pretend that multiple users are actually the same person and estimate one parameter that applies to everyone, or (ii) to assign different parameters to different users and update the parameter estimate based on data from the corresponding user only. The risk of method (i) is that a single parameter is not able to incorporate the heterogeneity among users, while the risk of method (ii) is inefficiency in learning parameters.

Our novel approach incorporates the network information in estimating each μ_j while handling the confounding by $\nu_j(t)$. To be specific, we handle non-stationarity for each individual by conditioning, while simultaneously accommodating information from neighbors. The key idea of conditioning is based on that the non-stationarity does not change across the arms, hence centering the context around the mean for the arms does not alter the problem of finding the maximum reward across the arms. This allows us to construct an estimator of μ_j that is robust to the effect of $\nu_j(t)$ while exploiting the user affinity information via graph.

The proposed SemiGraphTS algorithm is described in Algorithm 1. Key steps include parameter estimation and Thompson sampling steps.

Algorithm 1 Proposed algorithm (SemiGraphTS).

Fix $\lambda > 0$. Set $B_j(1) = \lambda l_{jj} I_d$, $y_j(1) = 0_d$ and $v_j = (4R+12)\sqrt{d \log\left\{(24T^4/\delta)(1+\lambda^{-1})\right\}} + \sqrt{\lambda}(1+\|\Delta_j\|)$ for $j = 1, \dots, n$.
for $t = 1, 2, \dots, T$ **do**
  Observe j_t.
  for $j = 1, 2, \dots, n$ **do**
    if $j \neq j_t$ **then**
      Update $B_j(t+1) \leftarrow B_j(t)$, $\bar{\mu}_j(t+1) \leftarrow \bar{\mu}_j(t)$, and $y_j(t+1) \leftarrow y_j(t)$.
    else
      $\widehat{\mu}_j(t) \leftarrow \bar{\mu}_j(t) - B_j(t)^{-1} \sum_{k \neq j} \lambda l_{jk} \bar{\mu}_k(t)$.
      $\Gamma_j(t) \leftarrow B_j(t) + \lambda^2 \sum_{k \neq j} l_{jk}^2 B_k(t)^{-1}$
      Sample $\widetilde{\mu}_j(t)$ from $\mathcal{N}_d(\widehat{\mu}_j(t), v_j^2 \Gamma_j(t)^{-1})$.
      Pull arm $a(t) = \mathrm{argmax}_i\{b_i(t)^T \widetilde{\mu}_j(t)\}$ and get reward $r_{a(t),j}(t)$.
      $\pi_i(t) \leftarrow \mathbb{P}(a(t) = i | \mathcal{F}_{t-1})$, $i = 1, \dots, N$.
      $\bar{b}(t) \leftarrow \sum_{i=1}^N \pi_i(t) b_i(t)$ and $X_t \leftarrow b_{a(t)}(t) - \bar{b}(t)$.
      Update $B_j(t+1) \leftarrow B_j(t) + X_t X_t^T + \sum_{i=1}^N \pi_i(t)(b_i(t) - \bar{b}(t))(b_i(t) - \bar{b}(t))^T$, $y_j(t+1) \leftarrow y_j(t) + 2X_t r_{a(t),j}(t)$, and $\bar{\mu}_j(t+1) \leftarrow B_j(t+1)^{-1} y_j(t+1)$.
    end if
  end for
end for

In the parameter estimation step, we propose a novel estimator $\widehat{\mu}_{j_t}(t)$ for the j_t-th user, which is constructed as follows. Define $\mathcal{T}_{j,t} = \{\tau : j_\tau = j, 1 \le \tau \le t\}$, i.e., $\mathcal{T}_{j,t}$ collects time indices when user j is served up to time t. We first calculate an unadjusted user-specific estimator $\bar{\mu}_k(t)$ $(k = 1, \dots, n)$ proposed by

$$\bar{\mu}_k(t) = B_k(t)^{-1} \sum_{\tau \in \mathcal{T}_{k,t-1}} 2X_\tau r_{a(\tau),k}(\tau), \tag{6}$$

where

$$B_k(t) = \widehat{\Sigma}_{k,t} + \Sigma_{k,t} + \lambda l_{kk} I_d, \tag{7}$$

$X_\tau = b_{a(\tau)}(\tau) - \mathbb{E}(b_{a(\tau)}(\tau)|\mathcal{F}_{\tau-1})$, $\widehat{\Sigma}_{k,t} = \sum_{\tau \in \mathcal{T}_{k,t-1}} X_\tau X_\tau^T$, and $\Sigma_{k,t} = \sum_{\tau \in \mathcal{T}_{k,t-1}} \mathbb{E}(X_\tau X_\tau^T|\mathcal{F}_{\tau-1})$, $k = 1, \dots, n$. The expectation in X_τ and $\Sigma_{k,t}$ originates from the randomness of $a(t)$ given $\mathcal{F}_{t-1}$. The definition of $\bar{\mu}_k(t)$ coincides with calculating a regularized version of Kim and Paik [13]'s estimator for each user independently. Then, the main proposed estimator $\widehat{\mu}_{j_t}(t)$ is given by

$$\widehat{\mu}_{j_t}(t) = \bar{\mu}_{j_t}(t) - \lambda B_{j_t}(t)^{-1} \sum_{k \neq j_t} l_{j_t k} \bar{\mu}_k(t). \tag{8}$$

Intuitively, $\widehat{\mu}_{j_t}(t)$ adjusts $\bar{\mu}_{j_t}(t)$ by the neighborhood counterpart according to the graph structure. The designation of (8) is motivated from Yang, Toni, and Dong [26] and carefully constructed so that the estimation error can be expressed in terms of three different types of martingales (with respect to $\mathcal{F}_{t-1}$), $\eta_{j_t,\tau}(\tau)$, X_τ, and D_τ as follows:

$$\widehat{\mu}_{j_t}(t) - \mu_{j_t} = B_{j_t}(t)^{-1} \left[c_{j_t} - \lambda \Delta_{j_t} + \sum_{k=1}^n \left\{ M_{j_t k} \sum_{\tau \in \mathcal{T}_{k,t-1}} \left(X_\tau \eta_{j_t,\tau}(\tau) + A_k(\tau) \right) \right\} \right],$$

where $A_k(t) = \sum_{\tau \in \mathcal{T}_{k,t-1}} D_\tau \mu_k + \sum_{\tau \in \mathcal{T}_{k,t-1}} 2X_\tau \left(v_k(\tau) + \bar{b}(\tau)^T \mu_k \right)$, $D_\tau = X_\tau X_\tau^T - \mathbb{E}(X_\tau X_\tau^T|\mathcal{F}_{\tau-1})$, $M_{jk} = I_d$ if $j = k$ and $\lambda l_{jk} B_k(t)^{-1}$ if $j \neq k$, and c_j is a constant term bounded by λ. Centering induces X_τ which in turn absorbs non-stationary term, $v_k(\tau)$. Detailed proof of sketch is provided in the next Section. The tuning parameter λ controls the influence of the graph structure. For a larger λ, (8) indicates that adjacent nodes more profoundly affect on $\widehat{\mu}_{j_t}$. Our regret analysis does not make any assumptions based on λ, except for $\lambda > 0$.

In the Thompson sampling step, we propose to sample $\widetilde{\mu}_{j_t}(t)$ from $\mathcal{N}_d(\widehat{\mu}_{j_t}(t), v_{j_t}^2 \Gamma_{j_t}(t)^{-1})$, where

$$\Gamma_j(t) = B_j(t) + \lambda^2 \sum_{k \neq j} l_{jk}^2 B_k(t)^{-1}. \tag{9}$$

The form of $\Gamma_j(t)$ is analogous to [26]'s variance formula that was developed under the stationary reward model ($v_j(t) = 0$). If the parameters for all users are estimated simultaneously under regularization by a graph Laplacian and each error term $\eta_{i,j}(t)$ independently follows a mean-zero homoscedastic distribution, it can be shown that $\Gamma_1(t), \Gamma_2(t), \cdots, \Gamma_n(t)$ are approximations for the block diagonals of the precision matrix of the estimate of μ, where $\mu \in \mathbb{R}^{nd}$ concatenates $\mu_1, \cdots, \mu_n$ into one vector. Detailed derivation of (9) is described in Appendix B. Since each $B_k(t)$ is positive definite, it holds that $\Gamma_{j_t}(t)^{-1} < B_{j_t}(t)^{-1}$. This intuitively means that $\Gamma_{j_t}(t)$ contains more information than $B_{j_t}(t)$ by incorporating the neighborhood information. As a result, our proposed sampling searches over narrower region around $\widehat{\mu}_{j_t}(t)$ than the sampling with variance $v_{j_t}^2 B_{j_t}(t)^{-1}$. This leads to an improvement of regret up to a factor less than one compared to an algorithm without graph, as we will see in the next Section. Finally, we select the arm $a(t)$ that satisfies $a(t) = \mathrm{argmax}_i\{b_i(t)^T \widetilde{\mu}_{j_t}(t)\}$.

It is worth mentioning that the proposed estimator $\widehat{\mu}_{j_t}$ and Thompson sampling step are *local*, in a sense that we run the procedure only for user j_t at each time, not for the entire users. The idea of local update appears natural because we have no updated information about the other nodes at time t.

The terms related to the conditional expectation can be calculated as follows. We define $\pi_i(t)$ as the probability of choosing the i-th arm at time t, that is, $\pi_i(t) = \mathbb{P}(a(t)=i|\mathcal{F}_{t-1})$. This is determined by the posterior distribution of $\widetilde{\mu}_{j_t}(t)$, which calls for the evaluation of an integral of a multivariate normal density on a polytope. One may employ well-known approximation algorithms for the integral, for example, Wilhelm and Manjunath [25] and Botev [4]. In our experiments on both synthetic and real data, the Monte Carlo approximation performed well. Once $\pi_i(t)$ is obtained, we can calculate $\mathbb{E}(b_{a(t)}(t)|\mathcal{F}_{t-1}) = \mathbb{E}(\sum_{i=1}^N I(a(t)=i)b_i(t)|\mathcal{F}_{t-1}) = \sum_{i=1}^N \pi_i(t)b_i(t)$. Similarly, $\mathbb{E}(X_tX_t^T|\mathcal{F}_{t-1}) = \sum_{i=1}^N \pi_i(t)(b_i(t) - \bar{b}(t))(b_i(t) - \bar{b}(t))^T$, where $\bar{b}(t) = \mathbb{E}(b_{a(t)}(t)|\mathcal{F}_{t-1})$.

The computation complexity of the proposed algorithm is $O(d^2N + d^2\deg(j_t) + M(d^2 + dN))$ if we use the Monte Carlo approximation for evaluating $\pi_i(t)$, where M is the number of Monte Carlo samples. Note that the complexity does not depend on n; thus, the proposed algorithm is scalable for large graphs, provided that the average degree of nodes is in a moderate range. To see why, first, $\widehat{\mu}_{j_t}(t)$ and $\Gamma_{j_t}(t)$ in (8) requires $O(d^2\deg(j_t))$ computations given $\bar{\mu}_k(t)$. As for $\bar{\mu}_k(t)$ and $B_k(t)$, note that $B_j(t) = B_j(t-1)$ and $\bar{\mu}_j(t) = \bar{\mu}_j(t-1)$ if $j \neq j_t$. Thus, $\bar{\mu}_k(t)$ and $B_k(t)$ is computed only for $k = j_t$, which requires $O(d^2N)$ operations. In addition, the Thompson sampling step and the approximation for $\pi_i(t)$ cost $O(M(d^2+dN))$. To compare with the fastest algorithms in similar settings, Kim and Paik [13] and Yang, Toni, and Dong [26] require $O(d^2N + M(d^2 + dN))$ and $O(d^2\deg(j_t))$ operations, respectively. Although the proposed algorithm has slightly increased order, in the Experiments Section, we demonstrate that the actual runtime of the proposed method is comparable to those fastest algorithms.

4. Regret analysis

We present the high-probability regret upper bound for the proposed SemiGraphTS algorithm. A sketch of proof is provided for a key step. The complete proof can be found in Appendices Appendix C and Appendix D in the Supplement Material. The regret bound for SemiGraphTS is described in the following theorem.

Theorem 1. *Assume* (2) *and* $\delta \in (0,1)$. *Under the semi-parametric linear reward model* (1)*, with probability* $1-\delta$*, the cumulative regret from* SemiGraphTS *(Algorithm 1) achieves*

$$R(T) \leq \sum_{j=1}^{n} O\left(\Psi_{j,T}\left\{\sqrt{d\log(|\mathcal{T}_{j,T}|)}+\sqrt{\lambda}\|\Delta_j\|\right\}\times\min\left\{\sqrt{d\log(dT)},\sqrt{\log(NT)}\right\}\sqrt{d|\mathcal{T}_{j,T}|\log(|\mathcal{T}_{j,T}|)}\right), \tag{10}$$

where $\Psi_{j,T} = \sum_{t\in\mathcal{T}_{j,T}}\|X_t\|_{\Gamma_j(t)^{-1}} / \sum_{t\in\mathcal{T}_{j,T}}\|X_t\|_{B_j(t)^{-1}}$.

We note that $\Psi_{j,T} \in (0,1)$ due to $\Gamma_{j_t}(t)^{-1} < B_{j_t}(t)^{-1}$. A simpler representation of our regret is $\widetilde{O}(\max_j\Psi_{j,T}\cdot d\sqrt{nT}\min\{\sqrt{d},\sqrt{\log(N)}\})$, if we assume $|\mathcal{T}_{j,T}| \approx T/n$ (each j_t is uniformly chosen at random). Compared to the regret bound derived in Yang, Toni, and Dong [26] for the linear graph bandit model, ours has an additional $\min\{\sqrt{d\log(dT)},\sqrt{\log(NT)}\}$ due to the Thompson sampling; other parts are the same, although our model has additional nonparametric intercept $\nu_j(t)$. Running Kim and Paik [13]'s algorithm for each user independently under the same setting leads to the same form of regret bound with (10), except for the term $\Psi_{j,T}\{\sqrt{d\log(|\mathcal{T}_{j,T}|)}+\sqrt{\lambda}\|\Delta_j\|\}$ is replaced with $\sqrt{d\log(|\mathcal{T}_{j,T}|)} + \sqrt{\lambda}\|\mu_j\|$. Since $\Psi_{j,T} \in (0,1)$, the regret bound of the propose algorithm is strictly lower than that from running Kim and Paik [13] independently, provided $\|\Delta_j\| \leq \|\mu_j\|$.

The outline of the proof for Theorem 1 follows Agrawal and Goyal [3] and Kim and Paik [13]. Major modifications are made at establishing a high-probability bound for $\widehat{\mu}_{j_t}(t) - \mu_{j_t}$, as stated in the theorem below.

Theorem 2. *Assume that the settings for the semi-parametric linear reward model* (1) *hold along with* (2). *Let* $E^{\widehat{\mu}}(t)$ *be an event satisfying*

$$E^{\widehat{\mu}}(t) = \left\{\forall i : |b_i^c(t)^T(\widehat{\mu}_{j_t}(t) - \mu_{j_t})| \leq s_{i,j_t}^c(t)\alpha(t)\right\},$$

where $b_i^c(t) = b_i(t) - \bar{b}(t)$, $s_{i,j_t}^c(t) = \|b_i^c(t)\|_{\Gamma_{j_t}(t)^{-1}}$ *and*

$$\alpha(t) = (4R+12)\sqrt{2d\log\left\{\frac{24t^4}{\delta}\left(1+\frac{1}{\lambda}\right)\right\}} + \sqrt{2\lambda}(1+\|\Delta_{j_t}\|).$$

For all $\delta \in (0,1)$ *and* $t \geq 1$, $\mathbb{P}(E^{\widehat{\mu}}(t)) \geq 1 - \delta/t^2$.

The proof for Theorem 2 carefully leverages the structures of $B_j(t)$ and $\Gamma_j(t)$. First, the lemma below enables us to induce $s_{i,j_t}^c(t)$ from $|b_i^c(t)^T(\widehat{\mu}_{j_t}(t) - \mu_{j_t})|$ while encapsulating the other terms into quadratic forms associated with $B_{j_t}(t)^{-1}$.

Lemma 3. *For any* $x, y \in \mathbb{R}^d$ *and* $j = 1, \ldots, n$,

$$x^T B_j(t)^{-1} y \leq \sqrt{2}\|x\|_{\Gamma_j(t)^{-1}}\|y\|_{B_j(t)^{-1}}.$$

Proof. For simplicity, let $B_j = B_j(t)$ and $\Gamma_j = \Gamma_j(t)$ for all j. By the Cauchy-Schwartz inequality, $x^T B_j^{-1} y = x\Gamma_j^{-\frac{1}{2}} \Gamma_j^{\frac{1}{2}} B_j^{-1} y \le \|x\|_{\Gamma_j^{-1}} \sqrt{y^T B_j^{-1} \Gamma_j B_j^{-1} y}$. Note that $B_j^{-1} \le (\lambda l_{jj})^{-1} I_d$. Then, by (7) and (9),

$$B_j^{-1}\Gamma_j = I_d + \sum_{k\neq j} \lambda^2 l_{jk}^2 B_j^{-1} B_k^{-1} \le I_d + \sum_{k\neq j} l_{jk}^2/(l_{jj} l_{kk}) I_d. \tag{11}$$

By (3), we have $\sum_{k\neq j} l_{jk}^2/(l_{jj}l_{kk}) \le 1$. So, the bound in (11) is further bounded by $2I_d$. Therefore, $\sqrt{y^T B_j^{-1}\Gamma_j B_j^{-1} y} \le \sqrt{2}\|y\|_{B_j^{-1}}$. This concludes the proof. □

Then, we utilize the lemma below to simplify random quadratic forms caused by neighboring users' intermediate estimators $\bar{\mu}_k(t)$ $(k \neq j_t)$.

Lemma 4. *For any $x \in \mathbb{R}^d$ and $j, k = 1, \dots, n$,*

$$\|B_k(t)^{-1} x\|_{B_j(t)^{-1}} \le \|x\|_{B_k(t)^{-1}} / \sqrt{\lambda^2 l_{jj} l_{kk}}.$$

Proof. By (7), it suffices to show $(uI_d + A)(vI_d + B)(uI_d + A) \ge uv(uI_d + A)$ for any scalars $u, v > 0$ and positive semi-definite matrices A, B. Observe that $(uI_d + A)(vI_d + B)(uI_d + A) = v(uI_d + A)^2 + (uI_d + A)B(uI_d + A) \ge v(uI_d + A)^2 = u^2 v(I_d + u^{-1}A)^2 \ge u^2 v(I_d + u^{-1}A) = uv(uI_d + A)$, which completes the proof. □

Finally, we separately bound each of the simplified terms by employing the technique of Abbasi-Yadkori, Pal, and Szepesvári [2]. We apply a union bound argument to obtain a uniform bound.

Sketch of proof for Theorem 2. Detailed derivations for key inequalities are provided in Appendix C. Suppose that the semi-parametric reward model (1) holds. Fix t and δ. Let $\hat{\mu}_{j_t}(t)$, $B_k(t)$ and $\bar{\mu}_k(t)$ be as in (8), and (6). For simplification, we write as $b_\tau = b_{a(\tau)}(\tau)$ and $\eta_\tau = \eta_{a(\tau),j_\tau}(\tau)$ for $\tau = 1, \dots, t-1$, and $j = j_t$ with slight abuse of notation. By algebra and Lemma 3,

$$|b_i^c(t)^T(\hat{\mu}_j(t) - \mu_j)| \le \sqrt{2} s_{i,j}^c(t) \sum_{l=1}^{6} C_l, \tag{12}$$

where

$$\begin{aligned} C_1 &= \| \textstyle\sum_{k=1}^n \lambda l_{jk}\mu_k \|_{B_j(t)^{-1}}, & C_2 &= \| \textstyle\sum_{k\neq j} \lambda l_{jk} B_k(t)^{-1} \lambda l_{kk}\mu_k \|_{B_j(t)^{-1}}, \\ C_3 &= \| \textstyle\sum_{\tau\in\mathcal{T}_{j,t-1}} X_\tau \eta_\tau \|_{B_j(t)^{-1}}, & C_4 &= \| \textstyle\sum_{k\neq j} \lambda l_{jk} B_k(t)^{-1} \sum_{\tau\in\mathcal{T}_{j,t-1}} X_\tau\eta_\tau \|_{B_j(t)^{-1}}, \\ C_5 &= \|A_j(t)\|_{B_j(t)^{-1}}, & C_6 &= \| \textstyle\sum_{k\neq j} \lambda l_{jk} B_k(t)^{-1} A_k(t) \|_{B_j(t)^{-1}}, \end{aligned} \tag{13}$$

with

$$A_k(t) = \sum_{\tau\in\mathcal{T}_{k,t-1}} D_\tau \mu_k + \sum_{\tau\in\mathcal{T}_{k,t-1}} 2X_\tau \left(\nu_k(\tau) + \bar{b}(\tau)^T \mu_k\right),$$

$k = 1, \dots, n$, and $D_\tau = X_\tau X_\tau^T - \mathbb{E}(X_\tau X_\tau^T | \mathcal{F}_{\tau-1})$.

For C_1, we have $C_1 \le \sqrt{\lambda}\|\Delta_j\|$ from $B_j(t)^{-1} \le (\lambda l_{jj})^{-1} I_d$. For C_2, from Lemma 4, we have $C_2 \le \sqrt{\lambda} \sum_{k\neq j}(|l_{jk}|/\sqrt{l_{jj}})\|\mu_k\|$ and so $C_2 \le \sqrt{\lambda}$ by $\|\mu_k\| \le 1$ and (3). To bound C_3 and C_4, we first observe that applying Lemma 4 to C_4 yields

$$C_3 + C_4 \le 2\sum_{k=1}^n \frac{|l_{jk}|}{\sqrt{l_{jj}l_{kk}}} \left\| \sum_{\tau\in\mathcal{T}_{k,t-1}} X_\tau \eta_\tau \right\|_{B_k(t)^{-1}}. \tag{14}$$

Next, for each k, Lemma A.1 in Appendix A yields the following with probability at least $1 - \delta(|\mathcal{T}_{k,t-1}| + 1/n)/3t^3$:

$$\left\| \sum_{\tau\in\mathcal{T}_{k,t-1}} X_\tau \eta_\tau \right\|_{B_k(t)^{-1}} \le R\sqrt{d \log\left\{ \frac{24t^4}{\delta}\left(1 + \frac{1}{\lambda l_{kk}}\right)\right\}}. \tag{15}$$

Since $\sum_{k=1}^n (|\mathcal{T}_{k,t-1}| + 1/n) = t$, a union bound argument shows that event (15) holds for all $k = 1, \dots, n$ with probability at least $1 - \delta/t^2$. Under this event, (14) and along with (3) yields

$$C_3 + C_4 \le 4R\sqrt{d \log\left\{ \frac{24t^4}{\delta}\left(1 + \frac{1}{\lambda}\right)\right\}}. \tag{16}$$

Now, for C_5 and C_6, applying Lemma 4 to C_6 leads to

$$C_5 + C_6 \le \sum_{k=1}^{n} \frac{|l_{jk}|}{\sqrt{l_{jj} l_{kk}}} \left\| A_k(t) \right\|_{B_k(t)^{-1}}. \tag{17}$$

To bound $\|A_k(t)\|_{B_k(t)^{-1}}$, we first use the definition for a fixed k,

$$\left\| A_k(t) \right\|_{B_k(t)^{-1}} \le 2 \left\| \sum_{\tau \in \mathcal{T}_{k,t-1}} X_\tau \big(\nu_k(\tau) + \bar{b}(\tau)^T \mu_k \big) \right\|_{B_k(t)^{-1}} + \left\| \sum_{\tau \in \mathcal{T}_{k,t-1}} D_\tau \mu_k \right\|_{B_k(t)^{-1}}.$$

Using the fact that X_τ and D_τ are mean-zero random variables given $\mathcal{F}_{\tau-1}$, we can follow the techniques in Theorem 4.2 of Kim and Paik [13] to bound each term in the right-hand side of the equation above. Then, by a union bound argument,

$$\left\| A_k(t) \right\|_{B_k(t)^{-1}} \le 6 \sqrt{d \log \left\{ \frac{24 t^4}{\delta} \left(1 + \frac{1}{\lambda l_{kk}} \right) \right\}} \tag{18}$$

uniformly for all $k = 1, \dots, n$ with probability at least $1 - 2\delta/(3t^2)$. Combining (17), (18) and the definition of random-walk Laplacian (3), we have with probability at least $1 - 2\delta/(3t^2)$

$$C_5 + C_6 \le 12 \sqrt{d \log \left\{ \frac{24 t^4}{\delta} \left(1 + \frac{1}{\lambda} \right) \right\}}. \tag{19}$$

Finally, plugging the bounds of C_1, C_2, (16), and (19) into (12) completes the proof. □

Remark 1. In our proof, we utilized the definition of the random-walk normalized Laplacian (3) to deduce (i) $\sum_{k \ne j} l_{jk}^2 / (l_{jj} l_{kk}) \le 1$ in (11); (ii) $\sum_{k=1}^{n} |l_{jk}| / \sqrt{l_{jj} l_{kk}} = 2$ in both (14) and (17); and (iii) $l_{jj} = 1$.

Remark 2. We note that the SemiGraphTS algorithm and Theorem 1 is applicable to weighted and/or direct graphs by generalizing the definition of the random-walk Laplacian (3). First, we consider an undirected weighted graph $\mathcal{G}$. Let $W = [w_{jk}]_{1 \le j,k \le n}$ represent the (weighted) adjacency matrix of $\mathcal{G}$. The degree matrix D of $\mathcal{G}$ is $D = \mathrm{diag}(\sum_{k=1}^{n} w_{1k}, \dots, \sum_{k=1}^{n} w_{nk})$. Based on the definition of the random-walk normalized Laplacian $L = D^{-1}(D - W)$, each component of L is written as

$$l_{jj} = 1, \quad l_{jk} = -\frac{w_{jk}}{\sum_{k=1}^{n} w_{jk}} \quad (k \ne j). \tag{20}$$

This extended definition continues to satisfy the three properties mentioned in Remark 1. Thus, the entire proof for Theorem 1 remains valid. Next, assume that $\mathcal{G}$ represents a directed and weighted graph with the weighted adjacency matrix $W = [w_{jk}]_{1 \le j,k \le n}$. Here, w_{jk} denotes the arrow weight from node j to k, where $1 \le j,k \le n$. Let $D = \mathrm{diag}(\sum_{k=1}^{n} w_{1k}, \dots, \sum_{k=1}^{n} w_{nk})$ be the out-degree matrix of $\mathcal{G}$. Alternatively, one can define the reverse, where w_{jk} represents the weight from k to j, and D corresponds to the in-degree matrix. In either case, based on the definition, the random-walk normalized Laplacian L can once again be expressed as (20), and properties (i) through (iii) in Remark 1 still hold. Therefore, the proposed algorithm with L defined in (20) continues to satisfy Theorem 1.

Remark 3. Our proposed regret bound may not hold if other Laplacian representations are used. For instance, suppose that L was defined using the symmetric normalized Laplacian instead of the random-walk normalized representation (3). If the given graph represents a hub graph in which the first node is connected to all other nodes, then $\sum_{k=1}^{n} |l_{1k}| / \sqrt{l_{11} l_{kk}} \approx \sqrt{n}$, which is not bounded by a constant. Therefore, the final regret bound exhibits an increased order in n. See also Yang, Toni, and Dong [26] for further discussion.

Remark 4. In deriving the regret bound in Theorem 1, we assumed that $\pi_i(t)$ can be exactly computed, as in Kim and Paik [13]. This assumption appears reasonable since we can choose arbitrary precision to approximate $\pi_i(t)$. The additional regret caused by the uncertainty of finite Monte Carlo samples can be absorbed in the current bound; detailed discussion is provided in Appendix E.

5. Experiments

We compared the proposed SemiGraphTS with algorithms for (i) semi-parametric bandits without exploiting graph, (ii) linear bandits exploiting graph, and (iii) linear bandits without graph. For (i), we included running Kim and Paik [13] independently on n users to fully personalize recommendations ("SemiTS-Ind"), running a single instance of Kim and Paik [13] for all users to synchronize recommendations across users ("SemiTS-Sin"). For (ii), we considered a Laplacian regularization-based method (Yang, Toni, and Dong 26, namely "GraphUCB") and clustering-based methods (Li et al. 17, "SCLUB"; Li, Wu, and Wang 15, "DyClu"). For (iii), we included "LinTS-Ind" and "LinTS-Sin", running Agrawal and Goyal [3] in "independent" and "single" fashions. Every bandit algorithm involves a hyperparameter that controls the degree of exploration, either through the variance of $\tilde{\mu}(t)$ in the TS-type algorithms (e.g. v_j in our algorithm) or through the confidence width in the UCB-type algorithms. In graph-based and independent bandit algorithms, we use the same value across users, i.e., $v_j = v$. Another hyperparameter is λ, which controls the strength incorporating the graph structure. We tuned (v, λ) by a grid search for first t_0 rounds, with $v \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$ and

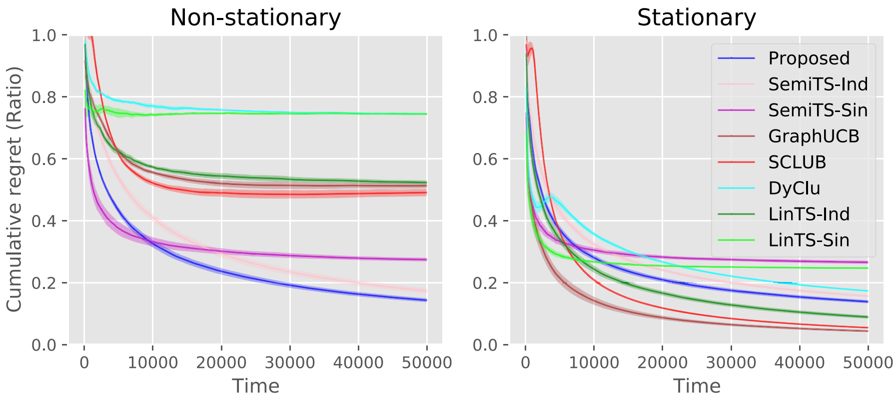

Fig. 1. Current cumulative regrets under the non-stationary scenario (left) and the stationary scenario (right). All regrets are relative to that of the random selection.

$\lambda \in \{5^{-3}, 5^{-2}, 5^{-1}, 5^{0}, 5^{1}\}$. Then, with the best combination of hyperparameters, we assessed each algorithm for over next T rounds. Other hyperparameters were set as default for each algorithm. All computations were conducted in a workstation with AMD Ryzen 3990X CPU and 256 GB RAM. All results were generated over five replications. In all Figures, we report the average in solid line and the confidence band (average $\pm 1.96 \times$ (standard deviation)/$\sqrt{5}$) in light band.

Synthetic dataset We generated data under (1). We considered $\nu_j(t)$ as $\nu_j(t) = -b_{a*(t)}(t)^T \mu_j$ to simulate a non-stationary scenario and $\nu_j(t) = 0$ for a stationary scenario. We fixed $n = 30, N = 10, d = 40$. For each time t, we chose j_t uniformly at random. We constructed the item features as $b_i(t) = (I(i=1)z_1(t)^T, I(i=2)z_2(t)^T, \dots, I(i=N)z_N(t)^T)^T$, where $z_i(t)$ follows a uniform distribution on d'-dimensional sphere ($d' = d/N$). A random error $\eta_{i,j}(t)$ was generated from $\mathcal{N}_d(0, 0.1^2)$. Such settings have been considered in Kim and Paik [13]. Next, the user network $\mathcal{G}$ was generated following the Erdös-Rényi (ER) model, in which the edges were generated independently and randomly with probability p. We set $p = 0.4$. Then we constructed the true user-specific parameters $\mu \in \mathbb{R}^{nd}$ according to $\mu = \mathrm{argmin}_{\mu' \in \mathbb{R}^{nd}} \left[\|\mu' - \mu_0\|^2 + \gamma \mu'^T (L \otimes I_d) \mu' \right]$, where $\mu_0 \in \mathbb{R}^{nd}$ is randomly initialized, L is the random-walk graph Laplacian of $\mathcal{G}$, and $\gamma \geq 0$ [27]. We put $t_0 = 5{,}000$ and $T = 50{,}000$.

Fig. 1 displays the result for the non-stationary scenario with $\gamma = 5$. This scenario satisfies all of our assumptions. As expected, tne proposed SemiGraphTS outperformed other algorithms. Compared to SemiTS-Ind that was the second-best, SemiGraphTS recorded the final cumulative regret decreased by 11.5 percent and the performance gap was larger in small T. This observation is anticipated; since both algorithms specify the reward model correctly, it is natural that the two algorithms will converge to the same result. However, because the proposed algorithm uses a more efficient estimator and exploration strategy by exploiting the graph structure while SemiTS-ind does not, the cumulative regret is significantly smaller than SemiTS-ind when T is small. The third best was SemiTS-Sin, although it performed the best in early rounds. Since SemiTS-Sin estimates only a small number of parameters, the fitted coefficients may have been converging fast to a biased target. Another observation is that SemiGraphTS outperformed the linear graph-based methods. This may suggest that our method could robustly leverage the graph structure when non-stationarity exists. As a next experiment, we tested the same setting but under the stationary scenario $\nu_j(t) = 0$. Note that both linear and semi-parametric algorithms have theoretical guarantees for this case. The result is reported in the right panel of Fig. 1. We see that the linear graph-based algorithms (GraphUCB and SCLUB) outperformed SemiGraphTS. Similarly, LinTS-Ind outperformed SemiTS-Ind. We hypothesize that accommodating the nuisance terms in semi-parametric algorithms may delay convergence of fitted coefficients, which is a price to pay for robustness. In Appendix F, we report additional simulation results based on another choice of $\nu_j(t)$.

For sensitivity analysis, we tested the performances of the algorithms against graph strength and graph misspecification. First, we assessed the performance of the proposed method with varying γ. In practice, the value of γ is unknown and varies largely across different applications. In the left panel of Fig. 2, we tracked the final cumulative regrets for $\gamma = 0$ ($\mu_1, \mu_2, \cdots, \mu_n$ are not enforced to be smooth over the given graph) through $\gamma = 20$ ($\mu_1, \mu_2, \cdots, \mu_n$ are almost the same), under the non-stationary scenario. A larger γ indicates a stronger similarity between μ_j's. For large-γ cases, SemiGraphTS was between those of SemiTS-Ind and SemiTS-Sin. For small-γ cases, SemiGraphTS was comparable to SemiTS-Ind and outperformed SemiTS-Sin with a large margin. In other words, the proposed method performed robustly well in the given range of γ. Second, we assessed the performance of the proposed method with varying number of sign-reversed nodes, varying the proportion of node js in which the signs of μ_j were reversed. In practice, graphs may be misspecified due to unobserved connections, etc. The misspecification proportion is unknown and probably varies across applications. The right panel of Fig. 2 shows the results. When the proportion was large, SemiGraphTS behaved comparably to SemiTS-Ind, while SemiTS-Sin performed poorly. It is natural that SemiTS-Ind performs close to ours when the proportion is high, since in this case the graph information would not be useful any more and the algorithms that do not use this information could perform better. In summary, the proposed method performed robustly well across the given range.

Scalability Fig. 3 reports the average runtime per step of each algorithm, varying the number of users n (left panel), the number of features d (middle panel), and the number of arms N (right panel), fixing other settings the same as in the non-stationary synthetic

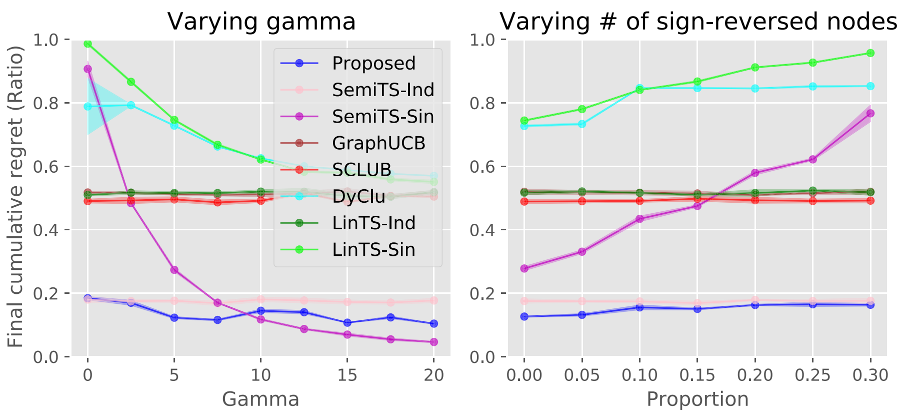

Fig. 2. Final cumulative regrets under the non-stationary scenario, while varying γ (left) and the proportion of sign-reversed nodes (right). All regrets are relative to that of the random selection.

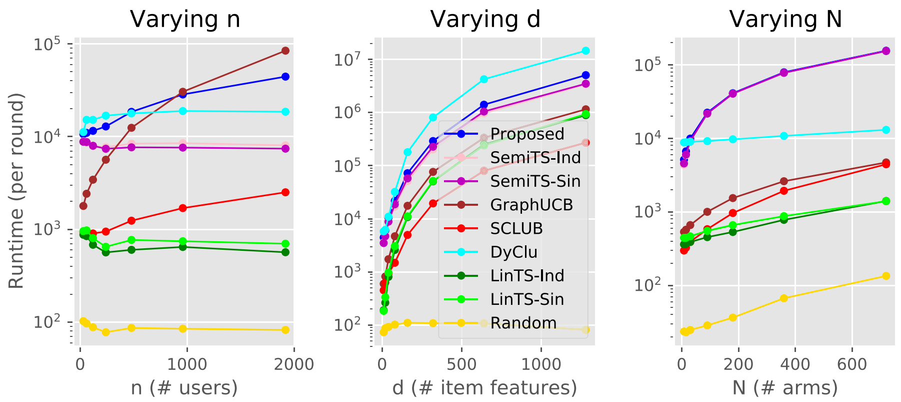

Fig. 3. Average runtimes of the algorithms over varying n (left), d (middle), and N (right).

experiment. SemiGraphTS was slightly slower than SemiTS-Ind. This difference is expected; the construction of $\hat{\mu}_{j_t}(t)$ and $\Gamma_{j_t}(t)$ depends on the degree of the node (user) to serve, which increases linearly with n in the ER graph we tested. A comparison of the semi-parametric methods with the linear methods revealed that each of the semi-parametric methods costed more time than its linear counterparts, mainly due to the Monte Carlo approximation of the arm selection probability. One exception was that SemiGraphTS was faster than GraphUCB as n increases. Overall, SemiGraphTS demonstrated comparable efficiency for large graphs when d and N are in a moderate range.

Real data example The LastFM and Delicious datasets are from a music streaming service last.fm and social bookmarking web service Delicous, released by Cantador, Brusilovsky, and Kuflik [5].[1] The LastFM dataset consists of $n = 1{,}892$ nodes (users) connected by $|E| = 12{,}717$ edges, and 17,632 items (artists) described by 11,946 tags. In the delicious dataset, there are $n = 1{,}861$ nodes (users) with $|E| = 7{,}668$ edges, and 69,226 items (artists) described by 53,388 tags. The datasets contain aggregated tables for the frequencies of (user, artist) pairs, representing the number of times a user listened to any music of an artist. For each dataset, we generated an artificial history of $t_0 = 5{,}000$ and $T = 50{,}000$ rounds following Casa-bianchi et al. [6] and Gentile, Li, and Zappella [10]. In short, we randomly sampled one user to serve and $N = 25$ artists for each round. As item features, we used the first $d = 25$ principal component scores resulting from a term-frequency-inverse-document-frequency (TF-IDF) matrix of artists versus tags, treating artists as "documents" and tags as "words." We set the reward to 1 if the selected user ever listened to a selected artist and 0 otherwise.

Fig. 4 displays the cumulative rewards of the considered algorithms for the two datasets, relative to that of the random selection policy. In the LastFM dataset (left panel), SemiGraphTS produced the best final cumulative reward, 16.7 percent higher value compared to the second-best algorithms. In particular, SemiGraphTS uniformly outperformed SemiTS-Ind and SemiTS-Sin, which we believe that the proposed method might have exploited the graph structure successfully. Compared to the linear graph-based algorithms, SemiGraphTS underperformed GraphUCB in early stages but eventually outperformed them. This result is somewhat anticipated from the synthetic experiment; the presence of nuisance term might have slowed down the learning process of the proposed method but enhanced the robustness of against the change of timely trends. In the Delicious dataset (right panel), Sem-

[1] URLs: http://ir.ii.uam.es/hetrec2011/.

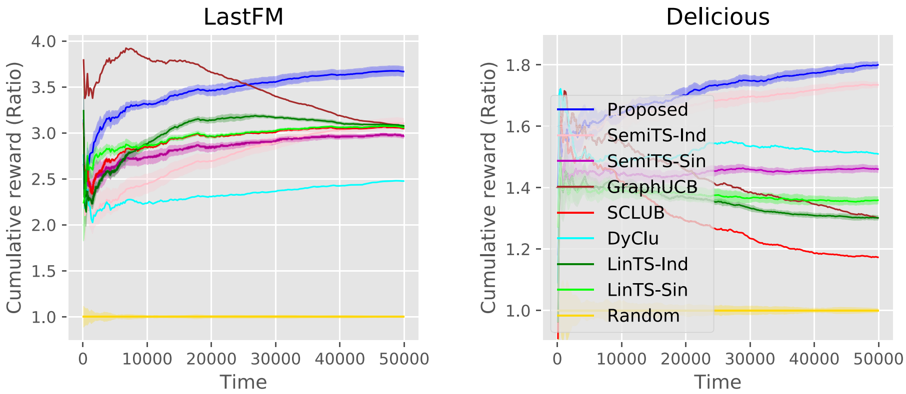

Fig. 4. Current cumulative rewards for the LastFM and Delicious datasets, normalized by the random selection policy.

iGraphTS algorithm again attained the highest final cumulative reward, despite exhibiting lower performance than several other algorithms in the initial stages. Notably, it outperformed other algorithms, except SemiTS-Ind, by a significant margin. This may again suggest the robustness of semi-parametric models against nonstationarity. The performance gap between SemiGraphTS and SemiTS-Ind was relatively small, at 3.7%. This finding is consistent with prior literature which highlights the crucial role of personalization in this dataset, and indicates that leveraging the graph provides only a marginal benefit [6,24]. A hyperparameter sensitivity analysis on both LastFM and Delicious datasets can be found in Appendix G.

To summary the synthetic and real-data experiments, the proposed algorithm appears to robustly achieve desirable performances.

6. Concluding remarks

This study proposes SemiGraphTS, the first algorithm for the semi-parametric contextual bandit MAB problem for multiple users equipped with a graph encoding similarity between user preferences. SemiGraphTS is well suited to more realistic problems in which individual baseline rewards change over time. Experiments demonstrate the potential advantage of SemiGraphTS.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. 2020R1A2C1A01011950 and RS-2023-00252026) and the Institute of Information & communications Technology Planning & evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2020-0-01336, Artificial Intelligence Graduate School Program (UNIST); No. 2022-0-00469, Development of Core Technologies for Task-oriented Reinforcement Learning for Commercialization of Autonomous Drones).

CRediT authorship contribution statement

Young-Geun Choi: Methodology, Software, Validation, Formal Analysis, Origianl draft preparation. **Gi-Soo Kim:** Conceptualization, Software, Review & Editing, Supervision. **Seunghoon Paik:** Software. **Myunghee Cho Paik:** Conceptualization, Review & Editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Myunghee Cho Paik has patent #(KR)10-2021-0103852 pending to Myunghee Cho Paik, Young-Geun Choi, Gi-Soo Kim, Seunghoon Paik. Young-Geun Choi has patent #(KR)10-2021-0103852 pending to Myunghee Cho Paik, Young-Geun Choi, Gi-Soo Kim, Seunghoon Paik. Gi-Soo Kim has patent #(KR)10-2021-0103852 pending to Myunghee Cho Paik, Young-Geun Choi, Gi-Soo Kim, Seunghoon Paik. Seunghoon Paik has patent #(KR)10-2021-0103852 pending to Myunghee Cho Paik, Young-Geun Choi, Gi-Soo Kim, Seunghoon Paik.

Data availability

I have attached the code at the “Attached Files” step. The link for the data for the real-world example is included in the manuscript.

Appendix

In Appendix A, we introduce lemmas for theoretical derivation. In Appendix B, we discuss the rationale behind the variance formula (9) utilized in the Thompson sampling step of our algorithm. In Appendix C, we complete the proof for Theorem 2. In

Appendix D, we provide the proof for Theorem 1. Finally, in Appendix E, we discuss the derivation of the regret bound that addresses the approximation to exact $\pi_i(t)$ by Monte Carlo sampling. In Appendix F, we provide more simulation results. In Appendix G, we analyze a hyperparameter sensitivity of the proposed algorithm on both LastFM and Delicious datasets.

Appendix A. Auxiliary lemmas

Lemma 5 *(Simplified version of Corollary 4.3 in de la Peña, Klass, and Lai [7]). Let $X_\tau \in \mathbb{R}^d$ and $c_\tau \in \mathbb{R}$ be random variables for $\tau = 1, \ldots, t$. Let $A(t) \in \mathbb{R}^{d\times d}$ be a symmetric and positive semi-definite matrix. Suppose that, for all $u \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\exp\left\{u^T \sum_{\tau=1}^{t} X_\tau c_\tau - \frac{1}{2}u^T A(t) u\right\}\right] \le 1.$$

Then, for any $\delta \in (0,1)$ and any symmetric positive definite matrix $Q \in \mathbb{R}^{d\times d}$, the following holds with probability at least $1-\delta$:

$$\left\|\sum_{\tau=1}^{t} X_\tau c_\tau\right\|^2_{(Q+A(t))^{-1}} \le \log\left\{\frac{\det(Q+A(t))/\det(Q)}{\delta^2}\right\}.$$

The lemma below is Lemma 7 in de la Peña, Klass, and Lai [8]. See also Lemma A.3 of Kim and Paik [13] for proof.

Lemma 6. *Let $\{\mathcal{F}_\tau\}_{\tau=1}^t$ be a filtration. Let $X_\tau \in \mathbb{R}^d$ and $c_\tau \in \mathbb{R}$ be $\mathcal{F}_\tau$-measurable random variables such that $\mathbb{E}(X_\tau|\mathcal{F}_{\tau-1}) = 0$, $X_\tau \perp c_\tau|\mathcal{F}_{\tau-1}$, $\|X_\tau\| \le B$, and $\|c_\tau\| \le 1$ for some constant B, $\tau = 1, \ldots, t$. Then, for any $u \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\exp\left\{u^T \sum_{\tau=1}^{t} X_\tau c_\tau - \frac{1}{2}u^T\left(\sum_{\tau=1}^{t} X_\tau X_\tau^T + \sum_{\tau=1}^{t}\mathbb{E}(X_\tau X_\tau^T|\mathcal{F}_{\tau-1})\right)u\right\}\right] \le 1.$$

Lemma 7 *(Azuma-Hoeffding inequality). If $\{M_t\}_{t=0}^T$ is a supermartingale satisfying $|M_t - M_{t-1}| \le c_t$ for all t almost surely, then for any $a > 0$,*

$$\mathbb{P}\left(|M_T - M_0| \ge a\right) \le \exp\left(-\frac{a^2}{2\sum_{t=1}^T c_t^2}\right).$$

Lemma 8 *(Abramowitz and Stegun [1]). If Z is a standard normal random variable, then for any $u \ge 1$,*

$$\frac{1}{2\sqrt{\pi}u}\exp\left(-\frac{u^2}{2}\right) \le \mathbb{P}(|Z| > u) \le \frac{1}{\sqrt{\pi}u}\exp\left(-\frac{u^2}{2}\right).$$

Appendix B. Motivation for the variance formula (9)

We follow the lines of Appendices B and C in Yang, Toni, and Dong [26]. Suppose that we estimate the parameters for all users simultaneously by a semi-parametric estimator with regularization by a graph Laplacian L. Then, we may consider to estimate $\mu := (\mu_1^T, \ldots, \mu_n^T)^T \in \mathbb{R}^{nd}$ by

$$\hat{\mu}^* = \mathbf{M}_t^{-1}\mathbf{\Phi}_t\mathbf{Y}_t.$$

Here, $\mathbf{M}_t$ is a $nd \times nd$ matrix in which the (j,k)-th $d\times d$ block $\mathbf{M}_{jk,t}$ is

$$\mathbf{M}_{jj,t} = B_j(t) = \Sigma_{j,t} + \hat{\Sigma}_{j,t} + \lambda l_{jj} I_d, \quad \mathbf{M}_{jk,t} = \lambda l_{jk} I_d \ (j \ne k).$$

Next, $\mathbf{\Phi}_t$ is a $nd \times (t-1)$ matrix whose τ-th column $\boldsymbol{\phi}_\tau \in \mathbb{R}^{nd}$ is

$$\boldsymbol{\phi}_\tau = [\underbrace{0_d^T, \ \ldots \ , 0_d^T}_{j_\tau - 1 \text{ times}}, \ 2X_\tau^T, \ \underbrace{0_d^T, \ \ldots \ , 0_d^T}_{n - j_\tau \text{ times}}]^T,$$

and $\mathbf{Y}_t \in \mathbb{R}^{t-1}$ is the reward vector, i.e.,

$$\mathbf{Y}_t = [r_{a(1),j_1}(1), r_{a(2),j_2}(2), \ldots, r_{a(t-1),j_{t-1}}(t-1)]^T.$$

Assume that each error term $\eta_{i,j}(\tau)$ independently follows a distribution with mean zero and variance σ^2. Then, as in equation 34 of Yang, Toni, and Dong [26], the precision matrix of $\hat{\mu}^*$ in a fixed design setting, denoted by $\mathbf{\Lambda}_t$, is

$$\mathbf{\Lambda}_t = \frac{1}{\sigma^2}\mathbf{M}_t\mathbf{A}_t^{-1}\mathbf{M}_t,$$

where $\mathbf{A}_t = \mathbf{\Phi}_t \mathbf{\Phi}_t^T \in \mathbb{R}^{nd\times nd}$. Note that the j-th $d\times d$ diagonal block of $\mathbf{A}_t$ is $\mathbf{A}_{j,t} := 4\sum_{\tau\in\mathcal{T}_{j,t-1}} X_\tau X_\tau^T = 4\widehat{\Sigma}_{j,t}$. From equation 38 in Yang, Toni, and Dong [26], the j-th $d\times d$ diagonal block of $\mathbf{\Lambda}_t$ is

$$\mathbf{\Lambda}_{j,t} := \frac{1}{\sigma^2}\left[\mathbf{M}_{jj,t}\mathbf{A}_{j,t}^{-1}\mathbf{M}_{jj,t} + \sum_{k\neq j}\mathbf{M}_{jk,t}\mathbf{A}_{k,t}^{-1}\mathbf{M}_{kj,t}\right].$$

From construction, we recall that $\mathbf{M}_{jj,t} = B_j(t) = \Sigma_{j,t} + \widehat{\Sigma}_{j,t} + \lambda l_{jj} I_d$ and $\mathbf{M}_{jk,t} = \lambda l_{jk} I_d$. Moreover, from the definition of $\mathbf{\Phi}_t$, we have $\mathbf{A}_{j,t} = 4\widehat{\Sigma}_{j,t}$. Thus,

$$\mathbf{\Lambda}_{j,t} = \frac{1}{4\sigma^2}\left[B_j(t)\widehat{\Sigma}_{j,t}^{-1}B_j(t) + \sum_{k\neq j}\lambda^2 l_{jk}^2\widehat{\Sigma}_{k,t}^{-1}\right].$$

Now, we claim that we can approximate $\mathbf{\Lambda}_{j,t}$ by $\frac{1}{2\sigma^2}\Gamma_j(t)$, where $\Gamma_j(t)$ is defined in (9). Note that $\widehat{\Sigma}_{j,t} = \sum_{\tau\in\mathcal{T}_{j,t-1}} X_\tau X_\tau^T$ may be numerically unstable when its inverse is taken, since it is not of full rank at early stages. Therefore, we may want a perturbed version of $\widehat{\Sigma}_{j,t}$, say $\widehat{\Sigma}_{j,t} + cI_d$ for a constant c. In addition, we remark that $\widehat{\Sigma}_{j,t} \approx \Sigma_{j,t}$ when $|\mathcal{T}_{j,t-1}|$ is large. To see this, observe that

$$\begin{aligned}
&\frac{1}{|\mathcal{T}_{j,t-1}|}(\widehat{\Sigma}_{j,t} - \Sigma_{j,t}) \\
&= \frac{1}{|\mathcal{T}_{j,t-1}|}\sum_{\tau\in\mathcal{T}_{j,t-1}}\sum_{i=1}^{N}\left[\{I(a(\tau)=i) - \pi_i(\tau)\}\{b_i(\tau) - \bar{b}(\tau)\}\{b_i(\tau) - \bar{b}(\tau)\}^T\right].
\end{aligned}$$

Since $\{I(a(\tau)=i) - \pi_i(\tau)\}_\tau$ is a martingale adapted to $\{\mathcal{F}_\tau\}_\tau$ and $\|b_i(\tau) - \bar{b}(\tau)\|$ is bounded, $\frac{1}{|\mathcal{T}_{j,t-1}|}(\widehat{\Sigma}_{j,t} - \Sigma_{j,t}) \to 0$ in probability as $|\mathcal{T}_{j,t-1}| \to \infty$ by the Azuma-Hoeffding inequality (Lemma 7). Therefore, taking $c = \lambda l_{jj}/2$ and approximating $\widehat{\Sigma}_{j,t}$ by $(\widehat{\Sigma}_{j,t} + \Sigma_{j,t})/2$, we obtain $\widehat{\Sigma}_{j,t} \approx \frac{\widehat{\Sigma}_{j,t}+\Sigma_{j,t}+\lambda l_{jj}I_d}{2} = B_j(t)/2$. So,

$$\begin{aligned}
\mathbf{\Lambda}_{j,t} &\approx \frac{1}{4\sigma^2}\left[B_j(t)\left(\frac{B_j(t)}{2}\right)^{-1}B_j(t) + \sum_{k\neq j}\lambda^2 l_{jk}^2\left(\frac{B_k(t)}{2}\right)^{-1}\right] \\
&= \frac{1}{2\sigma^2}\left[B_j(t) + \sum_{k\neq j}\lambda^2 l_{jk}^2 B_k(t)^{-1}\right] \\
&= \frac{1}{2\sigma^2}\Gamma_j(t).
\end{aligned}$$

We handle the constant factor $1/2\sigma^2$ in v_j in the Thompson sampling variance. Therefore, the use of $\mathbf{\Lambda}_{j,t}$ is approximately equivalent to the use of $\Gamma_j(t)$.

We note that while other approximations may be considered, our approximation $\widehat{\Sigma}_{j,t} \approx \frac{\widehat{\Sigma}_{j,t}+\Sigma_{j,t}+\lambda l_{jj}I_d}{2} = B_j(t)/2$ could be a favorable choice, as it not only simplifies the expression that leads to $\Gamma_j(t)$ in (9) but also exhibits reasonably good empirical performance.

Appendix C. Proof of Theorem 2

The proof of Theorem 2 follows the sketch in the Regret Analysis Section.

C.1. Proof of (11)

By the semi-parametric reward assumption, for $k = 1,\dots,n$,

$$\begin{aligned}
\bar{\mu}_k(t) &= B_k(t)^{-1}\sum_{\tau\in\mathcal{T}_{k,t-1}} 2X_\tau\left\{\nu(\tau) + b_\tau^T\mu_k + \eta_\tau\right\} \\
&= B_k(t)^{-1}\Bigg\{\sum_{\tau\in\mathcal{T}_{k,t-1}} 2X_\tau\nu(\tau) + \sum_{\tau\in\mathcal{T}_{k,t-1}} 2X_\tau X_\tau^T\mu_k \\
&\qquad + \sum_{\tau\in\mathcal{T}_{k,t-1}} 2X_\tau\bar{b}(\tau)^T\mu_k + \sum_{\tau\in\mathcal{T}_{k,t-1}} 2X_\tau\eta_\tau\Bigg\} \\
&= B_k(t)^{-1}\left\{\left(\widehat{\Sigma}_{k,t} + \Sigma_{k,t}\right) + \left(\widehat{\Sigma}_{k,t} - \Sigma_{k,t}\right) + \lambda l_{kk}I_d - \lambda l_{kk}I_d\right\}\mu_k \\
&\quad + B_k(t)^{-1}\left\{\sum_{\tau\in\mathcal{T}_{k,t-1}} 2X_\tau\left(\nu(\tau) + \bar{b}(\tau)^T\mu_k\right) + \sum_{\tau\in\mathcal{T}_{k,t-1}} 2X_\tau\eta_\tau\right\}
\end{aligned}$$

$$= \mu_k - \lambda l_{kk} B_k(t)^{-1}\mu_k + B_k(t)^{-1}A_k(t) + B_k(t)^{-1}\sum_{\tau\in\mathcal{T}_{k,t-1}} 2X_\tau\eta_\tau.$$

By the relation above, the proposed estimator satisfies

$$\begin{aligned}
&\hat{\mu}_j(t) - \mu_j \\
&= -\left[B_j(t)^{-1}\sum_{k=1}^{n}\lambda l_{jk}\mu_k\right] + \left[B_j(t)^{-1}\sum_{k\neq j}\lambda^2 l_{jk}l_{kk}B_k(t)^{-1}\mu_k\right] \\
&\quad + 2\left[B_j(t)^{-1}\sum_{\tau\in\mathcal{T}_{k,t-1}} X_\tau\eta_\tau\right] - 2\left[B_j(t)^{-1}\sum_{k\neq j}\lambda l_{jk}B_k(t)^{-1}\sum_{\tau\in\mathcal{T}_{k,t-1}} X_\tau\eta_\tau\right] \\
&\quad + \left[B_j(t)^{-1}A_j(t)\right] - \left[B_j(t)^{-1}\sum_{k\neq j}\lambda l_{jk}B_k(t)^{-1}A_k(t)\right].
\end{aligned}$$

Now, left-multiply $b_i^c(t)^T$ on each side of the equation above and applying Lemma 3 on the six terms in the right-hand side yields the desired result. □

C.2. Proof of (16)

Fix k ($k = 1,\dots,n$). Note that when $\mathcal{F}_{\tau-1}$ and $a(\tau)$ are given, X_τ is fixed and η_τ is R-sub-Gaussian. Then, from (7), for all $u \in \mathbb{R}^d$,

$$\mathbb{E}\left[\exp\left\{u^T X_\tau\left(\frac{\eta_\tau}{R}\right) - \frac{1}{2}u^T X_\tau X_\tau^T u\right\}\middle|\mathcal{F}_{\tau-1}, a(\tau)\right] \le 1.$$

This leads to

$$\mathbb{E}\left[\exp\left\{u^T\sum_{\tau\in\mathcal{T}_{k,t-1}} X_\tau c_\tau - \frac{1}{2}u^T\hat{\Sigma}_{k,t}u\right\}\right] \le 1,$$

which satisfies the assumption of Lemma 5 with the choice of $X_\tau = X_\tau$, $c_\tau = n_\tau/R$, $Q = \lambda l_{kk} I_d + \Sigma_{k,t}$ and $A(t) = \hat{\Sigma}_{k,t}$. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta(|\mathcal{T}_{k,t-1}| + 1/n)/(3t^3)$,

$$\left\|\sum_{\tau\in\mathcal{T}_{k,t-1}} X_\tau\eta_\tau\right\|_{B_k(t)^{-1}} \le R\sqrt{\log\left\{\frac{\det(B_k(t))/\det(\lambda l_{kk}I_d + \Sigma_{k,t})}{(\delta(|\mathcal{T}_{k,t-1}| + 1/n)/3t^3)^2}\right\}}. \tag{C.1}$$

We may assume $|\mathcal{T}_{k,t-1}| \ge 1$, otherwise the left-hand side of (C.1) is zero. The determinant-trace inequality for $\det(B_k(t))$ yields

$$\begin{aligned}
&\det(B_k(t)) \le \left(\frac{\mathrm{tr}(B_k(t))}{d}\right)^d \\
&= \left(\frac{\mathrm{tr}(\lambda l_{kk}I_d) + \sum_{\tau\in\mathcal{T}_{k,t-1}}\mathrm{tr}(X_\tau X_\tau^T + \mathbb{E}(X_\tau X_\tau^T|\mathcal{F}_{\tau-1}))}{d}\right)^d \\
&\le \left(\lambda l_{kk} + \frac{8|\mathcal{T}_{k,t-1}|}{d}\right)^d,
\end{aligned}$$

where we used $\|X_\tau\| \le 2$. On the other hand, since $\Sigma_{k,t}$ is positive semi-definite, we have $\det(\lambda l_{kk}I_d + \Sigma_{k,t}) \ge \det(\lambda l_{kk}I_d) = (\lambda l_{kk})^d$. Then, for $d \ge 2$ and $t \ge 1$,

$$\frac{\det(B_j(t))}{\det(\lambda l_{kk}I_d + \Sigma_{k,t})} \le \left(1 + \frac{8|\mathcal{T}_{k,t-1}|}{d\,\lambda l_{kk}}\right)^d \le 8^d|\mathcal{T}_{k,t-1}|^d\left(1 + \frac{1}{\lambda l_{kk}}\right)^d.$$

Since $d \ge 2$, $t > 1$, $0 < \delta < 1$ and $1 \le |\mathcal{T}_{k,t-1}| \le t$, the right-hand side of (C.1) is further simplified by

$$\begin{aligned}
R\sqrt{\log\left\{\frac{\det(B_k(t))/\det(\lambda l_{kk}I_d)}{(\delta(|\mathcal{T}_{k,t-1}| + 1/n)/3t^3)^2}\right\}} &\le R\sqrt{\log\left\{\frac{t^6}{\delta^2}\frac{8^d 9|\mathcal{T}_{k,t-1}|^d\left(1+\frac{1}{\lambda l_{kk}}\right)^d}{(|\mathcal{T}_{k,t-1}| + 1/n)^2}\right\}} \\
&\le R\sqrt{\log\left\{\frac{t^{3d}}{\delta^d}\frac{8^d 3^d|\mathcal{T}_{k,t-1}|^d\left(1+\frac{1}{\lambda l_{kk}}\right)^d}{(|\mathcal{T}_{k,t-1}| + 1/n)^2}\right\}}
\end{aligned}$$

$$\le R\sqrt{d\log\left\{\frac{24t^3}{\delta}|\mathcal{T}_{k,t-1}|\left(1+\frac{1}{\lambda l_{kk}}\right)\right\}}$$
$$\le R\sqrt{d\log\left\{\frac{24t^4}{\delta}\left(1+\frac{1}{\lambda l_{kk}}\right)\right\}}.$$

Combining (C.1) and the result above, for any $0<\delta<1$, we have

$$\left\|\sum_{\tau\in\mathcal{T}_{k,t-1}}X_\tau\eta_\tau\right\|_{B_k(t)^{-1}}\le R\sqrt{d\log\left\{\frac{24t^4}{\delta}\left(1+\frac{1}{\lambda l_{kk}}\right)\right\}}$$

with probability at least $\ge 1-\delta(|\mathcal{T}_{k,t-1}|+1/n)/(3t^3)$. This concludes the derivation. □

C.3. Proof of (17)

Fix k ($k=1,\dots,n$). Recall the definition of $A_k(t)$,

$$\|A_k(t)\|_{B_k(t)^{-1}}\le 2\left\|\sum_{\tau\in\mathcal{T}_{k,t-1}}X_\tau\big(\nu(\tau)+\bar{b}(\tau)^T\mu_k\big)\right\|_{B_k(t)^{-1}}+\left\|\sum_{\tau\in\mathcal{T}_{k,t-1}}D_\tau\mu_k\right\|_{B_k(t)^{-1}}. \tag{C.2}$$

For the first term of the right-hand side of (C.2), Lemma 6 yields

$$\mathbb{E}\left[\exp\left\{u^T\sum_{\tau\in\mathcal{T}_{k,t-1}}X_\tau c_\tau-\frac{1}{2}u^T\left(\widehat{\Sigma}_{k,t}+\Sigma_{k,t}\right)u\right\}\right]\le 1$$

for any $u\in\mathbb{R}^d$, where $c_\tau=(\nu(\tau)+\bar{b}(\tau)^T\mu_k)/2$. Then, we can apply Lemma 5 with $A(t)=\widehat{\Sigma}_{k,t}+\Sigma_{k,t}$ and $Q=\lambda l_{kk}I_d$ to obtain the following inequality with probability at least $1-\delta(|\mathcal{T}_{k,t-1}|+1/n)/3t^3$:

$$\left\|\sum_{\tau\in\mathcal{T}_{k,t-1}}X_\tau\left(\nu(\tau)+\bar{b}(\tau)^T\mu_k\right)\right\|_{B_k(t)^{-1}}\le 2\sqrt{\log\left\{\frac{\det(B_k(t))/\det(\lambda l_{kk}I_d)}{(\delta(|\mathcal{T}_{k,t-1}|+1/n)/3t^3)^2}\right\}}. \tag{C.3}$$

We can bound (C.3) similarly as in bounding the right-hand side of (C.1). Therefore, with probability at least $1-\delta(|\mathcal{T}_{k,t-1}|+1/n)/3t^3$,

$$\left\|\sum_{\tau\in\mathcal{T}_{k,t-1}}X_\tau\left(\nu(\tau)+\bar{b}(\tau)^T\mu_k\right)\right\|_{B_k(t)^{-1}}\le 2\sqrt{d\log\left\{\frac{24t^4}{\delta}\left(1+\frac{1}{\lambda l_{kk}}\right)\right\}}. \tag{C.4}$$

For the second term of the right-hand side of (C.2), we let $Y_{k,\tau}=D_\tau\mu_k$ and observe $Y_{k,\tau}\in\mathbb{R}^d$, $\mathbb{E}(Y_{k,\tau}|\mathcal{F}_{\tau-1})=0$. It is straightforward from Lemma 4.4 and its proof in Kim and Paik [13] to derive

$$\mathbb{E}\left[\exp\left\{u^T\sum_{\tau\in\mathcal{T}_{k,t-1}}\frac{1}{\sqrt{2}}Y_{k,\tau}-\frac{1}{2}u^T\left(\widehat{\Sigma}_{k,t}+\Sigma_{k,t}\right)u\right\}\right]\le 1$$

for any $u\in\mathbb{R}^d$, which again satisfies the assumption of Lemma 5 with the choice of $X_\tau=Y_{k,\tau}$, $c_\tau=1/\sqrt{2}$ and $A(t)=\widehat{\Sigma}_{k,t}+\Sigma_{k,t}$. Then, putting $Q=\lambda l_{kk}I_d$, we have with probability at least $1-\delta(|\mathcal{T}_{k,t-1}|+1/n)/3t^3$,

$$\left\|\sum_{\tau\in\mathcal{T}_{k,t-1}}D_\tau\mu_k\right\|_{B_k(t)^{-1}}\le\sqrt{2}\sqrt{\log\left\{\frac{\det(B_k(t))/\det(\lambda l_{kk}I_d)}{(\delta(|\mathcal{T}_{k,t-1}|+1/n)/3t^3)^2}\right\}}$$
$$\le\sqrt{2}\sqrt{d\log\left\{\frac{24t^4}{\delta}\left(1+\frac{1}{\lambda l_{kk}}\right)\right\}}. \tag{C.5}$$

Plugging (C.4) and (C.5) into (C.2) yields bounds for each user:

$$\mathbb{P}\left[\|A_k(t)\|_{B_k(t)^{-1}}\le 6\sqrt{d\log\left\{\frac{24t^4}{\delta}\left(1+\frac{1}{\lambda l_{kk}}\right)\right\}}\right]\ge 1-\frac{2\delta(|\mathcal{T}_{k,t-1}|+1/n)}{3t^3}.$$

Finally, applying the union bound argument yields

$$\mathbb{P}\left[\forall k=1,\dots,n\,:\,\|A_k(t)\|_{B_k(t)^{-1}}\le 6\sqrt{d\log\left\{\frac{24t^4}{\delta}\left(1+\frac{1}{\lambda l_{kk}}\right)\right\}}\right]\ge 1-\frac{2\delta}{3t^2}, \tag{C.6}$$

which completes the proof. $\square$

Appendix D. Proof of Theorem 1

The proof incorporates the lines of Agrawal and Goyal [3] and Kim and Paik [13] with the proposed estimation and Thompson sampling steps. Throughout the Section, we write as $j = j_t$, $b_\tau = b_{a(\tau)}(\tau)$ and $\eta_\tau = \eta_{a(\tau),j_\tau}(\tau)$ for brevity. We reserve k ($k = 1,\dots,n$) to denote user index. The proof has six steps:

(a) (Theorem 2) To establish a high-probability upper bound of $|b_i^c(t)(\widehat{\mu}_j(t) - \mu_j)|$.
(b) (Lemma 9) To establish a high-probability upper bound of $|b_i^c(t)(\widetilde{\mu}_j(t) - \widehat{\mu}_j(t))|$ given $\mathcal{F}_{t-1}$.
(c) (Definition 1) To divide arms at each time t into saturated arms and unsaturated arms.
(d) (Lemma 10) To bound the probability of playing saturated arms by a function of playing unsaturated arms.
(e) (Lemma 11) To bound $regret(t)$ given $\mathcal{F}_{t-1}$ for each t.
(f) To bound $R(T)$ and complete the proof.

We begin with step (b).

Lemma 9. *Let $E^{\widetilde{\mu}}(t)$ be an event defined by*

$$E^{\widetilde{\mu}}(t) = \left\{ \forall i : |b_i^c(t)^T(\widetilde{\mu}_j(t) - \widehat{\mu}_j(t))| \le v_j s_{i,j}^c(t) \min\{\sqrt{4d\log(2dT)}, \sqrt{4\log(2NT)}\} \right\},$$

for all $t \ge 1$, $\mathbb{P}(E^{\widetilde{\mu}}(t)|\mathcal{F}_{t-1}) \ge 1 - 1/T^2$.

Proof. We first show $|b_i^c(t)^T(\widetilde{\mu}_j(t) - \widehat{\mu}_j(t))| \le v_j s_{i,j}^c(t)\sqrt{4d\log(2dT)}$. Given $\mathcal{F}_{t-1}$, the values of $b_i^c(t)$, $\Gamma_j(t)$, and $\widehat{\mu}_j(t)$ are fixed. Then, for $i = 1,\dots,N$, we have

$$\begin{aligned}
\left|b_i^c(t)^T(\widetilde{\mu}_j(t) - \widehat{\mu}_j(t))\right| &= \left|v_j b_i^c(t)^T\Gamma_j(t)^{-\frac{1}{2}} \cdot \frac{1}{v_j}\Gamma_j(t)^{\frac{1}{2}}(\widetilde{\mu}_j(t) - \widehat{\mu}_j(t))\right| \\
&\le v_j s_{i,j}^c(t)\left\|\frac{1}{v_j}\Gamma_j(t)^{\frac{1}{2}}(\widetilde{\mu}_j(t) - \widehat{\mu}_j(t))\right\|_2 \\
&= v_j s_{i,j}^c(t)\sqrt{\sum_{l=1}^d Z_l(t)^2},
\end{aligned} \tag{D.1}$$

where $Z_l(t)|\mathcal{F}_{t-1}$ $(l = 1,\dots,d)$ identically and independently follow the standard normal distribution. We apply Lemma 8 with the choice of $u = \sqrt{2\log(2dT^2)}$. Noting $\sqrt{2\log(2dT^2)} \le \sqrt{2\log(2^2d^2T^2)} = \sqrt{4\log(2dT)}$,

$$\begin{aligned}
&\mathbb{P}\left(|Z_l(t)| > \sqrt{4\log(2dT)}\middle|\mathcal{F}_{t-1}\right) \le \mathbb{P}\left(|Z_l(t)| > \sqrt{2\log(2dT^2)}\middle|\mathcal{F}_{t-1}\right) \\
&\le \frac{1}{\sqrt{2\pi\log(2dT^2)}} \cdot \frac{1}{2dT^2} \le \frac{1}{2dT^2},
\end{aligned}$$

for each $l = 1,\dots,d$. Then, by a union bound argument,

$$\mathbb{P}\left(\forall l = 1,\dots,d : |Z_k(t)| > \sqrt{4\log(2dT)}\middle|\mathcal{F}_{t-1}\right) \le \frac{1}{2T^2}. \tag{D.2}$$

Therefore, combining (D.1) and (D.2) yields

$$\mathbb{P}\left(\forall i : \left|b_i^c(t)^T(\widetilde{\mu}_j(t) - \widehat{\mu}_j(t))\right| \le v_j s_{i,j}^c(t)\sqrt{4d\log(2dT)}\right) \ge 1 - \frac{1}{2T^2}.$$

On the other hand, by the observation that $b_i^c(t)^T(\widetilde{\mu}_j(t) - \widehat{\mu}_j(t))|\mathcal{F}_{t-1}$ $(i = 1,\dots,N)$ identically and independently follow the standard normal distribution, one can apply a similar technique to derive $|b_i^c(t)^T(\widetilde{\mu}_j(t) - \widehat{\mu}_j(t))| \le v_j s_{i,j}^c(t)\sqrt{4\log(2NT)}$ with probability at least $1 - 1/(2T^2)$ given $\mathcal{F}_{t-1}$. Combining the two bounds, we obtain the desired result. $\square$

In step (c), we divide arms at each time t into saturated arms and unsaturated arms. Note that $C(t)$ implicitly depends on j_t.

Definition 1. Define $C(t)$, the set of saturated arms, by

$$C(t) = \{i : b_i^c(t)^T\mu_j + g_j(T)s_{i,j}^c(t) < b_{a^*(t)}(t)^T\mu_j\},$$

where $g_k(T) = \alpha_k(T) + v_k\min\{\sqrt{4d\log(2dT)}, \sqrt{4\log(2NT)}\}$ and $\alpha_k(T) = (4R + 12)\cdot \sqrt{d\log\left\{(24T^4/\delta)(1+\lambda^{-1})\right\}} + \sqrt{\lambda}(1 + \|\Delta_k\|)$, $k = 1,\dots,n$.

In step (d), we establish that the probability of playing saturated arms is bounded by the probability of playing unsaturated arms up to constant multiplication and addition.

Lemma 10. *Given $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ is true,*

$$\mathbb{P}\left(a(t)\in C(t)|\mathcal{F}_{t-1}\right)\leq\frac{1}{p}\mathbb{P}\left(a(t)\notin C(t)|\mathcal{F}_{t-1}\right)+\frac{1}{pT^2},$$

where $p=1/(4e\sqrt{\pi})$.

Proof. Since $a(t)=\operatorname{argmax}_{1\leq i\leq N}\{b_i^c(t)^T\widetilde{\mu}_j(t)\}$ by definition, if $b_{a^*(t)}(t)^T\widetilde{\mu}_j(t)>b_i^c(t)^T\widetilde{\mu}_j(t)$ for every $i\in C(t)$, then $a(t)\notin C(t)$. This implies

$$\mathbb{P}\left(a(t)\notin C(t)|\mathcal{F}_{t-1}\right)\geq\mathbb{P}\left(\forall i\in C(t):b_{a^*(t)}(t)^T\widetilde{\mu}_j(t)>b_i^c(t)^T\widetilde{\mu}_j(t)|\mathcal{F}_{t-1}\right). \tag{D.3}$$

On the other hand, when $E^{\widetilde{\mu}}(t)$ is additionally true,

$$\begin{aligned}b_i^c(t)^T\widetilde{\mu}_j(t)&\leq b_i^c(t)^T\mu_j+g_j(T)s_{i,j}^c(t) &&(\text{Def. of } E^{\hat{\mu}}(t)\ \&\ E^{\widetilde{\mu}}(t))\\ &\leq b_{a^*(t)}(t)^T\mu_j, &&(\text{Def. of } C(t)),\end{aligned}$$

which implies that

$$\begin{aligned}&\mathbb{P}\left(b_{a^*(t)}(t)^T\mu_j<b_{a^*(t)}(t)^T\widetilde{\mu}_j(t)|\mathcal{F}_{t-1}\right)\\ &\leq\mathbb{P}\left(\forall i\in C(t):b_i^c(t)^T\widetilde{\mu}_j(t)<b_{a^*(t)}(t)^T\widetilde{\mu}_j(t)|\mathcal{F}_{t-1}\right)+\left(1-\mathbb{P}\left(E^{\widetilde{\mu}}(t)|\mathcal{F}_{t-1}\right)\right)\end{aligned} \tag{D.4}$$

The left-hand side of (D.4) can be lower-bounded, because the normality of $\widetilde{\mu}_j(t)$ and Lemma 8 yields

$$\begin{aligned}&\mathbb{P}\left(b_{a^*(t)}(t)^T\widetilde{\mu}_j(t)>b_{a^*(t)}(t)^T\mu_j\middle|\mathcal{F}_{t-1}\right)\\ &=\mathbb{P}\left(\frac{b_{a^*(t)}(t)^T(\widetilde{\mu}_j(t)-\hat{\mu}_j(t))}{v_js_{a^*(t),j}^c(t)}>\frac{b_{a^*(t)}(t)^T(\mu_j-\hat{\mu}_j(t))}{v_js_{a^*(t),j}^c(t)}\middle|\mathcal{F}_{t-1}\right)\\ &\geq\mathbb{P}\left(Z(t)>\frac{\alpha_j(T)}{v_j}\middle|\mathcal{F}_{t-1}\right)\\ &\geq\frac{1}{4\sqrt{\pi}u}\exp\left(-\frac{u^2}{2}\right),\end{aligned}$$

where $u=\alpha_j(T)/v_j$ and $Z(t)|\mathcal{F}_{t-1}$ is a standard normal random variable. Note that $u\leq 1$ by the construction. Therefore,

$$\mathbb{P}\left(b_{a^*(t)}(t)^T\widetilde{\mu}_j(t)>b_{a^*(t)}(t)^T\mu_j|\mathcal{F}_{t-1}\right)\geq\frac{1}{4e\sqrt{\pi}}=p. \tag{D.5}$$

Combining (D.3), (D.4), (D.5) and Lemma 9, we have

$$\mathbb{P}\left(a(t)\notin C(t)|\mathcal{F}_{t-1}\right)+\frac{1}{T^2}\geq p,$$

which implies

$$\frac{\mathbb{P}\left(a(t)\in C(t)|\mathcal{F}_{t-1}\right)}{\mathbb{P}\left(a(t)\notin C(t)|\mathcal{F}_{t-1}\right)+\frac{1}{T^2}}\leq\frac{1}{p}.$$

This completes the proof. □

Before proceeding to bound the cumulative regret, we bound each $regret(t)$ given $\mathcal{F}_{t-1}$ in step (e).

Lemma 11. *Given $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ is true,*

$$\mathbb{E}\left(regret(t)|\mathcal{F}_{t-1}\right)\leq\frac{5g_j(T)}{p}\mathbb{E}\left(s_{a(t),j}^c(t)|\mathcal{F}_{t-1}\right)+\frac{4g_j(T)}{pT^2}.$$

Proof. Let $\bar{a}(t)=\operatorname{argmin}_{i\notin C(t)}s_{i,j}^c(t)$. If $\mathcal{F}_{t-1}$ is given, then $\bar{a}(t)$ is deterministic. This value is also well-defined due to $a^*(t)\notin C(t)$. Under $\mathcal{F}_{t-1}$ such that both $E^{\hat{\mu}}(t)$ and $E^{\widetilde{\mu}}(t)$ holds,

$$\begin{aligned}&b_{a^*(t)}(t)^T\mu_j\\ &=b_{a^*(t)}(t)^T\mu_j-b_{\bar{a}(t)}(t)^T\mu_j+b_{\bar{a}(t)}(t)^T\mu_j\end{aligned}$$

$$
\begin{aligned}
&\le g_j(T)s^c_{\bar{a}(t),j}(t) + b_{\bar{a}(t)}(t)^T\mu_j && (\bar{a}(t)\notin C(t) \text{ \& def. of } C(t))\\
&\le g_j(T)s^c_{\bar{a}(t),j}(t) + b_{\bar{a}(t)}(t)^T\widetilde{\mu}_j(t) + g_j(T)s^c_{\bar{a}(t),j}(t) && (\text{Def. of } E^{\widehat{\mu}}(t) \text{ and } E^{\widetilde{\mu}}(t))\\
&\le 2g_j(T)s^c_{\bar{a}(t),j}(t) + b_{a(t)}(t)^T\widetilde{\mu}_j(t) && (\text{Def. of } a(t))\\
&\le 2g_j(T)s^c_{\bar{a}(t),j}(t) + b_{a(t)}(t)^T\mu_j + g_j(T)s^c_{a(t),j}(t) && (\text{Def. of } E^{\widehat{\mu}}(t) \text{ and } E^{\widetilde{\mu}}(t)),
\end{aligned}
$$

which yields

$$
regret(t) \le 2g_j(T)s^c_{\bar{a}(t),j}(t) + g_j(T)s^c_{a(t),j}(t).
$$

Then, under $\mathcal{F}_{t-1}$ such that $E^{\widehat{\mu}}(t)$ holds, the following holds from inequality above, Lemma 9 and $|regret(t)| \le 2$:

$$
\begin{aligned}
&\mathbb{E}\left(regret(t)|\mathcal{F}_{t-1}\right)\\
&= \mathbb{E}\left(regret(t)I(E^{\widetilde{\mu}}(t))|\mathcal{F}_{t-1}\right) + \mathbb{E}\left(regret(t)\{1-I(E^{\widetilde{\mu}}(t))\}|\mathcal{F}_{t-1}\right)\\
&\le 2g_j(T)s^c_{\bar{a}(t),j}(t) + g_j(T)\mathbb{E}\left(s^c_{a(t),j}(t)|\mathcal{F}_{t-1}\right) + 2\left(1-\mathbb{P}\left(E^{\widetilde{\mu}}(t)|\mathcal{F}_{t-1}\right)\right)\\
&\le 2g_j(T)s^c_{\bar{a}(t),j}(t) + g_j(T)\mathbb{E}\left(s^c_{a(t),j}(t)|\mathcal{F}_{t-1}\right) + \frac{2}{T^2}.
\end{aligned} \tag{D.6}
$$

We now further bound $s^c_{\bar{a}(t),j}(t)$. Observe that

$$
\begin{aligned}
&s^c_{\bar{a}(t),j}(t)\\
&= s^c_{\bar{a}(t),j}(t)\left\{\mathbb{P}(a(t)\in C(t)|\mathcal{F}_{t-1}) + \mathbb{P}(a(t)\notin C(t)|\mathcal{F}_{t-1})\right\}\\
&= s^c_{\bar{a}(t),j}(t)\left\{\frac{2}{p}\mathbb{P}(a(t)\notin C(t)|\mathcal{F}_{t-1}) + \frac{1}{pT^2}\right\} && (\text{Lemma } 10)\\
&= \frac{2}{p}\mathbb{E}\left(s^c_{\bar{a}(t),j}(t)I(a(t)\notin C(t))|\mathcal{F}_{t-1}\right) + \frac{s^c_{\bar{a}(t),j}(t)}{pT^2}\\
&\le \frac{2}{p}\mathbb{E}\left(s^c_{a(t),j}(t)I(a(t)\notin C(t))|\mathcal{F}_{t-1}\right) + \frac{s^c_{\bar{a}(t),j}(t)}{pT^2} && (\text{Def. of } \bar{a}(t))\\
&\le \frac{2}{p}\mathbb{E}\left(s^c_{a(t),j}(t)|\mathcal{F}_{t-1}\right) + \frac{1}{pT^2}. && (s^c_{i,k}(t)\le 1 \text{ for any } i,k,t)
\end{aligned}
$$

Combining the inequality above and (D.6) conclude the proof. □

In step (f), we complete the proof.

Proof for Theorem 1. Let

$$
M_t := regret(t)I(E^{\widehat{\mu}}(t)) - \frac{5g_{j_t}(T)}{p}s_{a(t),j_t}(t) - \frac{4g_{j_t}(T)}{pT^2},\ t=1,\dots,T,
$$

with $M_0 = 0$.

We apply martingale arguments for each user $k = 1,\dots,n$, and aggregate them by union bound. Fix k and let $T_k = |\mathcal{T}_{k,T}|$. Due to Lemma 11 and $s^c_{i,k}(t) \le 1$, $\{M_t\}_{t\in\{0\}\cup\mathcal{T}_{k,T}}$ is a supermartingale process satisfying $|M_t| \le 10g_k(T)/p$. We apply Lemma 7 with the choice of $c_t = 10g_k(T)/p$ and $a = (10g_k(T)/p)\sqrt{2T_k\log(2T/(\delta T_k))}$ that satisfies $\exp(-a^2/(2\sum_t c_t^2)) = \delta T_k/(2T)$. This yields

$$
\begin{aligned}
&\sum_{t\in\mathcal{T}_{k,T}} regret(t)I(E^{\widehat{\mu}}(t)) \le\\
&\frac{5g_k(T)}{p}\sum_{t\in\mathcal{T}_{k,T}} s^c_{a(t),k}(t) + \frac{4g_k(T)}{pT} + \frac{10g_k(T)}{p}\sqrt{2T_k\log(2T/(\delta T_k))}
\end{aligned} \tag{D.7}
$$

with probability at least $1-\delta T_k/(2T)$. Since $T_1 + \ldots + T_n = T$, a union bound argument over $k = 1,\dots,n$ leads to

$$
\begin{aligned}
&\sum_{t=1}^{T} regret(t)I(E^{\widehat{\mu}}(t)) \le\\
&\sum_{k=1}^{n}\left[\frac{5g_k(T)}{p}\sum_{t\in\mathcal{T}_{k,T}} s_{a(t),k}(t) + \frac{4g_k(T)}{pT} + \frac{10g_k(T)}{p}\sqrt{2T_k\log(2T/(\delta T_k))}\right]
\end{aligned} \tag{D.8}
$$

with probability at least $1-\delta/2$.

On the other hand, we apply a union bound argument to Theorem 2 over $t = 1, \dots, T$ and replace δ with $3\delta/\pi^2$, which yields $\mathbb{P}(E^{\hat{\mu}}(t)$ for all $t = 1, \dots, T) \geq 1 - \delta/2$. Then, $regret(t) I(E^{\hat{\mu}}(t)) = regret(t)$ for every t with probability at least $1 - \delta/2$.

Therefore, with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} regret(t) \leq \sum_{k=1}^{n} \left[\frac{5g_k(T)}{p} \sum_{t \in \mathcal{T}_{k,T}} s^c_{a(t),k}(t) + \frac{4g_k(T)}{pT} + \frac{10g_k(T)}{p} \sqrt{2T_k \log(2T/(\delta T_k))} \right].$$

Now, by Lemma 12 below and the definitions of $g_k(T)$ and p,

$$R(T) \leq \sum_{k=1}^{n} O\Bigg(\Psi_{k,T} \left\{ \sqrt{d \log(|\mathcal{T}_{k,T}|)} + \sqrt{\lambda} \|\Delta_k\| \right\} \times$$

$$\min\left\{ \sqrt{d \log(d|\mathcal{T}_{k,T}|)}, \sqrt{\log(N|\mathcal{T}_{k,T}|)} \right\} \sqrt{d|\mathcal{T}_{k,T}| \log(|\mathcal{T}_{k,T}|)} \Bigg)$$

with probability at least $1 - \delta$, which completes the proof. $\square$

Lemma 12.

$$\sum_{t \in \mathcal{T}_{k,T}} s^c_{a(t),k}(t) = O\left(\Psi_{k,T} \sqrt{d|\mathcal{T}_{k,T}| \log(|\mathcal{T}_{k,T}|)} \right).$$

Proof. We recall that $\Psi_{k,T} = \sum_{t \in \mathcal{T}_{k,T}} \|X_t\|_{\Gamma_k(t)^{-1}} / \sum_{t \in \mathcal{T}_{k,T}} \|X_t\|_{B_k(t)^{-1}}$ and that $\Psi_{k,T} \in (0,1)$ due to $\Gamma_k(t)^{-1} < B_k(t)^{-1}$ for all j, t. Since $\sum_{t \in \mathcal{T}_{k,T}} s^c_{a(t),k}(t) = \sum_{t \in \mathcal{T}_{k,T}} \|X_t\|_{\Gamma_k(t)^{-1}}$ by the definitions of $s^c_{i,k}(t)$ and X_t, we have

$$\sum_{t \in \mathcal{T}_{k,T}} s^c_{a(t),k}(t) = \Psi_{k,T} \sum_{t \in \mathcal{T}_{k,T}} \|X_\tau\|_{B_k(t)^{-1}}.$$

We now claim $\sum_{t \in \mathcal{T}_{k,T}} \|X_\tau\|_{B_k(t)^{-1}} = O\left(\sqrt{d|\mathcal{T}_{k,T}| \log(|\mathcal{T}_{k,T}|)} \right)$. This has been proved in similar settings [2,3,24,13]; for completeness, we present the proof. Define $s_{i,k}(t) = \|b^c_i(t)\|_{B_j(k)^{-1}}$. Note that $s_{a(t),j_t}(t) = X_t$ and $\sum_{t \in \mathcal{T}_{k,T}} \|X_\tau\|_{B_k(t)^{-1}} =$. Then, $\|X_\tau\|_{B_k(t)^{-1}} = \sum_{t \in \mathcal{T}_{k,T}} s_{a(t),k}(t)$. Following the lines for equation 60 of Vaswani, Schmidt, and Lakshmanan [24], we can derive

$$\log\left[\det(B_k(t+1))\right] \geq \log\left[\det(\lambda l_{kk} I_d)\right] + \sum_{\tau \in \mathcal{T}_{k,t}} \log\left(1 + s_{a(\tau),k}(\tau)^2\right). \tag{D.9}$$

On the other hand, the trace of $B_k(t+1)$ is

$$\operatorname{tr}\left(B_k(t+1)\right) \leq 8|\mathcal{T}_{k,T}| + \lambda l_{kk} d, \tag{D.10}$$

where we used $\|X_\tau\| \leq 2$ by construction. Plugging (D.9) and (D.10) into the determinant-trace inequality $\left\{\operatorname{tr}\left(B_k(t+1)\right)/d\right\}^d \geq \det\left(B_k(t+1)\right)$, equivalently $d \log\left\{\operatorname{tr}\left(B_k(t+1)\right)/d\right\} \geq \log\det\left(B_k(t+1)\right)$, we obtain

$$d \log\left(\frac{8|\mathcal{T}_{k,T}|}{d} + \lambda l_{kk} \right) \geq d \log(\lambda l_{kk}) + \sum_{\tau \in \mathcal{T}_{k,t}} \log\left(1 + s_{a(\tau),k}(\tau)^2\right),$$

or,

$$\sum_{\tau \in \mathcal{T}_{k,t}} \log\left(1 + s_{a(\tau),k}(\tau)^2\right) \leq d \log\left(1 + \frac{8|\mathcal{T}_{k,T}|}{d\lambda l_{kk}}\right).$$

Now, we bound $\sum_{\tau \in \mathcal{T}_{k,t}} s_{a(\tau),k}(\tau)^2$ by the result above. First, we have $s_{a(\tau),k}(\tau)^2 \in [0, 1/(\lambda l_{kk})]$ because

$$s_{a(\tau),k}(\tau)^2 = b_\tau^T B_k(\tau)^{-1} b_\tau \leq b_\tau^T \left(\lambda l_{kk} I_d\right)^{-1} b_\tau \leq (\lambda l_{kk})^{-1}.$$

Considering a function $f(t) = \log(1+t) / \left[\lambda l_{kk} \log\left(1 + (\lambda l_{kk})^{-1}\right)\right]$, f satisfies $t \leq f(t)$ for all $t \in [0, 1/(\lambda l_{kk})]$. Therefore,

$$\begin{aligned}\sum_{\tau \in \mathcal{T}_{k,t}} s_{a(\tau),k}(\tau)^2 &\leq \frac{1}{\lambda l_{kk} \log\left(1 + (\lambda l_{kk})^{-1}\right)} \sum_{\tau \in \mathcal{T}_{k,t}} \log\left(1 + s_{a(\tau),k}(\tau)^2\right) \\ &\leq \frac{d}{\lambda l_{kk} \log\left(1 + (\lambda l_{kk})^{-1}\right)} \log\left(1 + \frac{|\mathcal{T}_{k,T}|}{d\lambda l_{kk}}\right).\end{aligned}$$

Algorithm 2 A special case of the SemiGraphTS algorithm that approximates $\pi_i(t)$ by the Monte Carlo sampling (SemiGraphTS-MC).

Fix $\lambda > 0$ and M. Set $B_j(1) = \lambda l_{jj} I_d$, $y_j(1) = 0_d$ and $v_j = (4R+12)\sqrt{d \log\left\{(24T^4/\delta)(1+\lambda^{-1})\right\}} + \sqrt{\lambda}(1+\|\Delta_j\|)$ for $j = 1, \ldots, n$.
for $t = 1, 2, \ldots, T$ **do**
  Observe j_t.
  for $j = 1, 2, \ldots, n$ **do**
    if $j \neq j_t$ **then**
      Update $B_j(t+1) \leftarrow B_j(t)$, $\bar{\mu}_j(t+1) \leftarrow \bar{\mu}_j(t)$, and $y_j(t+1) \leftarrow y_j(t)$.
    else
      $\widehat{\mu}_j(t) \leftarrow \bar{\mu}_j(t) - B_j(t)^{-1} \sum_{k\neq j} \lambda l_{jk} \bar{\mu}_k(t)$.
      $\Gamma_j(t) \leftarrow B_j(t) + \lambda^2 \sum_{k\neq j} l_{jk}^2 B_k(t)^{-1}$
      for $m = 1, 2, \cdots, M$ **do**
        Sample $\widetilde{\mu}_j^m(t)$ from $\mathcal{N}_d(\widehat{\mu}_j(t), v_j^2 \Gamma_j(t)^{-1})$
      end for
      for $i = 1, 2, \cdots, N$ **do**
        Compute $\widehat{\pi}_i(t) = \frac{1}{M}\sum_{m=1}^M I\left\{i = \mathrm{argmax}_k \{b_k(t)^T \widetilde{\mu}_j^m(t)\}\right\}$
      end for
      Sample $a(t)$ from $\mathrm{Multinom}(\widehat{\pi}_1(t), \cdots, \widehat{\pi}_N(t))$.
      $\bar{b}(t) \leftarrow \sum_{i=1}^N \widehat{\pi}_i(t) b_i(t)$ and $X_t \leftarrow b_{a(t)}(t) - \bar{b}(t)$.
      Update $B_j(t+1) \leftarrow B_j(t) + X_t X_t^T + \sum_{i=1}^N \widehat{\pi}_i(t)(b_i(t) - \bar{b}(t))(b_i(t) - \bar{b}(t))^T$, $y_j(t+1) \leftarrow y_j(t) + 2X_t r_{a(t),j}(t)$, and $\bar{\mu}_j(t+1) \leftarrow B_j(t+1)^{-1} y_j(t+1)$.
    end if
  end for
end for

Finally, from the Cauchy-Schwartz inequality and the result above,

$$\sum_{\tau \in \mathcal{T}_{k,t}} s_{a(\tau),k}(\tau) \leq \sqrt{8|\mathcal{T}_{k,T}|}\sqrt{\sum_{\tau \in \mathcal{T}_{k,t}} s_{a(\tau),k}(\tau)^2}$$
$$\leq \sqrt{\frac{d|\mathcal{T}_{k,T}|}{\lambda l_{kk} \log\left(1 + (\lambda l_{kk})^{-1}\right)} \log\left(1 + \frac{|\mathcal{T}_{jk,t}|}{d\lambda l_{kk}}\right)}.$$

Since $l_{kk} = 1$ by the definition of the random-walk Laplacian,

$$\sum_{\tau \in \mathcal{T}_{k,t}} s_{a(\tau),k}(\tau) \leq \sqrt{\frac{d|\mathcal{T}_{k,T}|}{\lambda \log\left(1 + \frac{1}{\lambda}\right)} \log\left(1 + \frac{8|\mathcal{T}_{k,T}|}{d\lambda}\right)},$$

which proves the claim and concludes the proof. □

Appendix E. Regret bound when $\pi_i(t)$ is approximated by Monte Carlo sampling

In this section, we analyze the additional regret induced by approximation and show that the regret upper bound of the alternative algorithm has the same order as the bound of SemiGraphTS.

Our discussion is based on Algorithm 2, a special case of the SemiGraphTS algorithm (Algorithm 1), that explicitly states that we use the Monte Carlo approximated values of $\pi_i(t)$ for action selection. Before action selection, Algorithm 2 computes first the Monte Carlo approximates of $\pi_i(t)$. We denote the approximated value as $\widehat{\pi}_i(t)$. Then, Algorithm 2 samples the arm from a multinomial distribution with size 1, say $\mathrm{Multinom}(\widehat{\pi}_1(t), \cdots, \widehat{\pi}_N(t))$. In comparison, Algorithm 1 samples $a(t) \sim \mathrm{Multinom}(\pi_1(t), \cdots, \pi_N(t))$.

We now discuss the regret bound for Algorithm 2. We highlight the key differences from following the lines of Appendix D. Let the filtration $\mathcal{F}_{t-1}$ further include all Monte Carlo samples up to time $t-1$.

For step (a), Theorem 2 directly holds with $\pi_i(t)$'s replaced with $\widehat{\pi}_i(t)$'s since the approximated values $\widehat{\pi}_i(t)$'s are now the true probabilities of the arm selection.

Steps (b)-(e) in Appendix D exploited that the arm is selected from the exact probability. In other words, those results were derived if we select arm according to $\tilde{a}(t) = \mathrm{argmax}_{1\leq i\leq N}\{b_i(t)^T \widetilde{\mu}_{j_t}(t)\}$ (i.e., $\widetilde{a}(t) \sim \mathrm{Multinom}(\pi_1(t), \cdots, \pi_N(t))$). Now we show through an inductive argument that the remaining proofs are still valid with the new arm selection $a(t) \sim \mathrm{Multinom}(\widehat{\pi}_1(t), \cdots, \widehat{\pi}_N(t))$.

Suppose that until round $t-1$, we have sampled arms $a(\tau) \sim \mathrm{Multinom}(\widehat{\pi}_1(\tau), \cdots, \widehat{\pi}_N(\tau))$, $\tau = 1, \cdots, t-1$. Then we have the desired high-probability upper bound for the estimate $\hat{\mu}_j(t)$ for every $j = 1, \cdots, n$ (Theorem 2). Now suppose that at round t, we sample the arm $\tilde{a}(t) = \mathrm{argmax}_{1\leq i\leq N}\{b_i(t)^T \widetilde{\mu}_{j_t}(t)\}$. Then the proofs (b)-(e) go through, and by Lemma 13 we have,

$$\mathbb{E}\left((b_{a^*(t)}(t)^T \mu_{j_t} - b_{\tilde{a}(t)}(t)^T \mu_{j_t}) I(E^{\hat{\mu}}(t)) | \mathcal{F}_{t-1}\right)$$
$$\leq \frac{5g_{j_t}(T)}{p}\mathbb{E}\left(s^c_{\tilde{a}(t),j_t}(t) | \mathcal{F}_{t-1}\right) + \frac{4g_{j_t}(T)}{pT^2}. \tag{E.1}$$

Then, given $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ is true,

$$
\begin{aligned}
&\mathbb{E}\left(regret(t)|\mathcal{F}_{t-1}\right)\\
&=\mathbb{E}\left(b_{a^*(t)}(t)^T\mu_{j_t}-b_{a(t)}(t)^T\mu_{j_t}|\mathcal{F}_{t-1}\right)\\
&=\mathbb{E}\left((b_{a^*(t)}(t)^T\mu_{j_t}-b_{\tilde{a}(t)}(t)^T\mu_{j_t})+(b_{\tilde{a}(t)}(t)^T\mu_{j_t}-b_{a(t)}(t)^T\mu_{j_t})|\mathcal{F}_{t-1}\right)\\
&\le\frac{5g_{j_t}(T)}{p}\mathbb{E}\left(s^c_{\tilde{a}(t),j_t}(t)|\mathcal{F}_{t-1}\right)+\frac{4g_{j_t}(T)}{pT^2}+\mathbb{E}\left(b_{\tilde{a}(t)}(t)^T\mu_{j_t}-b_{a(t)}(t)^T\mu_{j_t}|\mathcal{F}_{t-1}\right)\\
&=\frac{5g_{j_t}(T)}{p}\mathbb{E}\left(s^c_{a(t),j_t}(t)|\mathcal{F}_{t-1}\right)+\frac{4g_{j_t}(T)}{pT^2}+\mathbb{E}\left(b_{\tilde{a}(t)}(t)^T\mu_{j_t}-b_{a(t)}(t)^T\mu_{j_t}|\mathcal{F}_{t-1}\right)\\
&\quad+\frac{5g_{j_t}(T)}{p}\mathbb{E}\left(s^c_{\tilde{a}(t),j_t}(t)-s^c_{a(t),j_t}(t)|\mathcal{F}_{t-1}\right).
\end{aligned}
$$

As compared to Lemma 13 for Algorithm 1, we have two additional terms to bound; for step (f), those terms appear in the final cumulative regret. We claim below that the cumulative sum of the two additional terms have lower order than the original regret bound of Algorithm 1. We first have,

$$
\begin{aligned}
&\mathbb{E}\left(b_{\tilde{a}(t)}(t)^T\mu_{j_t}-b_{a(t)}(t)^T\mu_{j_t}|\mathcal{F}_{t-1}\right)\\
&=\mathbb{E}\left(\sum_{i=1}^N b_i(t)^T\mu_{j_t}I(\tilde{a}(t)=i)-\sum_{i=1}^N b_i(t)^T\mu_{j_t}I(a(t)=i)\middle|\mathcal{F}_{t-1}\right)\\
&=\sum_{i=1}^N b_i(t)^T\mu_{j_t}\pi_i(t)-\sum_{i=1}^N b_i(t)^T\mu_{j_t}\mathbb{E}(\hat{\pi}_i(t)|\mathcal{F}_{t-1})\\
&=\sum_{i=1}^N b_i(t)^T\mu_{j_t}\pi_i(t)-\sum_{i=1}^N b_i(t)^T\mu_{j_t}\pi_i(t)=0,
\end{aligned}
$$

which is due to unbiasedness of the Monte-Carlo estimate $\hat{\pi}_i(t)$. Since we also have $||b_{\tilde{a}(t)}(t)^T\mu_{j_t}-b_{a(t)}(t)^T\mu_{j_t}||\le 2$, we can show from the Azuma-Hoeffding inequality, with high probability,

$$
\sum_{t=1}^T\left\{b_{\tilde{a}(t)}(t)^T\mu_{j_t}-b_{a(t)}(t)^T\mu_{j_t}\right\}\le O(\sqrt{T}). \tag{E.2}
$$

Similarly, we have

$$
\begin{aligned}
&\frac{5g_{j_t}(T)}{p}\mathbb{E}\left(s^c_{\tilde{a}(t),j_t}(t)-s^c_{a(t),j_t}(t)|\mathcal{F}_{t-1}\right)\\
&=\frac{5g_{j_t}(T)}{p}\mathbb{E}\left(\sum_{i=1}^N s^c_{i,j_t}(t)I(\tilde{a}(t)=i)-\sum_{i=1}^N s^c_{i,j_t}(t)I(a(t)=i)\middle|\mathcal{F}_{t-1}\right)\\
&=\frac{5g_{j_t}(T)}{p}\sum_{i=1}^N s^c_{i,j_t}(t)(\pi_i(t)-\pi_i(t))=0.
\end{aligned}
$$

Hence with high probability,

$$
\sum_{t=1}^T\frac{5g_{j_t}(T)}{p}\left\{s^c_{\tilde{a}(t),j_t}(t)-s^c_{a(t),j_t}(t)\right\}\le\frac{5\max_j g_j(T)}{p}O(\sqrt{T}). \tag{E.3}
$$

We remark that the right-hand sides of (E.2) and (E.3) does not depend on d, n, N nor the graph structure. Therefore, our claim holds.

Appendix F. More simulation results

To further support the empirical findings in the non-stationary scenario in Section 5, we conducted additional experiments with a different choice for $\nu_j(t)$. We considered a random time-varying intercept $\nu_j(t)\sim\mathcal{N}(-\{(j \bmod 2)+1\},0.1^2)$. Except $\nu_j(t)$, we used the same data generating process as used in the main experiment. Fig. F.5 displays the cumulative regret under $\gamma=5$, and Fig. F.6 reports the sensitivity analysis results against the graph strength and graph misspecification. Overall, the results are consistent with the case of $\nu_j(t)=-b_{a*(t)}(t)^T\mu_j$ in the main body.

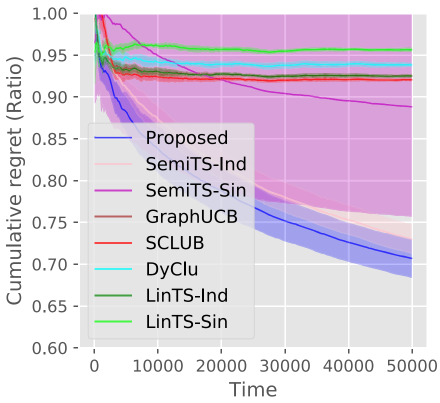

Fig. F.5. Current cumulative regrets under the non-stationary scenario with $v_j(t) \sim \mathcal{N}(-\{(j \bmod 2)+1\}, 0.1^2)$. All regrets are relative to that of the random selection.

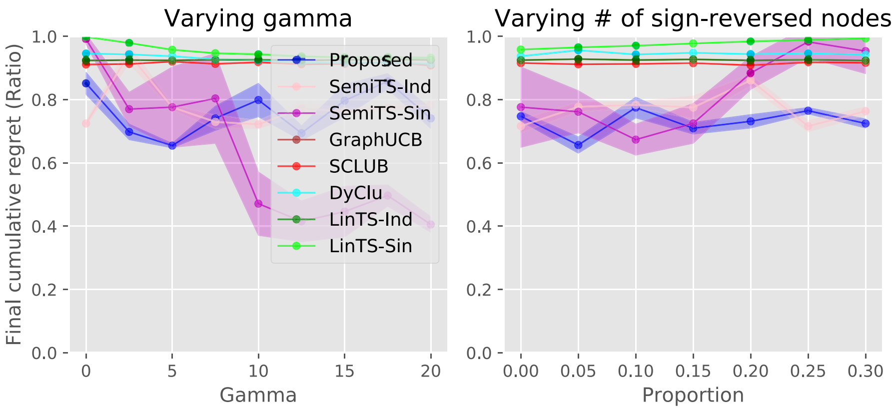

Fig. F.6. Final cumulative regrets under the non-stationary scenario with $v_j(t) \sim \mathcal{N}(-\{(j \bmod 2)+1\}, 0.1^2)$, while varying γ (left) and the proportion of sign-reversed nodes (right). All regrets are relative to that of the random selection.

Appendix G. Hyperparameter sensitivity on real datasets

We assessed how sensitive the performance of our proposed algorithm is to different hyperparameter values when applied to LastFM and Delicious datasets.

Table G.1 presents the cumulative reward (relative to a random selection algorithm) achieved by running our proposed algorithm for 50,000 rounds using different pairs of hyperparameters (v, λ). When we tried different values for one hyperparameter (either v or λ), we kept the other parameter fixed at the optimal value found in the original experiment in Section 5.

In Delicious dataset, the performance was relatively stable over different choice of v and λ. On the other hand, the performance in the LastFM dataset decayed in large values of v, while it was less sensitive to the choice of λ. This may suggest that tuning v, the degree of exploration, could be crucial in practice.

Table G.1
Hyperparameter sensitivity on two real datasets: Final cumulative rewards (normalized by the random selection policy) for different pairs of hyperparameters (v, λ). When we tried different values for one hyperparameter (either v or λ), we kept the other parameter fixed at the optimal value found in the original experiment in Section 5.

Varying v			Varying λ		
	LastFM	Delicious		LastFM	Delicious
$v = 5^{-3}$	3.70 ± 0.07	1.74 ± 0.02	$\lambda = 10^{-3}$	2.62 ± 0.05	1.35 ± 0.02
$v = 5^{-2}$	3.16 ± 0.06	1.81 ± 0.04	$\lambda = 10^{-2}$	2.76 ± 0.05	1.36 ± 0.01
$v = 5^{-1}$	3.11 ± 0.07	1.57 ± 0.01	$\lambda = 10^{-1}$	2.97 ± 0.03	1.56 ± 0.02
$v = 5^{0}$	2.27 ± 0.03	1.37 ± 0.01	$\lambda = 10^{0}$	3.35 ± 0.08	1.72 ± 0.02
$v = 5^{1}$	1.69 ± 0.04	1.35 ± 0.01	$\lambda = 10^{1}$	3.67 ± 0.06	1.83 ± 0.03

References

[1] M. Abramowitz, I. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, United States Department of Commerce, National Bureau of Standards, 1964.
[2] Y. Abbasi-Yadkori, D. Pal, C. Szepesvári, Improved algorithms for linear stochastic bandits, in: Advances in Neural Information Processing Systems, vol. 24, 2011, pp. 2312–2320.
[3] S. Agrawal, N. Goyal, Thompson sampling for contextual bandits with linear payoffs, in: Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 1164–1172.
[4] Z.I. Botev, The normal law under linear restrictions: simulation and estimation via minimax tilting, J. R. Stat. Soc., Ser. B, Stat. Methodol. 79 (1) (2017) 125–148.
[5] I. Cantador, P. Brusilovsky, T. Kuflik, 2nd workshop on information heterogeneity and fusion in recommender systems (HetRec 2011), in: Proceedings of the 5th ACM Conference on Recommender Systems, 2011.
[6] N. Casa-bianchi, C. Gentile, N. Cesa-Bianchi, C. Gentile, G. Zappella, A gang of bandits, in: Advances in Neural Information Processing Systems, vol. 26, 2013, pp. 737–745.
[7] V.H. de la Peña, M.J. Klass, T.L. Lai, Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws, Ann. Probab. 32 (3) (2004) 1902–1933.
[8] V.H. de la Peña, M.J. Klass, T.L. Lai, Theory and applications of multivariate self-normalized processes, Stoch. Process. Appl. 119 (12) (2009) 4210–4227.
[9] J. Delporte, A. Karatzoglou, T. Matuszczyk, S. Canu, Socially enabled preference learning from implicit feedback data, in: Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, 2013, pp. 145–160.
[10] C. Gentile, S. Li, G. Zappella, Online clustering of bandits, in: Proceedings of the 31st International Conference on Machine Learning, 2014, pp. 2296–2315.
[11] K. Greenewald, A. Tewari, P. Klasnja, S. Murphy, Action centered contextual bandits, in: Advances in Neural Information Processing Systems, vol. 31, 2017, pp. 5978–5986.
[12] E. Kaufmann, N. Korda, R. Munos, Thompson sampling: an asymptotically optimal finite-time analysis, in: International Conference on Algorithmic Learning Theory, 2012, pp. 199–213.
[13] G.S. Kim, M.C. Paik, Contextual multi-armed bandit algorithm for semiparametric reward model, in: Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 3389–3397.
[14] A. Krishnamurthy, Z.S. Wu, V. Syrgkanis, Semiparametric contextual bandits, in: Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 4330–4349.
[15] C. Li, Q. Wu, H. Wang, Unifying clustered and non-stationary bandits, in: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, vol. 130, 2021, pp. 1–26.
[16] L. Li, W. Chu, J. Langford, R.E. Schapire, A contextual-bandit approach to personalized news article recommendation, in: Proceedings of the 19th International Conference on World Wide Web, 2010, pp. 661–670.
[17] S. Li, W. Chen, S. Li, K.-S. Leung, Improved algorithm on online clustering of bandits, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019, pp. 2923–2929.
[18] P. Liao, K. Greenewald, P. Klasnja, S. Murphy, Personalized HeartSteps: a reinforcement learning algorithm for optimizing physical activity, in: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2020, pp. 1–22.
[19] Y. Peng, M. Xie, J. Liu, X. Meng, N. Li, C. Yang, T. Yao, R. Jin, A practical semi-parametric contextual bandit, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019, pp. 3246–3252.
[20] N. Rao, H.-F. Yu, P.K. Ravikumar, I.S. Dhillon, Collaborative filtering with graph information: consistency and scalable methods, in: Advances in Neural Information Processing Systems, vol. 28, 2015, pp. 2107–2115.
[21] S.L. Scott, A modern Bayesian look at the multi-armed bandit, Appl. Stoch. Models Bus. Ind. 26 (6) (2010) 639–658.
[22] L. Tang, R. Rosales, A. Singh, D. Agarwal, Automatic ad format selection via contextual bandits, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, 2013, pp. 1587–1594.
[23] A. Tewari, S.A. Murphy, From ads to interventions: contextual bandits in mobile health, in: Mobile Health, Springer International Publishing, 2017, pp. 495–517.
[24] S. Vaswani, M. Schmidt, L.V. Lakshmanan, Horde of bandits using Gaussian Markov random fields, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017, pp. 690–699.
[25] S. Wilhelm, B.G. Manjunath, tmvtnorm: a package for the truncated multivariate normal distribution generation of random numbers computation of marginal densities, R J. 2 (1) (2010) 25–29.
[26] K. Yang, L. Toni, X. Dong, Laplacian-regularized graph bandits: algorithms and theoretical analysis, in: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, 2020, pp. 3133–3143.
[27] Y. Yankelevsky, M. Elad, Dual graph regularized dictionary learning, IEEE Trans. Signal Inf. Process. Netw. 2 (4) (2016) 611–624.