Parallel Online Clustering of Bandits via Hedonic Game

Xiaotong Cheng [1] **Cheng Pan** [1] **Setareh Maghsudi** [1]

Abstract

Contextual bandit algorithms appear in several applications, such as online advertisement and recommendation systems like personalized education or personalized medicine. Individually-tailored recommendations boost the performance of the underlying application; nevertheless, providing individual suggestions becomes costly and even implausible as the number of users grows. As such, to efficiently serve the demands of several users in modern applications, it is imperative to identify the underlying users' clusters, i.e., the groups of users for which a single recommendation might be (near-)optimal. We propose CLUB-HG, a novel algorithm that integrates a game-theoretic approach into clustering inference. Our algorithm achieves Nash equilibrium at each inference step and discovers the underlying clusters. We also provide regret analysis within a standard linear stochastic noise setting. Finally, experiments on synthetic and real-world datasets show the superior performance of our proposed algorithm compared to the state-of-the-art algorithms.

1. Introduction

Multi-armed bandit (MAB) problems, which formalize the exploration-exploitation trade-off (Slivkins et al., 2019), have been under intensive investigations in the past decade (Li et al., 2019). In its seminal setting, an agent acts at each round and receives an instantaneous reward. The agent aims to develop an action selection policy to maximize its cumulative rewards over all rounds. The MAB framework finds application in facing decision-making problems in various areas such as clinical trials, dynamic pricing, computational advertising, web-page content optimization, and recommendation systems (Gentile et al., 2017; Bouneffouf et al., 2020).

[1]Department of Computer Science, University of Tübingen, Tübingen, Germany. Correspondence to: Xiaotong Cheng <xiaotong.cheng@uni-tuebingen.de>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

To develop decision-making strategies for large-scale applications, it is conventional to assume a linear structure between actions and rewards thanks to their simplicity and effectiveness (Li et al., 2019; Liu et al., 2022). In the stochastic linear bandit (Auer, 2002; Dani et al., 2008; Abbasi-Yadkori et al., 2011), each arm (action) has a feature vector known as "context". The expected payoff associated with each arm is an unknown linear function of the feature vector. The agent must infer the unknown linear function based on the context and payoff information and select accordingly. The contextual linear bandit is a reference model for adaptive recommendation systems (Li et al., 2010; Caron et al., 2012).

Standard contextual bandit algorithms usually focus on a single user and make recommendations only based on the individual's historical data (Li et al., 2010; Abbasi-Yadkori et al., 2011). However, such an approach suffers from several shortcomings in serving many users with multiple recommendable items in modern applications (Mahadik et al., 2020). On the one hand, the large number of users and items increases the computational burden significantly. On the other hand, the sparse and insufficient observations concerning the single agent yield a weak estimation of its characteristics and degrade the decision-making performance (Yang et al., 2020). To address such shortcomings, it is imperative to identify a few subgroups/communities within which the users share similar characteristics and utilize their collective effect for high-quality, speedy, and dynamic recommendations. In other words, discovering the clustering structure of users can reduce the computational burden and improve the quality of recommendations.

References (Gentile et al., 2014; Li et al., 2016) initially investigated the clustering of bandits, where the algorithms adaptively learn the clustering structure over users based on the estimated user similarities. The proposed algorithms use graph vertices to represent users and consider the users connected by edges to belong to the same cluster; Nevertheless, these algorithms have several limitations. First, since they use connected components to represent clusters, two clusters can become completely disjoint if all edges between them are removed. Besides, splitting clusters is irreversible, i.e., once users are (erroneously) separated into different clusters, they cannot be aggregated again (Gentile et al., 2017). Furthermore, most of the bandit-based

recommendation algorithms sequentially perform decision-making for a dynamically selected user, which is neither parallel nor distributed (Gentile et al., 2014; 2017; Li et al., 2016; McInerney et al., 2018; Li et al., 2019). That leads to high synchronization costs for users in the same cluster (Mahadik et al., 2020) and long delays in services.

To tackle the challenges above, we propose an algorithm for parallel online clustering of bandits, which formulates the clustering problem as a hedonic game. Game-theoretic concepts have been widely used in network clustering recently (Papadimitriou, 2001; Bu et al., 2017), and the hedonic game has emerged as a qualified model for clustering (Aziz et al., 2015; Feldman et al., 2015; Bilò et al., 2018; Aziz et al., 2019). Thus, we adopt a hedonic game-theoretic approach for clustering tasks and regard each user as an independent player. A self-organized clustering is obtained by following the decisions of each independent player (user). Our main contributions are as follows.

- We propose a parallel online clustering of bandits via a hedonic game (CLUB-HG) formulation, which combines a multi-agent contextual bandit algorithm with a novel hedonic game-theoretic clustering approach. To our best knowledge, this paper is the first attempt to adopt a game-theoretic approach to bandit clustering.
- We analyze the Nash equilibrium of the formulated clustering game at each inference step and provide rigorous proofs about achieving the Nash equilibrium. Besides, we prove that, under standard assumptions, for a sufficiently long horizon T, the clustering algorithm converges to the true underlying clustering structure, and the regret bound of the CLUB-HG algorithm is sublinear with T.
- We perform intensive experiments on both synthetic and real-world datasets and demonstrate the superior performance of our algorithm compared with state-of-the-art methods.

1.1. Related Work

A large body of literature investigates the clustering of bandits. Nguyen & Lauw (2014) propose a clustering-based contextual bandit algorithm based on the K-Means clustering technique. In Gentile et al. (2014), the authors develop the CLUB algorithm and demonstrate the effectiveness of using graph structures to represent the clustering structures of users. Li et al. (2016) and Gentile et al. (2017) extend this work. In Li et al. (2016), the collaborative effects on users and items are taken into account. Besides user clustering, the proposed algorithm also considers the clustering of items to address the case with many items. Gentile et al. (2017) study the clustering of the bandit problem in a context-aware manner. The proposed algorithm, CAB, adaptively matches user preferences in the face of a constantly evolving item universe, which differs from other literature by inducing different clusters according to item vectors. In addition, Li et al. (2019) generalize the setting in previous work by considering the non-uniform frequency distribution of user selections. Two operations for clustering, split, and merge, are designed to identify the underlying clusters.

In real-world applications, many users await recommendations simultaneously, which yields high synchronization costs and excessive latency. Therefore, it becomes crucial to develop clustering of bandit algorithms with efficient computational ability in a parallel or distributed manner. Korda et al. (2016) use a gossip-based protocol to create a distributed variant of the CLUB algorithm of Gentile et al. (2014) in peer-to-peer networks. DistUCB, a novel distributed algorithm proposed in Mahadik et al. (2020), eliminates the need for large buffer transfers in Korda et al. (2016) and produces better performance in identifying the cluster structure. Moreover, Liu et al. (2022) study the online clustering of bandits in a federated setting and proposes a bandit learning algorithm FCLUB that protects the privacy and is aware of communication requirements.

Most of the previous studies on the clustering of bandits compare the difference between feature vectors with a predefined threshold and performing split or merge functions. Thus, the clustering performance highly relies on threshold settings. To address this limitation, we adopt a game-theoretic approach for clustering tasks. The hedonic game is one of the most popular concepts in clustering, which is introduced in the field of economics as a model of coalition formation (Feldman et al., 2015). The clustering problem can be a "clustering game" (Bulò & Pelillo, 2009; Feldman et al., 2015), where the players correspond to the elements to be clustered, and the notion of clustering is equivalent to a classical game-theoretic equilibrium/stability concept (Bulò & Pelillo, 2009; Aziz et al., 2015; Feldman et al., 2015; Balliu et al., 2017). The players have preferences over coalitions that might merge them, and the outcome is the disjoint partitions of the agent set, referred to as a clustering or coalition structure (Feldman et al., 2015; Balliu et al., 2017). The clustering problem has been studied by a series of works considering the problem setting on fractional hedonic game (Aziz et al., 2015; 2019), the fixed and correlation clustering (Feldman et al., 2015), the social distance game variation (Balliu et al., 2017). Those works study the existence of the Nash equilibrium, the upper and lower bounds on the price of anarchy, and dynamics leading to the Nash equilibrium. The closest work to the clustering method in our framework is the correlation clustering game in Feldman et al. (2015). There, a player's utility depends on its similarity to other players in its cluster and its dissimilarity to players in other clusters, where the similarity is captured by the distance metric. The authors prove that a Nash equilibrium always exists as the convergence point of

the best response dynamics (Feldman et al., 2015).

2. Notation and Preliminaries

We use boldface lowercase letters and boldface uppercase letters to represent vectors and matrices respectively. For example, $\|\boldsymbol{x}\|_p$ denotes the $p-$norm of a vector $\boldsymbol{x}$. For a symmetric positive semi-definite matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, the weighted $2-$norm of vector $\boldsymbol{x} \in \mathbb{R}^d$ is defined by $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}$. The inner product is denoted by $\langle\cdot,\cdot\rangle$ and the weighted inner-product $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_{\boldsymbol{A}} = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y}$. **Table 1** in Appendix A summarizes the definitions and notations.

From the global perspective, there are n users (agents), denoted by the set $V = \{1, 2, \ldots, n\}$. Each user $i \in V$ has an unknown feature vector $\boldsymbol{\theta}_i \in \mathbb{R}^d$. For simplicity, we assume $\|\boldsymbol{\theta}_i\| \leq 1, \forall i \in V$. Besides, the similarity of user characteristics can be encoded as the closeness (distance) of $\boldsymbol{\theta}_i$, and the set of users can be partitioned into a small number of clusters $V_1, \ldots, V_m$ with $m \ll n$. Users belonging to the same cluster behave similarly and therefore share the same feature vector. The structure of the m clusters and the common user characteristics in each cluster are unknown.

At each time step $t = 1, \ldots, T$, each user learns and makes the decision independently, in other words, agents act in parallel. A set containing L actions $D_{i,t}$ (contexts / items) arrives for each user $i \in V$ to select, $D_{i,t} \subset D \subset \mathbb{R}^d$. Each agent i selects an action $\boldsymbol{x}_{i,t}$ and receives a payoff $\boldsymbol{a}_{i,t}$. We define the payoff of action $\boldsymbol{x}$ for user i to be

$$a_{i,t}(\boldsymbol{x}) = \boldsymbol{\theta}_i^T\boldsymbol{x} + \epsilon_{i,t}, \tag{1}$$

where $\boldsymbol{\theta}_i$ is a fixed but unknown feature vector of agent i and $\epsilon_{i,t}$ is a conditionally zero-mean, sub-Gaussian noise with variance parameter.[1] $\sigma^2(\boldsymbol{x}) \leq \sigma^2, \forall \boldsymbol{x}$. Therefore, conditioned on the past, the quantity $\boldsymbol{\theta}_i^T\boldsymbol{x}$ is indeed the expected payoff observed by user i for context vector $\boldsymbol{x}$. $V(i)$ is the cluster to which user i belongs. We have $V(i) = V(j)$ if and only if $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$. Below, we present two assumptions (Gentile et al., 2014; Li et al., 2019).

Assumption 1 (Well-separatedness)**.** For any two different feature vectors $\boldsymbol{\theta}_j \neq \boldsymbol{\theta}_{j'}$, there is a gap γ between them $\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j'}\| \geq \gamma > 0$.

Assumption 2 (Action regularity)**.** For each time step t, context vectors (action set) $D_{i,t} = \{\boldsymbol{x}_{i,t}^1, \ldots, \boldsymbol{x}_{i,t}^L\}$ are generated i.i.d. from a random vector X, $\|X\| \leq 1$, such that $\mathbb{E}[XX^T]$ is full rank with minimal eigenvalue $\lambda_{\min} > 0$. We assume that the lower bound of $\lambda_{\min}$ is known.

[1] A zero-mean random variable X is sub-Gaussian with variance parameter σ^2 if $\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{s^2\sigma^2}{2}\right)$ for all $s \in \mathbb{R}$. Any variable X with $\mathbb{E}(X) = 0$ and $|X| \leq b$ is sub-Gaussian with variance parameter upper bounded by b.

The instantaneous regret of each user i is given by

$$\mathcal{R}_{i,t} = \boldsymbol{\theta}_i^T\boldsymbol{x}_{i,t}^* - \mathbb{E}(a_{i,t}) = \boldsymbol{\theta}_i^T\boldsymbol{x}_{i,t}^* - \boldsymbol{\theta}_i^T\boldsymbol{x}_{i,t}, \tag{2}$$

where $\boldsymbol{x}_{i,t}^* = \arg\max_{\boldsymbol{x}\in D_{i,t}} \boldsymbol{\theta}_i^T\boldsymbol{x}$ is the optimal action for user i at t. The cumulative regret of the entire system is

$$R(T) = \sum_{t=1}^{T}\sum_{i=1}^{n} \mathcal{R}_{i,t}. \tag{3}$$

The goal is to minimize the total cumulative regret of the whole system with high probability.

3. Online Clustering of Bandits via Hedonic Game

In this section, we introduce our algorithm, Online Clustering of Bandits via Hedonic Game (CLUB-HG) for the online clustering of bandits. In contrast to the previous works, we model the clustering problem as a hedonic game in which each user is a player, and a clustering form is one coalitional structure. Each player decides whether to leave its current coalition and join a new one. This approach results in self-organized clustering as the outcome of the decisions of independent players. **Algorithm 1** presents the pseudocode.

CLUB-HG maintains a profile $(\boldsymbol{M}_{i,t}, \boldsymbol{b}_{i,t})$ for each user i, where $\boldsymbol{M}_{i,t}$ is the Gramian matrix and $\boldsymbol{b}_{i,t}$ is the moment vector of regressand by regressors (Li et al., 2019). $\boldsymbol{\omega}_{i,t}$ denotes the estimation of $\boldsymbol{\theta}_i$ at time step t. The user profiles are initialized at the beginning of the algorithm and the clustering is initialized to be one single cluster containing all users (line 2), $m_1 = 1$ and $\hat{V}_{1,1} = V$. We denote by $\hat{V}_{1,t}, \ldots, \hat{V}_{m_t,t}$ the clusters obtained from the hedonic clustering game (Algorithm 2) at time step t. The clusters $\hat{V}_{1,t}, \ldots, \hat{V}_{m_t,t}$ (current clusters) are indeed meant to estimate the underlying true partition $V_1, \ldots, V_m$ (underlying true clusters). For each current cluster j, the algorithm maintains a profile $(\overline{\boldsymbol{M}}_{j,t}, \bar{\boldsymbol{b}}_{j,t}, \hat{V}_{j,t})$ as well, where $\hat{V}_{j,t}$ contains all user indexes of current cluster j and

$$\overline{\boldsymbol{M}}_{j,t-1} = \boldsymbol{I} + \sum_{i\in\hat{V}_{j,t}} (\boldsymbol{M}_{i,t-1} - \boldsymbol{I})$$

$$\bar{\boldsymbol{b}}_{j,t-1} = \sum_{i\in\hat{V}_{j,t}} \boldsymbol{b}_{i,t-1},$$

contain the aggregate information of cluster j.

We assume that all users in a same cluster select the action based on their shared knowledge on their current common cluster. In this case, all users in the same current cluster make the same decision, which is selected by one-time calculation. Hence, the computational cost can be reduced (compared to the case that each user makes a different decision based on their own calculations). More specifically, at

Algorithm 1 CLUB-HG

1: **Input**: Exploration parameter: $\alpha_j(t)$, hedonic clustering accuracy parameter β_t.
2: **Initialization**
$\boldsymbol{b}_{i,0} = \mathbf{0} \in \mathbb{R}^d$ and $\boldsymbol{M}_{i,0} = \boldsymbol{I} \in \mathbb{R}^{d\times d}$, $\forall i \in V$;
Clusters $\hat{V}_{1,1} = V$, number of clusters $m_1 = 1$;
3: **for** $t = 1, 2, \cdots, T$ **do**
4: **for** $j = 1, 2, \cdots, m_t$ **do**
5: Set

$$\overline{\boldsymbol{M}}_{j,t-1} = \boldsymbol{I} + \sum_{i\in\hat{V}_{j,t}} (\boldsymbol{M}_{i,t-1} - \boldsymbol{I})$$
$$\bar{\boldsymbol{b}}_{j,t-1} = \sum_{i\in\hat{V}_{j,t}} \boldsymbol{b}_{i,t-1}$$
$$\bar{\boldsymbol{\omega}}_{j,t-1} = \overline{\boldsymbol{M}}^{-1}_{j,t-1}\bar{\boldsymbol{b}}_{j,t-1}$$

6: Select action $\bar{\boldsymbol{x}}_{j,t} = \arg\max_{\boldsymbol{x}\in D_{j,t}}(\bar{\boldsymbol{\omega}}^T_{j,t-1}\boldsymbol{x} + \alpha_j(t)\|\boldsymbol{x}\|_{\overline{\boldsymbol{M}}^{-1}_{j,t-1}})$
7: **for** $i \in \hat{V}_{j,t}$ **do**
8: Take action $\boldsymbol{x}_{i,t} = \bar{\boldsymbol{x}}_{j,t}$
9: Receive the payoff $a_{i,t}$
10: Update weights:

$$\boldsymbol{M}_{i,t} = \boldsymbol{M}_{i,t-1} + \boldsymbol{x}_{i,t}\boldsymbol{x}^T_{i,t},$$
$$\boldsymbol{b}_{i,t} = \boldsymbol{b}_{i,t-1} + a_{i,t}\boldsymbol{x}_{i,t},$$
$$\boldsymbol{\omega}_{i,t} = \boldsymbol{M}^{-1}_{i,t}\boldsymbol{b}_{i,t}$$

11: **end for**
12: **end for**
13: Run **Hedonic Clustering Game** in **Algorithm 2**, obtain the updated clustering $\hat{V}_{1,t}, \ldots, \hat{V}_{m_t,t}$
14: **end for**

each iteration, each user obtains its current cluster information and selects an action accordingly. This only needs one calculation per cluster. In the algorithm, it shows that each current cluster $\hat{V}_{j,t}$, $j = 1, \ldots, m_t$ aggregates the users' information, obtains the estimated feature vector $\bar{\boldsymbol{\omega}}_{j,t-1}$ (Line 5), and selects an action for all users in cluster j accordingly from the action set $D_{j,t}$ (Line 6). After receiving the payoff $a_{i,t}$, each user updates its profile from $(\boldsymbol{M}_{i,t-1}, \boldsymbol{b}_{i,t-1})$ to $(\boldsymbol{M}_{i,t}, \boldsymbol{b}_{i,t})$ and obtains the new estimation of feature vector $\boldsymbol{\omega}_{i,t}$. Subsequently, the clustering structure will be also updated according to users' new estimation. Here we formulate a hedonic clustering game model to update the clustering structure. Alternatively, a fully distributed version of Algorithm 1 is presented in Appendix B.

3.1. Hedonic Clustering Game

Hedonic game provides a natural framework to study clustering (Feldman et al., 2015). In the hedonic game, a set of players express their preference over coalitions they want to join and the outcome of the hedonic game is disjoint coalitions of the players. Considering the players and coalitions as users and clusters, respectively, we formulate the clustering problem as a hedonic game. The formal presentation of such a game is a tuple $\mathcal{H} = \prec V, \{\hat{V}(i)\}_{i\in V}, \{C_i\}_{i\in V} \succ$, where $V = \{1, \ldots, n\}$ denotes the set of players, $\hat{V}(i)$ is the player i's coalition, i.e., it is the set of players that user i is estimated to be in the same current cluster with, and $C_i \in \mathbb{R}$ is the expected cost of player i. Typically, when forming coalitions, the objective of each player i is to minimize its cost C_i. Here we use a similar cost function as (Feldman et al., 2015),

$$C_i = \sum_{k\in\hat{V}(i)} \beta_t d(k,i) + \sum_{k\notin\hat{V}(i)} (\Omega - d(i,k)), \quad (4)$$

where $d(i,k) = \|\boldsymbol{\omega}_i - \boldsymbol{\omega}_k\| \in [0, \Omega]$ measures the difference of estimated feature vectors $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_k$, and β_t is a parameter that controls the inter-cluster distance of clustering structure. It is worth mentioning that different from the setting in (Feldman et al., 2015), the exact distance between users is unknown in the formulated game $\mathcal{H}$. Hence we calculate $d(i,k)$ based on the estimated feature vectors obtained from the bandit learning process. Specifically, $d(i,k) \approx 0$ means that i and k are very similar, and vice versa. Besides, we introduce the parameter β_t to control the precision of clustering. In general, users with similar estimated feature vectors will probably join the same coalitions (clusters). **Algorithm 2** shows the pseudocode for hedonic clustering game.

Algorithm 2 Hedonic Clustering Game

Input: Estimated feature vectors $\boldsymbol{\omega}_{i,t}$, $i \in V$
repeat
Select a proposer i uniformly at random
Take the best response strategy

$$\hat{V}(i) = \arg\min_{\hat{V}(i)} [\textstyle\sum_{k\in\hat{V}(i)} \beta_t d(k,i) + \sum_{k\notin\hat{V}(i)} (\Omega - d(i,k))],$$

where $d(i,k) = \|\boldsymbol{\omega}_{i,t} - \boldsymbol{\omega}_{k,t}\|$.
until Nash equilibrium
return The converged clusters $\hat{V}_{1,t}, \ldots, \hat{V}_{m_t,t}$

4. Performance Guarantees

In this section, we theoretically analyze our proposed CLUB-HG framework and provide performance guarantees.

4.1. Hedonic Clustering Game

First, we analyze the performance guarantees in the formulated hedonic clustering game, based on the results in (Feldman et al., 2015). Detailed proofs are available in Appendix C.

Theorem 1. *There always exists a Nash equilibrium for the formulated hedonic clustering game $\mathcal{H}$. Moreover, the best-response dynamics of this game always converge to a Nash equilibrium.*

Proof sketch. The hedonic clustering game is a finite potential game with potential function $\Phi(\mathcal{C}) = \sum_{i\in V}(\sum_{j\in\hat{V}(i)}\beta_t d(j,i) + \sum_{j\notin\hat{V}(i)}(\Omega - d(i,j)))$ given a clustering configuration $\mathcal{C}$. Consequently, best-response dynamics converge to a Nash equilibrium. □

Theorem 2. *The problem of computing a Nash equilibrium of the formulated hedonic clustering game $\mathcal{H}$ is PLS-Complete.*

Proof sketch. The proof follows by a reduction from the problem called POS-NAE-3SAT (Not-All-Equal 3-Satisfiability). An instance of NAE-3SAT consists of clauses with at most three literals NAE(w_1, w_2, w_3), where each w_i is a literal or a constant (0 or 1). And in POS-NAE-3SAT, the literals are not negative. Each clause is assigned with a positive weight. Such a clause is satisfied if it does not assign the same value to all the literals. The objective of such problem is to find an assignment that is a local maximum, i.e., its weight cannot be increased by flipping the value of a single literal. POS-NAE-3SAT is proved to be a PLS-complete problem (Schäffer, 1991). For a detailed proof please see Theorem 5.4 of Feldman et al. (2015). □

Lemma 1 (Property of Nash equilibrium)**.** *Any two nodes i and k share the same cluster at Nash equilibrium if and only if*

$$d(i,k) \leq \frac{\Omega}{1+\beta_t} \tag{5}$$

4.2. Regret Analysis

Our analysis relies on the high probability analysis in (Abbasi-Yadkori et al., 2011).

Theorem 3. *Let the CLUB-HG algorithm described in Algorithm 1 runs with a set of agents $V = \{1, 2, \ldots, n\}$, which can be partitioned into m clusters $V_1, \ldots, V_m$, where $\forall j \in \{1, \ldots, m\}$, users within cluster V_j host the same vector $\boldsymbol{\theta}_j$ with $\|\boldsymbol{\theta}_j\| \leq 1$. The parameter $\boldsymbol{\theta}_i$, $\forall i \in V$ and the underlying clustering information are unknown and need to be inferred. Let Assumption 1 and 2 hold. Let $\alpha_j(t) = \sigma\sqrt{2\log\left(\frac{\det\left(\overline{\boldsymbol{M}}_{j,t-1}\right)^{1/2}}{\delta}\right)} + 1$ and $\beta_t = \sqrt{\frac{t}{\log(t+1)}}$. Then with probability at least $1-\delta$, the cumulative regret satisfies*

$$R(T) \leq \tilde{O}(\frac{n}{\lambda_{\min}\gamma^2} + \sqrt{dmnT}(\sigma\sqrt{d}+1)), \tag{6}$$

as T grows large (more precisely, for $T > \max\{\frac{128}{3\delta\lambda_{\min}}\log\left(\frac{128}{3\delta\lambda_{\min}}\right) + 3, \frac{512}{\gamma^2\lambda_{\min}}\left(\frac{d\lambda_{\min}\gamma^2}{512} + 1 + 2\log\left(\frac{2}{\delta}\right) + \sigma^2 d\log\left(\frac{512\sigma^2}{\lambda_{\min}\gamma^2}\right)\right), \frac{4\Omega^2\log(T+1)}{\gamma^2}\}$). Because of $\tilde{\mathcal{O}}$ notation, the factors $\log(1/\delta)$, $\log(d)$, and $\log(T)$ do not appear.

Proof sketch. We sketch below the proof of Theorem 3. Complete proofs of Theorem 3 and associated Lemmas are available in Appendix D.

Define modified confidence ellipsoids: First we need a version of the confidence ellipsoid theorem given in (Abbasi-Yadkori et al., 2011).

Lemma 2. *Fix user i, it holds with probability $1-\delta$ that*

$$\|\boldsymbol{\omega}_{i,t-1} - \boldsymbol{\theta}_i\|_{\boldsymbol{M}_{i,t-1}} \leq \sigma\sqrt{2\log\left(\frac{\det(\boldsymbol{M}_{i,t-1})^{1/2}}{\delta}\right)} + 1. \tag{7}$$

In the rest of the proof, we assume (7) holds.

Calculate the closeness of the estimated cluster and the underlying cluster:
According to Lemma 1, if user i and k share the same estimated cluster at time step t, then it holds $\|\boldsymbol{\omega}_{i,t} - \boldsymbol{\omega}_{k,t}\| \leq \frac{\Omega}{1+\beta_t}$. To guarantee that for any user i and k which share the same underlying cluster can be classified into the same current cluster under hedonic game, the following condition must be satisfied

$$\begin{aligned}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_k\| &\leq \|\boldsymbol{\theta}_i - \boldsymbol{\omega}_{i,t}\| + \|\boldsymbol{\omega}_{i,t} - \boldsymbol{\omega}_{k,t}\| + \|\boldsymbol{\theta}_k - \boldsymbol{\omega}_{k,t}\| \\ &\leq \gamma.\end{aligned} \tag{8}$$

The guarantee for the correctness of the estimated clustering structure is formulated in following lemma.

Lemma 3. *Fix any user i, when $t > \max\{\frac{128}{3\delta\lambda_{\min}}\log\left(\frac{128}{3\delta\lambda_{\min}}\right) + 3, \frac{512}{\gamma^2\lambda_{\min}}\left(\frac{d\lambda_{\min}\gamma^2}{512} + 1 + 2\log\left(\frac{2}{\delta}\right) + \sigma^2 d\log\left(\frac{512\sigma^2}{\lambda_{\min}\gamma^2}\right)\right), \frac{4\Omega^2\log(T+1)}{\gamma^2}\}$, it holds with probability at least $1-\delta$ that the algorithm infers the underlying clusters correctly.*

Decompose the instantaneous regret:

$$\begin{aligned}\mathcal{R}_{i,t} &= \boldsymbol{\theta}_i^T\boldsymbol{x}_{i,t}^* - \boldsymbol{\theta}_i^T\boldsymbol{x}_{i,t}, \\ &= \boldsymbol{\theta}_i^T\boldsymbol{x}_{i,t}^* - \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T\boldsymbol{x}_{i,t}^* + \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T\boldsymbol{x}_{i,t}^* \\ &\quad - \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T\boldsymbol{x}_{i,t} + \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T\boldsymbol{x}_{i,t} - \boldsymbol{\theta}_i^T\boldsymbol{x}_{i,t}.\end{aligned} \tag{9}$$

Because $\boldsymbol{x}_{i,t} = \arg\max_{\boldsymbol{x}\in X_t}(\bar{\boldsymbol{\omega}}^T_{\hat{V}_t(i),t-1}\boldsymbol{x} + \alpha_{\hat{V}_t(i)}\|\boldsymbol{x}\|_{\overline{\boldsymbol{M}}^{-1}_{\hat{V}_t(i),t-1}})$,

$$\bar{\boldsymbol{\omega}}^T_{\hat{V}_t(i),t-1}\boldsymbol{x}_{i,t}+\alpha\|\boldsymbol{x}_{i,t}\|_{\overline{\boldsymbol{M}}^{-1}_{\hat{V}_t(i),t-1}} \geq \bar{\boldsymbol{\omega}}^T_{\hat{V}_t(i),t-1}\boldsymbol{x}^*_{i,t} + \alpha\|\boldsymbol{x}^*_{i,t}\|_{\overline{\boldsymbol{M}}^{-1}_{\hat{V}_t(i),t-1}}.$$

Therefore, the instantaneous regret satisfies

$$\mathcal{R}_{i,t} \leq \alpha_{\hat{V}_t(i)}(t)\|\boldsymbol{x}^*_{i,t}\|_{\overline{\boldsymbol{M}}^{-1}_{\hat{V}_t(i),t-1}} + 2\alpha_{\hat{V}_t(i)}(t)\|\boldsymbol{x}_{i,t}\|_{\overline{\boldsymbol{M}}^{-1}_{\hat{V}_t(i),t-1}}.$$

Regret analysis after discovering true underlying clusters: Based on Lemma 3, when $t > \max\{\frac{128}{3\delta\lambda_{\min}}\log\left(\frac{128}{3\delta\lambda_{\min}}\right) + 3, \frac{512}{\gamma^2\lambda_{\min}}\left(\frac{d\lambda_{\min}\gamma^2}{512} + 1 + 2\log\left(\frac{2}{\delta}\right) + \sigma^2 d\log\left(\frac{512\sigma^2}{\lambda_{\min}\gamma^2}\right)\right), \frac{4\Omega^2\log(T+1)}{\gamma^2}\}$, the good event that all the users are partitioned into the true underlying clusters is guaranteed. Therefore, select $\tau = \max\{\frac{128}{3\delta\lambda_{\min}}\log\left(\frac{128}{3\delta\lambda_{\min}}\right) + 3, \frac{512}{\gamma^2\lambda_{\min}}\left(\frac{d\lambda_{\min}\gamma^2}{512} + 1 + 2\log\left(\frac{2}{\delta}\right) + \sigma^2 d\log\left(\frac{512\sigma^2}{\lambda_{\min}\gamma^2}\right)\right), \frac{4\Omega^2\log(T+1)}{\gamma^2}\}$, we can calculate the regret per cluster since each user follows its cluster decision

$$R(T) = An\tau + \sum_{t=\tau+1}^{T}\sum_{j=1}^{m}\sum_{i=1}^{|V_j|}\mathcal{R}_{i,t}, \tag{10}$$

where A is the bound of the instantaneous reward $a_{i,t}$, $i \in V$. The cumulative regret consists of two parts: The first part considers the worst case when the estimated clusters are not consistent with the true underlying clusters while the second part calculates the accumulated regret after the true underlying clusters are discovered. Using Cauchy-Schwarz inequality and Lemma 2, we can conclude that the cumulative regret satisfies

$$\begin{aligned} R(T) &= An\tau + \sum_{t=\tau+1}^{T}\sum_{j=1}^{m}\sum_{i=1}^{|V_j|}\mathcal{R}_{i,t} \\ &\leq An\tau + 3\sum_{t=\tau+1}^{T}\sum_{j=1}^{m}\sum_{i=1}^{|V_j|}\alpha_{V(i)}(t)\|\boldsymbol{x}\|_{\overline{\boldsymbol{M}}^{-1}_{V(i),t-1}} \\ &\leq An\tau + 3\sqrt{mnT}\sqrt{2d\log\left(1+\frac{nT}{d}\right)}\times \\ &\left(\sigma\sqrt{d\log\left(1+\frac{nT}{d}\right)+2\log\left(\frac{1}{\delta}\right)}+1\right). \end{aligned} \tag{11}$$

□

4.3. Discussion

Analysis in Theorem 3 is carried out in the case when the number of rounds T is large enough. However, if the number of rounds is limited, then we show that the upper bound of regret is still sublinear with respect to T in following theorem.

Theorem 4. *Let the CLUB-HG algorithm described in Algorithm 1 run with a set of agents $V = \{1, 2, \ldots, n\}$, which can be partitioned into m clusters $V_1, \ldots, V_m$, where $\forall j \in \{1, \ldots, m\}$, users within cluster V_j host the same vector $\boldsymbol{\theta}_j$ with $\|\boldsymbol{\theta}_j\| \leq 1$. Assume that Assumption 1 and 2 hold. Let $\alpha_j(t) = \sigma\sqrt{2\log\left(\frac{\det\left(\overline{\boldsymbol{M}}_{j,t-1}\right)^{1/2}}{\delta}\right)} + 1$ and $\beta_t = \sqrt{\frac{t}{\log(t+1)}}$. When the sampling steps is not large enough such that $T \leq \max\{\frac{128}{3\delta\lambda_{\min}}\log\left(\frac{128}{3\delta\lambda_{\min}}\right) + 3, \frac{512}{\gamma^2\lambda_{\min}}\left(\frac{d\lambda_{\min}\gamma^2}{512} + 1 + 2\log\left(\frac{2}{\delta}\right) + \sigma^2 d\log\left(\frac{512\sigma^2}{\lambda_{\min}\gamma^2}\right)\right), \frac{4\Omega^2\log(T+1)}{\gamma^2}\}$, with probability at least $1-\delta$ the cumulative regret satisfies*

$$R(T) \leq \tilde{\mathcal{O}}(n\sqrt{dT}(\sigma\sqrt{d}+1)). \tag{12}$$

Remark 1. The cumulative regret bound in Theorem 3 reveals that the regret bound has two main terms: The first depends on n and is inversely proportional to $\lambda_{\min}\gamma^2$. It accounts for the hardness of inferring the true underlying clusters through hedonic game. The second term corresponds to the regret bound obtained after discovering the true clustering structure. Besides, it shows that the regret is sublinear in T. Moreover, the second term evaluates the theoretical regret bound of the LinUCB algorithm that has the underlying clustering information as a priori. Thus, Theorem 3 shows the performance difference between the CLUB-HG and the LinUCB algorithms given the clustering information.

Remark 2. Theorem 4 shows the regret bound when the number of learning/decision steps is limited for the users and the convergence to the true underlying clusters may not be reached. In this scenario, n is the dominating factor in the regret bound. The regret bound in Theorem 4 has an extra $\sqrt{n/m}$ factor compared to the regret bound in Theorem 3 which highlights the advantage of finding the true underlying clustering structure for $m \ll n$.

5. Experiments

We evaluate the performance of our algorithm using synthetic and real-world datasets. Besides, we compare the results to some state-of-the-art bandit and clustering of bandits algorithms.

5.1. Datasets

Synthetic Dataset. We evaluate the performance of our algorithm using synthetic and real-world datasets. Besides, we compare the results to some state-of-the-art bandit and clustering of bandits algorithms. Our produced synthetic dataset has $n = 20$ users that belong to four latent clusters, $m = 4$. The clusters contain an equal number of users. Besides, we set $L = 10$, $d = 5$ and $T = 1000$. The generation of user feature vectors $\boldsymbol{\theta}$ and the context vectors $\boldsymbol{x}$ follows the same procedure as the synthetic experiments in (Li et al., 2019): We generate the values of first $d - 1$ dimensions using a standard Gaussian distribution, and we use the value of the last dimension for normalization. We generate the values of payoff (reward) by perturbing the innerproduct $\boldsymbol{\theta}^T \boldsymbol{x}$ using an additive white noise term ϵ, which follows a uniform distribution over the interval $[-\sigma, \sigma]$ and here we set $\sigma = 0.1$.

Netflix Dataset. Netflix Dataset [2] (Bennett et al., 2007) contains more than 16 million ratings for 1350 movies by 143458 users. We extract 10^3 movies with the most ratings and $n = 200$ users that give a rating most frequently from the dataset. We compute feature vectors (dimension $d = 10$) for all users based on the obtained rating matrix using singular-value decomposition (SVD), which is similar to the method in (Li et al., 2019). Subsequently, we apply the K-Means algorithm to divide $n = 200$ users into $m = 10$ clusters. At each time step, we sample $L = 10$ items from the extracted 10^3 movies as the action (context) set. The payoff setting is the same as in the synthetic dataset.

MovieLens Dataset. MovieLens dataset (Harper & Konstan, 2015) describes 5-star rating and free-text tagging activity from MovieLens, a movie recommendation service.[3] It contains 25 million ratings and 1093360 tag applications across 62423 movies. There are 20 genres, and each movie can have at most six genre tags. We use the information of movie genres to generate action (context) sets. The dimension of the context vector is $d = 20$, each representing a genre. If a movie is associated with a particular genre, the respective dimensional feature is set as $\boldsymbol{x}(l) = \frac{1}{\sqrt{S}}$, with S as the total number of genres of the movie (Bilaj et al., 2023). Similarly, we extract 10^3 movies with the most ratings and $n = 200$ users who rate most times from the dataset. The feature vector of each user is the normalized summation of context vectors from his/her five "favorite" (with the highest ranking) movies. Then we use K-Means to divide $n = 200$ users into $m = 10$ clusters. We set $L = 10$, and we sample the action set of each agent from the extracted 10^3 movies. The payoff setting remains unchanged.

[2]Netflix Movie Rating Dataset from Netflix's 'Netflix Prize" competition on https://www.kaggle.com/datasets/rishitjavia/netflix-movie-rating-dataset?resource=download

[3]MovieLens 25M Movie Ratings Dataset on https://grouplens.org/datasets/movielens/

5.2. Algorithms

We compare our proposed CLUB-HG algorithm to a number of state-of-the-art algorithms for linear bandit and clustering of bandits. All regret plots are based on the average results of 20 independent runs.

- CLUB (Gentile et al., 2014): The algorithm starts with a complete graph and progressively erases edges based on the evolution of estimated feature vectors.
- SCLUB (Li et al., 2019): The algorithm uses sets to represent clusters and allows both split and merge operations on sets during the learning process.
- LinUCB-IND (Abbasi-Yadkori et al., 2011): The algorithm performs n LinUCB policies for each user $i \in V$ independently.
- GOB (Cesa-Bianchi et al., 2013): The algorithm receives a Laplacian matrix that encodes the true underlying graph G (clustering information).
- LinUCB-CLU (Abbasi-Yadkori et al., 2011): The algorithm receives the true clustering structure a priori as input and performs LinUCB for the clusters.

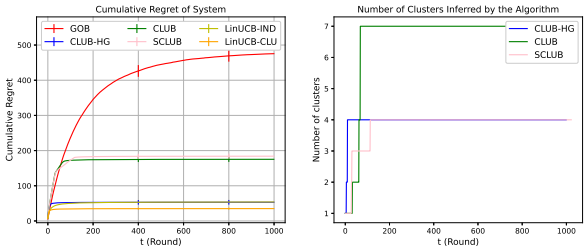
For consistency, we implement each algorithm in the same distributed manner as our proposed algorithm (The algorithm selects one same action for all users in the same current cluster). Furthermore, the upper confidence bound of linear bandit is set as $\alpha_j(t) = (\sigma\sqrt{2\log\left(\frac{\det\left(\boldsymbol{M}_{j,t-1}\right)^{1/2}}{\delta}\right)} + 1)$ for all algorithms. We implement the GOB algorithm only in the experiment of the synthetic dataset due to its high computational complexity.

5.3. Results

Figure 1 - 3 summarize the results. For all datasets and algorithms, we illustrate the cumulative regret of the system and the number of inferred clusters for performance evaluation. In the plot of cumulative regret, error bars indicate the standard deviations divided by $\sqrt{20}$.

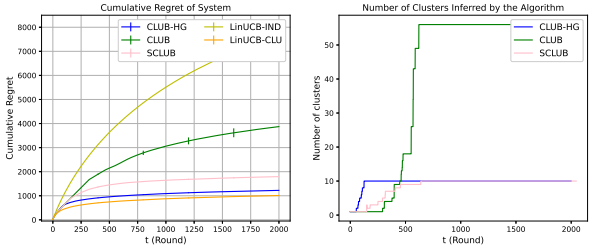
Based on the experimental results, our analysis and conclusions are as follows.

- Unsurprisingly, as the true underlying clustering structure is known for LinUCB-CLU a priori, LinUCB-CLU outperforms all the other algorithms on all datasets (synthetic dataset, Netflix dataset, and MovieLens dataset).
- On all datasets, our proposed CLUB-HG algorithm outperforms other algorithms for clustering of bandits (CLUB, SCLUB): CLUB-HG has the lowest cumu-



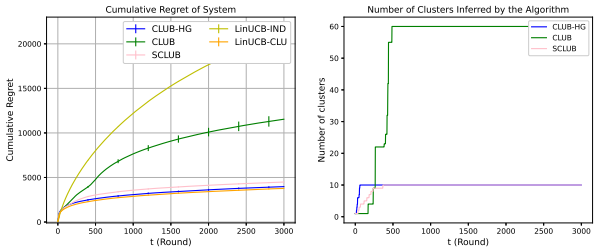
(a) Empirical cumulative regret. (b) Number of clusters

Figure 1. Results on synthetic dataset.



(a) Empirical cumulative regret. (b) Number of clusters.

Figure 2. Results on Netflix dataset.



(a) Empirical cumulative regret. (b) Number of clusters.

Figure 3. Results on MovieLens dataset.

lative regret and can identify the true clusters within the minimum time steps. That demonstrates the advancement of the CLUB-HG algorithm in clustering identification and action selection.

- In Figure 1, CLUB-HG algorithm has better performance than GOB. However, GOB knows the clustering structure apriori, which is unknown for our algorithm. The assumption in GOB that the system is a connected graph cannot be satisfied in the experiment since only users in the same cluster are connected. Therefore the whole system cannot be encoded as a connected graph. That leads to the degeneration of GOB's performance and explains the reason that GOB has higher regret. Moreover, our algorithm achieves the same or even better performance without the prior clustering information in this comparison.
- On all datasets, the number of clusters inferred by CLUB-HG and SCLUB converges to the correct number m of the underlying clusters, whereas it is not the case for the CLUB. In both Li et al. (2019) and our work, the convergence to the true underlying cluster is proved theoretically, whereas the CLUB algorithm does not necessarily converge. Since the cutting operations in the CLUB algorithm are irreversible, users that split out incorrectly in the early stages cannot join the correct cluster again in subsequent steps. This drawback explains why the number of clusters inferred by the CLUB algorithm is always more than others. Furthermore, it enlarges the regret and computational cost of the CLUB algorithm. In experiments on all real-world datasets, the CLUB algorithm has the highest regret among all algorithms for clustering of bandits.
- Different from the result in Figure 1, the regret of LinUCB-IND is higher than all the algorithms in online clustering of bandits (CLUB-HG, CLUB, and SCLUB) in Figure 2 and Figure 3. The reason is that the algorithms in clustering of bandits can utilize the shared information from all users in the same cluster to enhance decision-making performance. However, if the number of users, n, and the dimension of the feature vector, d, are relatively low (similar to the experiment on the synthetic dataset in Figure 1), that advantage remains negligible. However, in Netflix and MovieLens dataset, the feature vector dimension and the number of users are relatively high and large. This increases the difficulty in estimating user feature vectors in single-agent linear bandit algorithms (LinUCB-IND). Online clustering of bandits can compensate for this weakness by identifying clusters of users and sharing information. Therefore, they have lower computational costs and a better estimation of user features, thus lower regrets.
- In Figure 3, our proposed algorithm can even obtain comparable performance as LinUCB-CLU (the clustering information is known). As mentioned before, the advantage of sharing information and making a decision based on the cluster feature vector is not remarkable when n and d are relatively small, whereas, in the MovieLens dataset, the feature vector is sparse and has a much higher dimension. Due to the excellent underlying clustering identification in our proposed algorithm, CLUB-HG performs better even at the expense of some accuracy of user feature estimation and thus has comparable regret as the LinUCB-CLU.

6. Conclusion and Future Work

In this work, we propose the CLUB-HG algorithm for the online clustering of bandits using a game-theoretic approach. The proposed CLUB-HG algorithm effectively incorporates the hedonic game into the online clustering of bandits. We prove the convergence to the Nash equilibrium at each clus-

tering operation. Besides, we guarantee to achieve the true clustering structure after a certain number of iterations. Based on the superior performance of the game-theoretic clustering approach, similar users can share information to enhance the decision-making performance. Furthermore, we establish the theoretical regret bound of our proposed algorithm. The experiments on synthetic and real-world datasets demonstrate that our CLUB-HG algorithm outperforms existing approaches. Future work may include considering a more realistic scenario where the feature vectors of users are not the same but quite similar. Besides, one can investigate the mutual influence among users instead of assuming that similar users have the same preference.

Acknowledgement

This work was supported by the German Research Foundation (DFG) under Grant MA 7111/6-1 and by the Cyber Valley under Grant CyVy-RF-2021-20.

We thank Dr. Sofien Dhouib for the fruitful discussions and the valuable feedback.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Aziz, H., Gaspers, S., Gudmundsson, J., Mestre, J., and Taubig, H. Welfare maximization in fractional hedonic games. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Aziz, H., Brandl, F., Brandt, F., Harrenstein, P., Olsen, M., and Peters, D. Fractional hedonic games. *ACM Transactions on Economics and Computation (TEAC)*, 7(2):1–29, 2019.

Balliu, A., Flammini, M., Melideo, G., and Olivetti, D. Nash stability in social distance games. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Bennett, J., Lanning, S., et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, pp. 35. New York, 2007.

Bilaj, S., Dhouib, S., and Maghsudi, S. Hypothesis transfer in bandits by weighted models. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part IV*, pp. 284–299. Springer, 2023.

Bilò, V., Fanelli, A., Flammini, M., Monaco, G., and Moscardelli, L. Nash stable outcomes in fractional hedonic games: Existence, efficiency and computation. *Journal of Artificial Intelligence Research*, 62:315–371, 2018.

Bouneffouf, D., Rish, I., and Aggarwal, C. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8. IEEE, 2020.

Bu, Z., Li, H.-J., Cao, J., Wang, Z., and Gao, G. Dynamic cluster formation game for attributed graph clustering. *IEEE Transactions on Cybernetics*, 49(1):328–341, 2017.

Bulò, S. and Pelillo, M. A game-theoretic approach to hypergraph clustering. *Advances in Neural Information Processing Systems*, 22, 2009.

Caron, S., Kveton, B., Lelarge, M., and Bhagat, S. Leveraging side observations in stochastic bandits. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 142–151, 2012.

Cesa-Bianchi, N., Gentile, C., and Zappella, G. A gang of bandits. *Advances in Neural Information Processing Systems*, 26, 2013.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. 2008.

Feldman, M., Lewin-Eytan, L., and Naor, J. Hedonic clustering games. *ACM Transactions on Parallel Computing (TOPC)*, 2(1):1–48, 2015.

Gentile, C., Li, S., and Zappella, G. Online clustering of bandits. In *International Conference on Machine Learning*, pp. 757–765. PMLR, 2014.

Gentile, C., Li, S., Kar, P., Karatzoglou, A., Zappella, G., and Etrue, E. On context-dependent clustering of bandits. In *International Conference on Machine Learning*, pp. 1253–1262. PMLR, 2017.

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *Acm Transactions on Interactive Intelligent Systems (TIIS)*, 5(4):1–19, 2015.

Korda, N., Szorenyi, B., and Li, S. Distributed clustering of linear bandits in peer to peer networks. In *International Conference on Machine Learning*, pp. 1301–1309. PMLR, 2016.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, 2010.

Li, S., Karatzoglou, A., and Gentile, C. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 539–548, 2016.

Li, S., Chen, W., Li, S., and Leung, K.-S. Improved algorithm on online clustering of bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 2923—-2929, 2019.

Liu, X., Zhao, H., Yu, T., Li, S., and Lui, J. C. Federated online clustering of bandits. In *Uncertainty in Artificial Intelligence*, pp. 1221–1231. PMLR, 2022.

Mahadik, K., Wu, Q., Li, S., and Sabne, A. Fast distributed bandits for online recommendation systems. In *Proceedings of the 34th ACM international conference on supercomputing*, pp. 1–13, 2020.

McInerney, J., Lacker, B., Hansen, S., Higley, K., Bouchard, H., Gruson, A., and Mehrotra, R. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 31–39, 2018.

Nguyen, T. T. and Lauw, H. W. Dynamic clustering of contextual multi-armed bandits. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 1959–1962, 2014.

Papadimitriou, C. Algorithms, games, and the Internet. In *Proceedings of the thirty-third Annual ACM Symposium on Theory of Computing*, pp. 749–753, 2001.

Schäffer, A. A. Simple local search problems that are hard to solve. *SIAM journal on Computing*, 20(1):56–87, 1991.

Slivkins, A. et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Yang, L., Liu, B., Lin, L., Xia, F., Chen, K., and Yang, Q. Exploring clustering of bandits for online recommendation system. In *Fourteenth ACM Conference on Recommender Systems*, pp. 120–129, 2020.

Yi, X., Li, X., Yang, T., Xie, L., Chai, T., and Johansson, K. H. Distributed bandit online convex optimization with time-varying coupled inequality constraints. *IEEE Transactions on Automatic Control*, 66(10):4620–4635, 2020.

A. Table of Notations

Table 1. Notation

Problem-specific notations	
n	Number of users
m	Number of underlying clusters
V	User set
d	Dimension of feature vectors
T	Total number of rounds
$\boldsymbol{\theta}_i$	Feature vector of user i
$D_{i,t}$	Set of context vectors of user i at time t
L	Number of context vectors in set $D_{i,t}$
$\boldsymbol{x}_{i,t}$	The action of user i at time t
$a_{i,t}$	Payoff of user i at time t
$\epsilon_{i,t}$	Zero-mean sub-Gaussian noise
σ^2	Variance proxy of noise $\epsilon_{i,t}$
$V(i)$	The underlying cluster that user i belongs to
$\mathcal{R}_{i,t}$	Instantaneous regret of user i
$R(T)$	Cumulative regret of the system
m_t	The number of current clusters at time t
$\boldsymbol{M}_{i,t}$	The Gramian matrix of user i at time t
$\boldsymbol{b}_{i,t}$	The moment vector of regressand by regressors of user i at time t
$\boldsymbol{\omega}_{i,t}$	The estimation of $\boldsymbol{\theta}_i$ at time t
$\hat{V}_{j,t}$	The j-th current cluster at time t
$\overline{\boldsymbol{M}}_{j,t}$	The Gramian matrix of current cluster j at time t
$\bar{\boldsymbol{b}}_{j,t}$	The moment vector of regressand by regressors of current cluster j at time t
C_i	The cost function of user i in Hedonic clustering game
$d(i,k)$	The difference of estimated feature vectors $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_j$, $d(i,k) = \|\boldsymbol{\omega}_i - \boldsymbol{\omega}_j\|$

B. Distributed Online Clustering of Bandits via Hedonic Game

In this section, the distributed version of Algorithm 1 is presented in Algorithm 3. Particularly, the regret bound of Algorithm 3 is the same as the parallel algorithm Algorithm 1 stated in Theorem 3 and Theorem 4.

C. Proof of Theorem 1 and Lemma 1

C.1. Proof of Theorem 1

Proof. For a clustering configuration $\mathcal{C}$, consider the potential function $\Phi(\mathcal{C})$ with value

$$\Phi(\mathcal{C}) = \sum_{i \in V} \Big(\sum_{j \in \hat{V}(i)} \beta_t d(j,i) + \sum_{j \notin \hat{V}(i)} (\Omega - d(i,j)) \Big). \tag{13}$$

A best-response move performed by element i changes the value of Φ only through edges adjacent to i. The overall contribution of such edge is

$$2\Big(\sum_{j \in \hat{V}(i)} \beta_t d(j,i) + \sum_{j \notin \hat{V}(i)} (\Omega - d(i,j)) \Big), \tag{14}$$

which is two times the cost of i, hence it strictly decreases if agent i makes a best-response move. Hence, Φ decreases with every move performed by a player decreasing its cost. Since the hedonic clustering game is a finite game, best-response dynamics are guaranteed to converge to a Nash equilibrium. □

Algorithm 3 Distributed CLUB-HG

Input: Exploration parameter: $\alpha_{\hat{V}_t(i)}(t)$, $\forall i \in V$, hedonic clustering accuracy parameter β_t.
Initialization
$\boldsymbol{b}_{i,0} = \mathbf{0} \in \mathbb{R}^d$ and $\boldsymbol{M}_{i,0} = \boldsymbol{I} \in \mathbb{R}^{d\times d}$, $\forall i \in V$;
Clusters $\hat{V}_{1,1} = V$, number of clusters $m_1 = 1$, $\bar{\boldsymbol{b}}_{1,0} = \mathbf{0} \in \mathbb{R}^d$, $\overline{\boldsymbol{M}}_{1,0} = \boldsymbol{I} \in \mathbb{R}^{d\times d}$, $\bar{\boldsymbol{\omega}}_{1,0} = \overline{\boldsymbol{M}}_{1,0}^{-1}\bar{\boldsymbol{b}}_{1,0}$;
for $t = 1, 2, \cdots, T$ **do**
  for $i \in V$ **do**
    Take action $\boldsymbol{x}_{i,t} = \arg\max_{\boldsymbol{x}\in D_{i,t}}(\bar{\boldsymbol{\omega}}^T_{\hat{V}_{t-1}(i),t-1}\boldsymbol{x} + \alpha_{\hat{V}_{t-1}(i)}(t)\|\boldsymbol{x}\|_{\overline{\boldsymbol{M}}^{-1}_{\hat{V}_{t-1}(i),t-1}})$
    Receive the payoff $a_{i,t}$
    Update weights:

$$\begin{aligned}\boldsymbol{M}_{i,t} &= \boldsymbol{M}_{i,t-1} + \boldsymbol{x}_{i,t}\boldsymbol{x}_{i,t}^T,\\ \boldsymbol{b}_{i,t} &= \boldsymbol{b}_{i,t-1} + a_{i,t}\boldsymbol{x}_{i,t},\\ \boldsymbol{\omega}_{i,t} &= \boldsymbol{M}_{i,t}^{-1}\boldsymbol{b}_{i,t}\end{aligned}$$

  end for
  Run **Hedonic Clustering Game** in **Algorithm 2**, obtain the updated clustering $\hat{V}_{1,t}, \ldots, \hat{V}_{m_t,t}$
  for $j = 1, \ldots, m_t$ **do**
    Set

$$\begin{aligned}\overline{\boldsymbol{M}}_{j,t} &= \boldsymbol{I} + \sum_{i\in\hat{V}_{j,t}}(\boldsymbol{M}_{i,t} - \boldsymbol{I})\\ \bar{\boldsymbol{b}}_{j,t} &= \sum_{i\in\hat{V}_{j,t}}\boldsymbol{b}_{i,t}\\ \bar{\boldsymbol{\omega}}_{j,t} &= \overline{\boldsymbol{M}}_{j,t}^{-1}\bar{\boldsymbol{b}}_{j,t}\end{aligned}$$

  end for
end for

C.2. Proof of Lemma 1

Proof. At the Nash equilibrium, if user i and k share the same cluster, user k has no incentive to exclude user i from their cluster, thus,

$$\sum_{z\in\hat{V}(k),z\neq i}\beta_t d(z,k) + \beta_t d(i,k) + \sum_{z\notin\hat{V}(k)}(\Omega - d(z,k)) \leq \sum_{z\in\hat{V}(k),z\neq i}\beta_t d(z,k) + (\Omega - d(i,k)) + \sum_{z\notin\hat{V}(k)}(\Omega - d(z,k)), \tag{15}$$

which is equivalent to $d(i,k) \leq \frac{\Omega}{1+\beta_t}$.

Similarly, if i and k share different clusters, then at Nash equilibrium, user k has no incentive to include user i into its cluster, therefore we have,

$$\sum_{z\in\hat{V}(k),z\neq i}\beta_t d(z,k) + (\Omega - d(i,k)) + \sum_{z\notin\hat{V}(k)}(\Omega - d(z,k)) < \sum_{z\in\hat{V}(k),z\neq i}\beta_t d(z,k) + \beta_t d(i,k) + \sum_{z\notin\hat{V}(k)}(\Omega - d(z,k)), \tag{16}$$

thus, $d(i,k) > \frac{\Omega}{1+\beta_t}$. The same conclusion can be obtained by assuming that if i and k share different clusters at Nash equilibrium, then user k has no incentive to include the coalition containing i into its cluster. That completes the proof. □

D. Proof of Theorem 3

First, we introduce some lemmas and theorems (Abbasi-Yadkori et al., 2011) that we require for the proof of Theorem 3.

Theorem 5 (Self-Normalized Bound for Vector-Valued Martingales (Abbasi-Yadkori et al., 2011)). *Let $\{F_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\epsilon_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that ϵ_t is F_t-measurable and ϵ_t is conditionally σ-sub-Gaussian for some $\sigma > 0$, i.e.*

$$\forall \lambda \in \mathbb{R} \quad \mathbb{E}[e^{\lambda \epsilon_t} | F_{t-1}] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right). \tag{17}$$

Let $\{\boldsymbol{x}_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that $\boldsymbol{x}_t$ is F_{t-1}-measurable. Assume that $\boldsymbol{M}_0$ is a $d \times d$ positive definite matrix. For any $t \geq 0$, define

$$\boldsymbol{M}_t = \boldsymbol{M}_0 + \sum_{s=1}^{t} \boldsymbol{x}_s \boldsymbol{x}_s^T \qquad \boldsymbol{S}_t = \sum_{s=1}^{t} \epsilon_s \boldsymbol{x}_s. \tag{18}$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$,

$$\|\boldsymbol{S}_t\|^2_{\boldsymbol{M}_t^{-1}} \leq 2\sigma^2 \log\left(\frac{\det(\boldsymbol{M}_t)^{1/2} \det(\boldsymbol{M}_0))^{-1/2}}{\delta}\right). \tag{19}$$

Lemma 4. *(Abbasi-Yadkori et al., 2011) Let $\{\boldsymbol{x}_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$, $\boldsymbol{M}_0$ a $d \times d$ positive definite matrix and define $\boldsymbol{M}_t = \boldsymbol{M}_0 + \sum_{s=1}^{t} \boldsymbol{x}_s \boldsymbol{x}_s^T$. Then, we have that*

$$\log\left(\frac{\det(\boldsymbol{M}_n)}{\det(\boldsymbol{M}_0)}\right) \leq \sum_{t=1}^{n} \|\boldsymbol{x}_t\|^2_{\boldsymbol{M}_{t-1}^{-1}}. \tag{20}$$

Further, if $\|\boldsymbol{x}\|_2 \leq 1$ for all t then

$$\sum_{t=1}^{n} \min\{1, \|\boldsymbol{x}_t\|^2_{\boldsymbol{M}_{t-1}^{-1}}\} \leq 2(\log\det(\boldsymbol{M}_n) - \log\det(\boldsymbol{M}_0)) \leq 2(d\log((\mathrm{Tr}(\boldsymbol{M}_0) + n)/d) - \log\det(\boldsymbol{M}_0)), \tag{21}$$

and finally, if $\lambda_{\min}(\boldsymbol{M}_0) \geq 1$ then,

$$\sum_{t=1}^{n} \|\boldsymbol{x}_t\|^2_{\boldsymbol{M}_{t-1}^{-1}} \leq 2\log\left(\frac{\det(\boldsymbol{M}_n)}{\det(\boldsymbol{M}_0)}\right) \tag{22}$$

Lemma 5 (Determinant-Trace Inequality (Abbasi-Yadkori et al., 2011)). *Suppose $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t \in \mathbb{R}^d$ and for any $1 \leq s \leq t$, $\|\boldsymbol{x}_s\|_2 \leq 1$. Let $\boldsymbol{M}_t = \boldsymbol{I} + \sum_{s=1}^{t} \boldsymbol{x}_s \boldsymbol{x}_s^T$. Then,*

$$\det(\boldsymbol{M}_t) \leq (1 + t/d)^d \tag{23}$$

D.1. Proof of Lemma 2

Proof. Let $\boldsymbol{X}_i = \boldsymbol{x}_{i,1:t}$ contain the action of agent i's from the beginning to time step t and similarly $\boldsymbol{\epsilon}_i = (\epsilon_{i,1}, \epsilon_{i,2}, \ldots, \epsilon_{i,t})^T$, therefore $\boldsymbol{X}_i \in \mathbb{R}^{t \times d}$ and $\boldsymbol{\epsilon}_i \in \mathbb{R}^t$. Using

$$\begin{aligned}
\boldsymbol{\omega}_{i,t-1} &= \boldsymbol{M}_{i,t-1}^{-1} \boldsymbol{b}_{i,t-1} \\
&= (\boldsymbol{X}_i^T \boldsymbol{X}_i + I)^{-1} \boldsymbol{X}_i^T (\boldsymbol{X}_i \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i) \\
&= (\boldsymbol{X}_i^T \boldsymbol{X}_i + I)^{-1} \boldsymbol{X}_i^T \boldsymbol{\epsilon}_i + (\boldsymbol{X}_i^T \boldsymbol{X}_i + I)^{-1} (\boldsymbol{X}_i^T \boldsymbol{X}_i + I) \boldsymbol{\theta}_i - (\boldsymbol{X}_i^T \boldsymbol{X}_i + I)^{-1} \boldsymbol{\theta}_i \\
&= (\boldsymbol{X}_i^T \boldsymbol{X}_i + I)^{-1} \boldsymbol{X}_i^T \boldsymbol{\epsilon}_i + \boldsymbol{\theta}_i - (\boldsymbol{X}_i^T \boldsymbol{X}_i + I)^{-1} \boldsymbol{\theta}_i,
\end{aligned}$$

we get

$$|\boldsymbol{\theta}_i^T \boldsymbol{x}_{i,t} - \boldsymbol{\omega}_{i,t-1}^T \boldsymbol{x}_{i,t}| = \langle \boldsymbol{\theta}_i, \boldsymbol{x}_{i,t} \rangle_{\boldsymbol{M}_{i,t-1}^{-1}} - \langle \boldsymbol{X}_i^T \epsilon_i, \boldsymbol{x}_{i,t} \rangle_{\boldsymbol{M}_{i,t-1}^{-1}}.$$

Using Cauthy-Schwarz inequality, we have

$$|\boldsymbol{\theta}_i^T \boldsymbol{x}_{i,t} - \boldsymbol{\omega}_{i,t-1}^T \boldsymbol{x}_{i,t}| \leq \|\boldsymbol{x}_{i,t}\|_{\boldsymbol{M}_{i,t-1}^{-1}} (\left\|\boldsymbol{X}_i^T \boldsymbol{\epsilon}_i\right\|_{\boldsymbol{M}_{i,t-1}^{-1}} + \|\boldsymbol{\theta}_i\|_{\boldsymbol{M}_{i,t-1}^{-1}})$$
$$\leq \|\boldsymbol{x}_{i,t}\|_{\boldsymbol{M}_{i,t-1}^{-1}} (\left\|\boldsymbol{X}_i^T \boldsymbol{\epsilon}_i\right\|_{\boldsymbol{M}_{i,t-1}^{-1}} + \|\boldsymbol{\theta}_i\|_2).$$

The second inequality is from $\|\boldsymbol{\theta}_i\|^2_{\boldsymbol{M}_{i,t-1}^{-1}} \leq \frac{1}{\lambda_{\min}(\boldsymbol{M}_{i,t-1})} \|\boldsymbol{\theta}_i\|_2^2 \leq \|\boldsymbol{\theta}_i\|_2^2$. According to Theorem 5 for any $\delta > 0$, with probability at least $1 - \delta$,

$$\forall t \geq 0, \quad \left\|\boldsymbol{X}_i^T \boldsymbol{\epsilon}_i\right\|_{\boldsymbol{M}_{i,t-1}^{-1}} \leq \sigma \sqrt{2 \log\left(\frac{\det(\boldsymbol{M}_{i,t-1})^{1/2}}{\delta}\right)}.$$

Therefore, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\forall t \geq 0, \quad \forall \boldsymbol{x} \in \mathbb{R}^d \quad |\boldsymbol{\theta}_i^T \boldsymbol{x} - \boldsymbol{\omega}_{i,t-1}^T \boldsymbol{x}| \leq \|\boldsymbol{x}\|_{\boldsymbol{M}_{i,t-1}^{-1}} (\sigma \sqrt{2 \log\left(\frac{\det(\boldsymbol{M}_{i,t-1})^{1/2}}{\delta}\right)} + \|\boldsymbol{\theta}_i\|_2). \tag{24}$$

Plugging in $\boldsymbol{x} = \boldsymbol{M}_{i,t-1}(\boldsymbol{\omega}_{i,t-1} - \boldsymbol{\theta}_i)$, and $\|\boldsymbol{\theta}_i\| \leq 1$, we have (Abbasi-Yadkori et al., 2011)

$$\|\boldsymbol{\omega}_{i,t-1} - \boldsymbol{\theta}_i\|^2_{\boldsymbol{M}_{i,t-1}} \leq \|\boldsymbol{M}_{i,t-1}(\boldsymbol{\omega}_{i,t-1} - \boldsymbol{\theta}_i)\|_{\boldsymbol{M}_{i,t-1}^{-1}} (\sigma \sqrt{2 \log\left(\frac{\det(\boldsymbol{M}_{i,t-1})^{1/2}}{\delta}\right)} + 1), \tag{25}$$

using $\|\boldsymbol{M}_{i,t-1}(\boldsymbol{\omega}_{i,t-1} - \boldsymbol{\theta}_i)\|_{\boldsymbol{M}_{i,t-1}^{-1}} = \|\boldsymbol{\omega}_{i,t-1} - \boldsymbol{\theta}_i\|_{\boldsymbol{M}_{i,t-1}}$, we have

$$\|\boldsymbol{\omega}_{i,t-1} - \boldsymbol{\theta}_i\|_{\boldsymbol{M}_{i,t-1}} \leq \sigma \sqrt{2 \log\left(\frac{\det(\boldsymbol{M}_{i,t-1})^{1/2}}{\delta}\right)} + 1 \tag{26}$$

□

D.2. Proof of Lemma 3

Before going to the details to prove Lemma 3, we first present one necessary Lemma as follows.

Lemma 6. *Fix any user i, let $A(\delta) = \frac{128}{3\delta\lambda_{\min}} \log\left(\frac{128}{3\delta\lambda_{\min}}\right) + 3$. It holds with probability at least $1 - \delta$ that*

$$\|\boldsymbol{\omega}_{i,t} - \boldsymbol{\theta}_{i,t}\| \leq \frac{\sigma \sqrt{d \log(1 + t/d) + 2 \log\left(\frac{2}{\delta}\right)} + 1}{\sqrt{t\lambda_{\min}/8}} \leq \frac{\gamma}{4}, \tag{27}$$

when $t \geq \max\{A(\delta), B(\delta)\}$, where $B(\delta) = \frac{512}{\gamma^2 \lambda_{\min}} \left(\frac{d\lambda_{\min}\gamma^2}{512} + 1 + 2\log\left(\frac{2}{\delta}\right) + \sigma^2 d \log\left(\frac{512\sigma^2}{\lambda_{\min}\gamma^2}\right)\right)$.

Proof. By (Gentile et al., 2014; Li et al., 2019; Korda et al., 2016), with probability at least $1 - \delta$,

$$\lambda_{\min}(\boldsymbol{S}_{i,t}) \geq \lambda_{\min} t - 8 \log\left(\frac{t+3}{\delta}\right) - 2\sqrt{t \log\left(\frac{t+3}{\delta}\right)}, \tag{28}$$

where $\boldsymbol{S}_{i,t} = \sum_{s \leq t} \boldsymbol{x}_{i,s} \boldsymbol{x}_{i,s}^T$. Using $\log(x) = \log\left(\frac{x}{M}\right) + \log(M)$ and $\log(x) \leq x - 1$ when $x > 0$, by mathematical calculation, we can obtain that when $t \geq \frac{128}{3\delta\lambda_{\min}} \log\left(\frac{128}{3\delta\lambda_{\min}}\right) + 3 = A(\delta)$, it satisfies $\lambda_{\min}(\boldsymbol{S}_{i,t}) \geq \frac{t\lambda_{\min}}{8}$. Besides, according to (Abbasi-Yadkori et al., 2011), we have

$$\|\boldsymbol{\omega}_{i,t} - \boldsymbol{\theta}_i\|_{\boldsymbol{M}_{i,t-1}} \leq \sigma \sqrt{d \log(1 + t/d) + 2 \log\left(\frac{2}{\delta}\right)} + 1. \tag{29}$$

Similarly, with mathematical calculation, we can obtain when $t \geq \frac{512}{\gamma^2 \lambda_{\min}} \left(\frac{d\lambda_{\min}\gamma^2}{512} + 1 + 2\log\left(\frac{2}{\delta}\right) + \sigma^2 d \log\left(\frac{512\sigma^2}{\lambda_{\min}\gamma^2}\right)\right) = B(\delta)$. That completes the proof. □

Now we are ready to prove Lemma 3.

Proof. According to lemma 1, if user i and k belong to the same cluster, then we have $\|\boldsymbol{\omega}_{i,t} - \boldsymbol{\omega}_{k,t}\| \leq \frac{\Omega}{1+\beta_t}$. Based on Lemma 6, when $t > \max\{A(\delta), B(\delta)\}$, the following is satisfied

$$\|\boldsymbol{\theta}_i - \boldsymbol{\omega}_{i,t}\| \leq \frac{\gamma}{4}.$$

Therefore, we have

$$\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_k\| \leq \|\boldsymbol{\theta}_i - \boldsymbol{\omega}_{i,t}\| + \|\boldsymbol{\omega}_{i,t} - \boldsymbol{\omega}_{k,t}\| + \|\boldsymbol{\theta}_k - \boldsymbol{\omega}_{k,t}\| \leq \frac{\gamma}{2} + \frac{\Omega}{1+\beta_t}.$$

When $t > \max\{A(\delta), B(\delta), \frac{4\Omega^2 \log(T+1)}{\gamma^2}\}$,

$$\frac{\Omega}{1+\beta_t} = \frac{\Omega}{1+\sqrt{\frac{t}{\log(t+1)}}} \leq \frac{\Omega}{\sqrt{\frac{t}{\log(t+1)}}} \leq \frac{\Omega}{\sqrt{\frac{4\Omega^2 \log(T+1)}{\gamma^2 \log(t+1)}}} \leq \frac{\gamma\sqrt{\log(t+1)}}{2\sqrt{\log(T+1)}}.$$

Because $t \leq T$, $\frac{\Omega}{1+\beta_t} \leq \frac{\gamma}{2}$ and $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_k\| \leq \gamma$, which implies that if user i and k are estimated to belong to the same cluster by CLUB-HG, they belongs to the same cluster in the true clustering structure.

Then, we need to prove that if i and k belong to the same true underlying cluster, then the hedonic game will cluster them into the same current estimated cluster. Select suitable δ such that $\frac{2\sigma\sqrt{d\log(1+t/d)+2\log\left(\frac{2}{\delta}\right)}+2}{\sqrt{t\lambda_{\min}/8}} \leq \frac{\Omega}{\beta_t+1}$, then it is guaranteed that for any two user i and k who belong to the same underlying cluster, their estimation satisfies

$$\|\boldsymbol{\omega}_i - \boldsymbol{\omega}_k\| \leq \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_k\| + \|\boldsymbol{\theta}_i - \boldsymbol{\omega}_i\| + \|\boldsymbol{\theta}_k - \boldsymbol{\omega}_k\| \leq \frac{\Omega}{\beta_t + 1}, \tag{30}$$

which implies that they will be clustered to the same current cluster at the Nash equilibrium under hedonic game setting. Therefore, for any user i, it is guaranteed to be clustered into the correct underlying cluster when $t > \max\{\frac{128}{3\delta\lambda_{\min}}\log\left(\frac{128}{3\delta\lambda_{\min}}\right) + 3, \frac{512}{\gamma^2\lambda_{\min}}\left(\frac{d\lambda_{\min}\gamma^2}{512} + 1 + 2\log\left(\frac{2}{\delta}\right) + \sigma^2 d\log\left(\frac{512\sigma^2}{\lambda_{\min}\gamma^2}\right)\right), \frac{4\Omega^2\log(T+1)}{\gamma^2}\}$. That completes the proof. □

D.3. Proof of Theorem 3

For the instantaneous regret, we have

$$\begin{aligned}\mathcal{R}_{i,t} &= \boldsymbol{\theta}_i^T \boldsymbol{x}_{i,t}^* - \boldsymbol{\theta}_i^T \boldsymbol{x}_{i,t}, \\ &= \boldsymbol{\theta}_i^T \boldsymbol{x}_{i,t}^* - \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T \boldsymbol{x}_{i,t}^* + \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T \boldsymbol{x}_{i,t}^* - \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T \boldsymbol{x}_{i,t} + \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T \boldsymbol{x}_{i,t} - \boldsymbol{\theta}_i^T \boldsymbol{x}_{i,t}.\end{aligned} \tag{31}$$

Because $\boldsymbol{x}_{i,t} = \arg\max_{\boldsymbol{x}\in X_t}(\bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T \boldsymbol{x} + \alpha_{\hat{V}_t(i)}(t)\|\boldsymbol{x}\|_{\overline{\boldsymbol{M}}_{\hat{V}_t(i),t-1}^{-1}})$,

$$\bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T \boldsymbol{x}_{i,t} + \alpha_{\hat{V}_t(i)}(t)\|\boldsymbol{x}_{i,t}\|_{\overline{\boldsymbol{M}}_{\hat{V}_t(i),t-1}^{-1}} \geq \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T \boldsymbol{x}_{i,t}^* + \alpha_{\hat{V}_t(i)}(t)\left\|\boldsymbol{x}_{i,t}^*\right\|_{\overline{\boldsymbol{M}}_{\hat{V}_t(i),t-1}^{-1}}.$$

The instantaneous regret satisfies

$$\begin{aligned}\mathcal{R}_{i,t} &\leq \boldsymbol{\theta}_i^T \boldsymbol{x}_{i,t}^* - \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T \boldsymbol{x}_{i,t}^* + \alpha_{\hat{V}_t(i)}(t)\|\boldsymbol{x}_{i,t}\|_{\overline{\boldsymbol{M}}_{\hat{V}_t(i),t-1}^{-1}} + \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T \boldsymbol{x}_{i,t} - \boldsymbol{\theta}_i^T \boldsymbol{x}_{i,t} \\ &\leq \alpha_{\hat{V}_t(i)}(t)\left\|\boldsymbol{x}_{i,t}^*\right\|_{\overline{\boldsymbol{M}}_{\hat{V}_t(i),t-1}^{-1}} + 2\alpha_{\hat{V}_t(i)}(t)\|\boldsymbol{x}_{i,t}\|_{\overline{\boldsymbol{M}}_{\hat{V}_t(i),t-1}^{-1}}.\end{aligned} \tag{32}$$

The second inequality is satisfied when the estimated clustering structure converge to the true underlying clustering structure. According to Lemma 3, after $T > \max\{\frac{128}{3\delta\lambda_{\min}}\log\left(\frac{128}{3\delta\lambda_{\min}}\right) + 3, \frac{512}{\gamma^2\lambda_{\min}}\left(\frac{d\lambda_{\min}\gamma^2}{512} + 1 + 2\log\left(\frac{2}{\delta}\right) +\right.$

$\sigma^2 d\log\Big(\frac{512\sigma^2}{\lambda_{\min}\gamma^2}\Big)\Big), \frac{4\Omega^2\log(T+1)}{\gamma^2}\}$ time steps, the estimated cluster converge to the true underlying cluster. And according to Lemma 2, 4, we have

$$\sum_{t=\tau+1}^{T}\sum_{j=1}^{m}\sum_{i=1}^{|V_j|}\alpha_{V(i)}(t)\|\boldsymbol{x}\|_{\overline{\boldsymbol{M}}^{-1}_{V(i),t-1}} \le \sqrt{mT\sum_{t=1}^{T}\sum_{j=1}^{m}\sum_{i=1}^{|V_j|}|V_j|\alpha_{V(i)}(t)\|\boldsymbol{x}\|^2_{\overline{\boldsymbol{M}}^{-1}_{V(i),t-1}}}$$
$$\le \sqrt{mT\sum_{j=1}^{m}\sum_{t=\tau+1}^{T}\sum_{i=1}^{|V_j|}|V_j|\|\boldsymbol{x}\|^2_{\overline{\boldsymbol{M}}^{-1}_{V_t(i),t-1}}\left(\sigma\sqrt{2\log\left(\frac{\det\big(\overline{\boldsymbol{M}}_{V(i),t-1}\big)^{1/2}}{\delta}\right)}+1\right)^2},$$

with $\tau = \max\{\frac{128}{3\delta\lambda_{\min}}\log\Big(\frac{128}{3\delta\lambda_{\min}}\Big)+3, \frac{512}{\gamma^2\lambda_{\min}}\Big(\frac{d\lambda_{\min}\gamma^2}{512}+1+2\log\big(\frac{2}{\delta}\big)+\sigma^2 d\log\Big(\frac{512\sigma^2}{\lambda_{\min}\gamma^2}\Big)\Big), \frac{4\Omega^2\log(T+1)}{\gamma^2}\}$. On the other hand, because $\overline{\boldsymbol{M}}_{V(i),t-1}$ contains the aggregated information of all the users in the cluster, which means it adds $|V_j|$, $\forall i \in V_j$ samples at each time step. Therefore,

$$\begin{aligned}
&\sum_{t=\tau+1}^{T}\sum_{j=1}^{m}\sum_{i=1}^{|V_j|}\alpha_{V(i)}(t)\|\boldsymbol{x}\|_{\overline{\boldsymbol{M}}^{-1}_{V(i),t-1}}\\
&\le \sqrt{mT\sum_{j=1}^{m}\sum_{t=\tau+1}^{T}\sum_{i=1}^{|V_j|}|V_j|\|\boldsymbol{x}\|^2_{\overline{\boldsymbol{M}}^{-1}_{V_t(i),t-1}}\left(\sigma\sqrt{2\log\left(\frac{\det\big(\overline{\boldsymbol{M}}_{V(i),t-1}\big)^{1/2}}{\delta}\right)}+1\right)^2}\\
&= \sqrt{mT\sum_{j=1}^{m}\sum_{t=\tau+1}^{|V_j|T}\sum_{i=1}^{|V_j|}\|\boldsymbol{x}\|^2_{\overline{\boldsymbol{M}}^{-1}_{V_t(i),t-1}}\left(\sigma\sqrt{2\log\left(\frac{\det\big(\overline{\boldsymbol{M}}_{V(i),t-1}\big)^{1/2}}{\delta}\right)}+1\right)^2}\\
&= \sqrt{mT\sum_{t=\tau+1}^{|V(i)|T}\sum_{i=1}^{n}\|\boldsymbol{x}\|^2_{\overline{\boldsymbol{M}}^{-1}_{V_t(i),t-1}}\left(\sigma\sqrt{2\log\left(\frac{\det\big(\overline{\boldsymbol{M}}_{V(i),t-1}\big)^{1/2}}{\delta}\right)}+1\right)^2}\\
&\le \sqrt{mT\sum_{t=1}^{|V(i)|T}\sum_{i=1}^{n}\|\boldsymbol{x}\|^2_{\overline{\boldsymbol{M}}^{-1}_{V_t(i),t-1}}\left(\sigma\sqrt{2\log\left(\frac{\det\big(\overline{\boldsymbol{M}}_{V(i),t-1}\big)^{1/2}}{\delta}\right)}+1\right)^2}\\
&\le \sqrt{mnT}\sqrt{2d\log\left(1+\frac{nT}{d}\right)}\left(\sigma\sqrt{d\log\left(1+\frac{nT}{d}\right)+2\log\left(\frac{1}{\delta}\right)}+1\right).
\end{aligned}$$

The cumulative regret satisfies

$$\begin{aligned}
R(T) &= An\tau + \sum_{t=\tau+1}^{T}\sum_{j=1}^{m}\sum_{i=1}^{|V_j|}\mathcal{R}_{i,t}\\
&\le An\tau + 3\sum_{t=\tau+1}^{T}\sum_{j=1}^{m}\sum_{i=1}^{|V_j|}\alpha_{V(i)}(t)\|\boldsymbol{x}\|_{\overline{\boldsymbol{M}}^{-1}_{V(i),t-1}}\\
&\le An\tau + 3\sqrt{mnT}\sqrt{2d\log\left(1+\frac{nT}{d}\right)}\left(\sigma\sqrt{d\log\left(1+\frac{nT}{d}\right)+2\log\left(\frac{1}{\delta}\right)}+1\right),
\end{aligned} \tag{33}$$

where $\tau = \max\{\frac{128}{3\delta\lambda_{\min}}\log\Big(\frac{128}{3\delta\lambda_{\min}}\Big)+3, \frac{512}{\gamma^2\lambda_{\min}}\Big(\frac{d\lambda_{\min}\gamma^2}{512}+1+2\log\big(\frac{2}{\delta}\big)+\sigma^2 d\log\Big(\frac{512\sigma^2}{\lambda_{\min}\gamma^2}\Big)\Big), \frac{4\Omega^2\log(T+1)}{\gamma^2}\}$ and A is the bound of the instantaneous payoff $a_{i,t}$.

E. Proof of Theorem 4

Proof. Before we demonstrate detailed proofs, we need to clarify that $V(i)$ refers to the actual cluster that user i belongs to based on its feature $\boldsymbol{\theta}_i$, while $\hat{V}(i)$ refers to the current estimated cluster that user i belongs to based on its current

estimated feature $\boldsymbol{\omega}_{i,t}$. $\bar{\boldsymbol{\omega}}_{V(i),t} = \frac{1}{|V(i)|}\sum_{k\in V(i)}\boldsymbol{\omega}_{k,t}$ refers to the average estimated feature vector for cluster $V(i)$. For the instantaneous regret, we have

$$\begin{aligned}\mathcal{R}_{i,t} &= \boldsymbol{\theta}_i^T\boldsymbol{x}_{i,t}^* - \boldsymbol{\theta}_i^T\boldsymbol{x}_{i,t},\\ &= \boldsymbol{\theta}_i^T\boldsymbol{x}_{i,t}^* - \bar{\boldsymbol{\omega}}_{V(i),t-1}^T\boldsymbol{x}_{i,t}^* + \bar{\boldsymbol{\omega}}_{V(i),t-1}^T\boldsymbol{x}_{i,t}^* - \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T\boldsymbol{x}_{i,t}^*\\ &\quad + \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T\boldsymbol{x}_{i,t}^* - \bar{\boldsymbol{\omega}}_{V(i),t-1}^T\boldsymbol{x}_{i,t} + \bar{\boldsymbol{\omega}}_{V(i),t-1}^T\boldsymbol{x}_{i,t} - \boldsymbol{\theta}_i^T\boldsymbol{x}_{i,t}.\end{aligned}$$

Because $\boldsymbol{x}_{i,t} = \arg\max_{\boldsymbol{x}\in X_t}(\bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T\boldsymbol{x} + \alpha_{\hat{V}_t(i)}(t)\|\boldsymbol{x}\|_{\overline{\boldsymbol{M}}_{\hat{V}_t(i),t-1}^{-1}})$,

$$\bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T\boldsymbol{x}_{i,t} + \alpha_{\hat{V}_t(i)}(t)\|\boldsymbol{x}_{i,t}\|_{\overline{\boldsymbol{M}}_{\hat{V}_t(i),t-1}^{-1}} \geq \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}^T\boldsymbol{x}_{i,t}^* + \alpha_{\hat{V}_t(i)}(t)\|\boldsymbol{x}_{i,t}^*\|_{\overline{\boldsymbol{M}}_{\hat{V}_t(i),t-1}^{-1}}.$$

Let $C_{i,t}(\boldsymbol{x}) = |\boldsymbol{\theta}_i^T\boldsymbol{x} - \boldsymbol{\omega}_{i,t-1}^T\boldsymbol{x}|$, so we have $C_{i,t}(\boldsymbol{x}) \leq \|\boldsymbol{x}\|_{\boldsymbol{M}_{i,t-1}^{-1}}(\sigma\sqrt{2\log\Big(\frac{\det(\boldsymbol{M}_{i,t-1})^{1/2}}{\delta}\Big)} + 1)$ according to (24). Similarly, let $\overline{C}_{V(i),t}(\boldsymbol{x}) = |\boldsymbol{\theta}_i^T\boldsymbol{x} - \bar{\boldsymbol{\omega}}_{V(i),t-1}^T\boldsymbol{x}|$, we have

$$\begin{aligned}\overline{C}_{V(i),t}(\boldsymbol{x}) &= |\boldsymbol{\theta}_i^T\boldsymbol{x} - \bar{\boldsymbol{\omega}}_{V(i),t-1}^T\boldsymbol{x}|\\ &= |\bar{\boldsymbol{\theta}}_{V(i)}^T\boldsymbol{x} - \frac{1}{|V(i)|}\sum_{k\in V(i)}\boldsymbol{\omega}_{k,t}\boldsymbol{x}|\\ &= \frac{1}{|V(i)|}\sum_{k\in V(i)}|\boldsymbol{\theta}_k^T\boldsymbol{x} - \boldsymbol{\omega}_{k,t}\boldsymbol{x}|\\ &= \frac{1}{|V(i)|}\sum_{k\in V(i)}C_{k,t}(\boldsymbol{x}),\end{aligned}$$

where $\bar{\boldsymbol{\theta}}_{V(i)} = \frac{1}{|V(i)|}\sum_{k\in V(i)}\boldsymbol{\theta}_k$ is the average of the feature vectors of users in the real cluster $V(i)$ (users in $V(i)$ share the same feature vector according to the problem formulation), which can be regarded as the "feature vector" of cluster $V(i)$. The instantaneous regret satisfies

$$\begin{aligned}\mathcal{R}_{i,t} &\leq \boldsymbol{\theta}_i^T\boldsymbol{x}_{i,t}^* - \bar{\boldsymbol{\omega}}_{V(i),t-1}\boldsymbol{x}_{i,t}^* + \alpha_{\hat{V}_t(i)}(t)\|\boldsymbol{x}_{i,t}\|_{\overline{\boldsymbol{M}}_{\hat{V}_t(i),t-1}^{-1}} - \alpha_{\hat{V}_t(i)}(t)\|\boldsymbol{x}_{i,t}^*\|_{\overline{\boldsymbol{M}}_{\hat{V}_t(i),t-1}^{-1}}\\ &\quad + \bar{\boldsymbol{\omega}}_{V(i),t-1}\boldsymbol{x}_{i,t} - \boldsymbol{\theta}_i^T\boldsymbol{x}_{i,t} + (\bar{\boldsymbol{\omega}}_{V(i),t-1} - \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1})^T(\boldsymbol{x}_{i,t}^* - \boldsymbol{x}_{i,t})\\ &\leq \overline{C}_{V(i),t}(\boldsymbol{x}_{i,t}^*) + \overline{C}_{V(i),t}(\boldsymbol{x}_{i,t}) + \alpha_{\hat{V}_t(i)}(t)\|\boldsymbol{x}_{i,t}\|_{\overline{\boldsymbol{M}}_{\hat{V}_t(i),t-1}^{-1}} + (\bar{\boldsymbol{\omega}}_{V(i),t-1} - \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1})^T(\boldsymbol{x}_{i,t}^* - \boldsymbol{x}_{i,t}).\end{aligned} \tag{34}$$

According to Lemma 1, for agents i, j in the same cluster $d(i,j) = \|\boldsymbol{\omega}_{i,t-1} - \boldsymbol{\omega}_{j,t-1}\| \leq \frac{\Omega}{1+\beta_t}$. Based on this lemma we can also easily get

$$\Big\|\boldsymbol{\omega}_{i,t-1} - \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1}\Big\| \leq \frac{\Omega}{1+\beta_t}. \tag{35}$$

Thus, we have

$$\begin{aligned}(\bar{\boldsymbol{\omega}}_{V(i),t-1} - \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1})^T(\boldsymbol{x}_{i,t}^* - \boldsymbol{x}_{i,t}) \leq&(\bar{\boldsymbol{\omega}}_{V(i),t-1} - \boldsymbol{\theta}_{i,t-1})^T(\boldsymbol{x}_{i,t}^* - \boldsymbol{x}_{i,t}) + (\boldsymbol{\theta}_{i,t-1} - \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1})^T(\boldsymbol{x}_{i,t}^* - \boldsymbol{x}_{i,t})\\ \leq&(\bar{\boldsymbol{\omega}}_{V(i),t-1} - \boldsymbol{\theta}_{i,t-1})^T(\boldsymbol{x}_{i,t}^* - \boldsymbol{x}_{i,t}) + (\boldsymbol{\theta}_{i,t-1} - \boldsymbol{\omega}_{i,t-1})^T(\boldsymbol{x}_{i,t}^* - \boldsymbol{x}_{i,t})\\ &+ (\boldsymbol{\omega}_{i,t-1} - \bar{\boldsymbol{\omega}}_{\hat{V}_t(i),t-1})^T(\boldsymbol{x}_{i,t}^* - \boldsymbol{x}_{i,t})\\ \leq&\overline{C}_{V(i),t}(\boldsymbol{x}^*) + \overline{C}_{V(i),t}(\boldsymbol{x}_{i,t}) + C_{i,t}(\boldsymbol{x}^*) + C_{i,t}(\boldsymbol{x}_{i,t}) + \frac{2\Omega}{1+\beta_t}.\end{aligned}$$

Lemma 7. *For any $i \in V$ and $\boldsymbol{x} \in \mathbb{R}^d$, $C_{i,t}(\boldsymbol{x})$ satisfies*

$$\begin{aligned}\sum_{t=1}^T C_{i,t}(\boldsymbol{x}) &\leq \sqrt{T\sum_{t=1}^T C_{i,t}^2(\boldsymbol{x})}\\ &\leq\sqrt{T}\left(\sigma\sqrt{d\log\left(1+\frac{T}{d}\right) + 2\log\left(\frac{1}{\delta}\right)} + 1\right) \times \sqrt{d\log\left(1+\frac{T}{d}\right)},\end{aligned} \tag{36}$$

Proof. For any $i \in V$ and $\boldsymbol{x} \in \mathbb{R}^d$, $C_{i,t}(\boldsymbol{x})$ satisfies

$$\begin{aligned}
\sum_{t=1}^{T} C_{i,t}(\boldsymbol{x}) &\leq \sqrt{T \sum_{t=1}^{T} C_{i,t}^2(\boldsymbol{x})} \\
&\leq \sqrt{T \sum_{t=1}^{T} \|\boldsymbol{x}\|^2_{\boldsymbol{M}^{-1}_{i,t-1}} \left(\sigma \sqrt{2\log\left(\frac{\det(\boldsymbol{M}_{i,t-1})^{1/2}}{\delta}\right)} + 1\right)^2} \\
&\leq \left(\sigma \sqrt{2\log\left(\frac{\det(\boldsymbol{M}_{i,T})^{1/2}}{\delta}\right)} + 1\right) \sqrt{T \sum_{t=1}^{T} \|\boldsymbol{x}\|^2_{\boldsymbol{M}^{-1}_{i,t-1}}} \\
&\leq \left(\sigma \sqrt{2\log\left(\frac{\det(\boldsymbol{M}_{i,T})^{1/2}}{\delta}\right)} + 1\right) \sqrt{T \log\left(\frac{\det(\boldsymbol{M}_{i,T})}{\det(\boldsymbol{M}_{i,0})}\right)} \\
&\leq \sqrt{T} \sqrt{d\log\left(1 + \frac{T}{d}\right)} \left(\sigma \sqrt{d\log\left(1 + \frac{T}{d}\right) + 2\log\left(\frac{1}{\delta}\right)} + 1\right),
\end{aligned}$$

where the first inequality holds by Cauthy-Schwarz and the last step is based on Lemma 4 and Lemma 5. □

According to Lemma 7, $C_{i,t}$ satisfies

$$\sum_{t=1}^{T} C_{i,t}(\boldsymbol{x}) \leq \sqrt{T} \sqrt{d\log\left(1 + \frac{T}{d}\right)} \left(\sigma \sqrt{d\log\left(1 + \frac{T}{d}\right) + 2\log\left(\frac{1}{\delta}\right)} + 1\right).$$

Following the same process as Lemma 7, we have

$$\sum_{t=1}^{T} \overline{C}_{V(i),t}(\boldsymbol{x}) \leq \sqrt{T} \sqrt{d\log\left(1 + \frac{T}{d}\right)} \left(\sigma \sqrt{d\log\left(1 + \frac{T}{d}\right) + 2\log\left(\frac{1}{\delta}\right)} + 1\right),$$

And

$$\begin{aligned}
\sum_{t=1}^{T} \alpha_{\hat{V}_t(i)}(t) \|\boldsymbol{x}\|_{\overline{\boldsymbol{M}}^{-1}_{\hat{V}_t(i),t-1}} &\leq \sqrt{T \sum_{t=1}^{T} \alpha^2_{\hat{V}_t(i)}(t) \|\boldsymbol{x}\|^2_{\overline{\boldsymbol{M}}^{-1}_{\hat{V}_t(i),t-1}}} \\
&\leq \sqrt{T \sum_{t=1}^{T} \|\boldsymbol{x}\|^2_{\overline{\boldsymbol{M}}^{-1}_{\hat{V}_t(i),t-1}} \left(\sigma \sqrt{2\log\left(\frac{\det\left(\overline{\boldsymbol{M}}_{\hat{V}_t(i),t-1}\right)^{1/2}}{\delta}\right)} + 1\right)^2} \\
&\leq \sqrt{T} \sqrt{d\log\left(1 + \frac{nT}{d}\right)} \left(\sigma \sqrt{d\log\left(1 + \frac{nT}{d}\right) + 2\log\left(\frac{1}{\delta}\right)} + 1\right),
\end{aligned}$$

Putting all together, we have

$$\begin{aligned}
R_i(T) &= \sum_{t=1}^{T} (2\overline{C}_{V(i),t}(\boldsymbol{x}^*_{i,t}) + 2\overline{C}_{V(i),t}(\boldsymbol{x}_{i,t}) + C_{i,t}(\boldsymbol{x}^*) + C_{i,t}(\boldsymbol{x}_{i,t}) + \frac{4}{1+\beta_t} + \alpha_{\hat{V}_t(i)}(t) \|\boldsymbol{x}_{i,t}\|_{\bar{M}^{-1}_{\hat{V}_t(i),t-1}}) \\
&\leq 6\sqrt{T} \sqrt{d\log\left(1 + \frac{T}{d}\right)} \left(\sigma \sqrt{d\log\left(1 + \frac{T}{d}\right) + 2\log\left(\frac{1}{\delta}\right)} + 1\right) + \sum_{t=1}^{T} \frac{2\Omega}{1+\beta_t} \\
&\quad + \sqrt{T} \sqrt{d\log\left(1 + \frac{nT}{d}\right)} \left(\sigma \sqrt{d\log\left(1 + \frac{nT}{d}\right) + 2\log\left(\frac{1}{\delta}\right)} + 1\right),
\end{aligned} \tag{37}$$

select $\beta_t = \sqrt{\frac{t}{\log(t+1)}}$, and because $\sum_{t=s}^{T} \frac{1}{t^\kappa} \leq \frac{T^{1-\kappa}}{1-\kappa}$, $\forall \kappa \in [0,1)$ (Yi et al., 2020), then

$$\sum_{t=1}^{T} \frac{2\Omega}{1+\beta_t} \leq \sum_{t=1}^{T} \frac{2\Omega}{\sqrt{\frac{t}{\log(t+1)}}} \leq 2\Omega\sqrt{\log(T+1)} \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 4\Omega\sqrt{T\log(T+1)}.$$

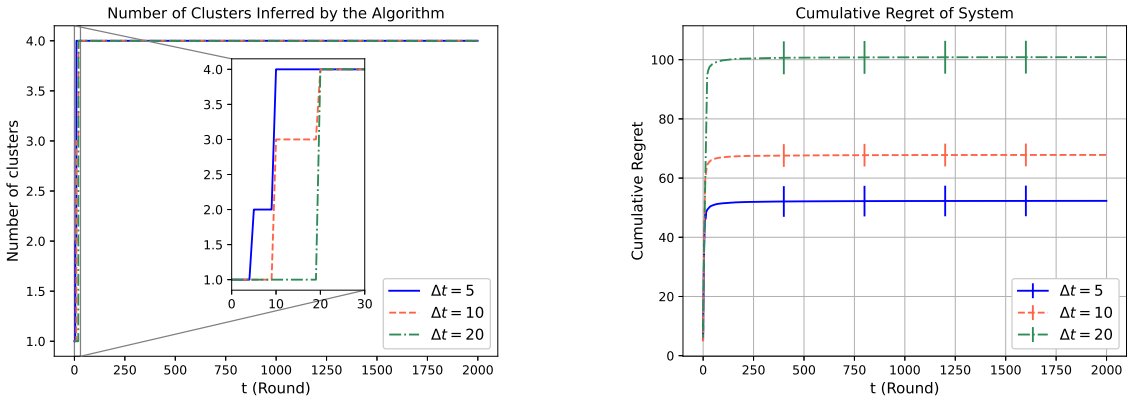
The cumulative regret satisfies

$$\begin{aligned} R(T) \leq\ & 6n\sqrt{T}\sqrt{d\log\left(1+\frac{T}{d}\right)}\left(\sigma\sqrt{d\log\left(1+\frac{T}{d}\right)+2\log\left(\frac{1}{\delta}\right)}+1\right)+4\Omega n\sqrt{T\log T} \\ & + n\sqrt{T}\sqrt{d\log\left(1+\frac{nT}{d}\right)}\left(\sigma\sqrt{d\log\left(1+\frac{nT}{d}\right)+2\log\left(\frac{1}{\delta}\right)}+1\right). \end{aligned} \tag{38}$$

That completes the proof. □

F. Discussion

We propose a simple tweak on CLUB-HG (Algorithm 1) to speed up the clustering operation using hedonic game. According to Theorem 2, the Nash equilibrium can be computed in polynomial time. To empirically reduce the computational cost, we can perform the hedonic game not at every time step, but only every Δt time steps (i.e. $t \in \{..., t', t'+\Delta T, t'+2\Delta T, ...\}$). We test this "lazy" version of CLUB-HG algorithm under different Δt on synthetic dataset and the result is shown in Figure 4.



(a) Number of clusters

(b) Empirical cumulative regret

Figure 4. Results of changing Δt (synthetic dataset).

In Figure 4, the regret increases with the growth of Δt. However, even the regret of the CLUB-HG algorithm with $\Delta t = 20$ is much lower than the regret of CLUB and SCLUB algorithm as shown in Figure 1. In general, this "lazy" version of CLUB-HG does not produce much additional regret, as it still outperforms most of the state-of-the-art algorithms, but significantly reduces the computation time. Furthermore, according to Figure 4a, if Δt is large enough, the algorithm can converge to the true clustering structure within one inference step. Therefore, choosing proper and small values, such as $\Delta t = 10, 20$, can speed up CLUB-HG and keep its superiority in performance at the same time.