Demystifying Online Clustering of Bandits: Enhanced Exploration Under Stochastic and Smoothed Adversarial Contexts

香港中文大學 The Chinese University of Hong Kong

Zhuohua Li†, Maoli Liu†, Xiangxiang Dai, and John C.S. Lui*

The Chinese University of Hong Kong

Introduction

- **We address an open problem in online clustering of bandits**, an extension of contextual linear bandits that groups similar users into clusters, utilizing shared features to improve learning efficiency.

Problem Formulation

- There are u users. Each user $i \in [u]$ is associated with an *unknown* preference vector $\boldsymbol{\theta}_i \in \mathbb{R}^d$.
- The users are separated into m ($m \ll u$) disjoint clusters, such that:
  - Users i, j in the same cluster satisfy $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$.
  - Users i, j from different clusters satisfy $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| \geq \gamma$.
- At each round $t = 1, 2, \ldots, T$, the learner receives a user index $i_t \in [u]$ and a finite set of arms $\mathcal{A}_t \subset \mathcal{A} \subset \mathbb{R}^d$ where $|\mathcal{A}_t| = K$. Each arm $a \in \mathcal{A}$ is associated with a feature vector $\boldsymbol{x}_a \in \mathbb{R}^d$. The learner assigns an appropriate cluster V_t for user i_t, recommends an arm $a_t \in \mathcal{A}_t$, and receives a reward $r_t = \boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}_{i_t} + \eta_t$, where η_t is noise.
- Let $a_t^* = \arg\max_{a \in \mathcal{A}_t} \boldsymbol{x}_a^\top \boldsymbol{\theta}_{i_t}$ be the optimal arm at time t. The goal is to minimize the expected cumulative regret:

$$\mathbb{E}[R(T)] = \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{x}_{a_t^*}^\top \boldsymbol{\theta}_{i_t} - \boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}_{i_t}\right)\right]$$

Open Problem

- **Existing algorithms (Gentile et al., 2014) rely on strong data diversity assumptions:**
  1. At each time t, vectors $\{\boldsymbol{x}_a\}_{a \in \mathcal{A}_t}$ are i.i.d. sampled from a fixed distribution $\boldsymbol{X}$ with $\lambda_{\min}(\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]) = \lambda_x$;
  2. For any unit vector $\boldsymbol{z} \in \mathbb{R}^d$, $(\boldsymbol{z}^\top \boldsymbol{X})^2$ is σ^2-sub-Gaussian with $\sigma^2 \leq \frac{\lambda_x^2}{8\log(4K)}$.
- **Open problem posed by Gentile et al. (2014): Can we remove the i.i.d. and other statistical assumptions?**
- Some follow-up work weakens these assumptions but suffers from deteriorated regret bounds.

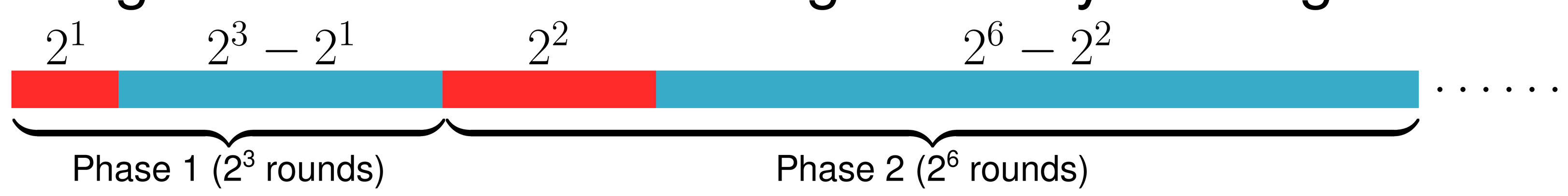
Contributions

- **Regret under Weaker Assumptions:** We solve the open problem by proposing UniCLUB and PhaseUniCLUB, which rely solely on Assumption 1 with regret $\widetilde{O}(\sqrt{T})$.
- **Removal of i.i.d. Assumption:** We remove the i.i.d. assumption by introducing the *smoothed adversarial context setting* and proposing SACLUB with regret $\widetilde{O}(\sqrt{T})$.

Methodologies

- **Key Idea: Add additional uniform exploration to ensure accurate clustering while maintaining low regret.**
- **UniCLUB: When the cluster gap γ is *known***
  - Use γ to determine when clustering is reliable.
  - Uniform exploration until all $\boldsymbol{\theta}_i$ are estimated accurately.

- **PhaseUniCLUB: When the cluster gap γ is *unknown***
  - It is unclear how much uniform exploration is sufficient.
  - Divide the time horizon into phases and perform an appropriate amount of uniform exploration in each phase.
  - Phase lengths and exploration budgets are carefully designed to balance clustering accuracy and regret.



Smoothed Adversarial Context Setting

- **Key Idea: The intrinsic diversity of contexts makes explicit exploration unnecessary.**
- Each feature vector is first arbitrarily chosen by an adversary and then perturbed by a noise vector sampled from a truncated multivariate Gaussian distribution.
- This setup is more practical and aligns more closely with the original setting of contextual linear bandits.

Theoretical Results

Theorem 1. *With Assumption 1 and assuming the cluster gap γ is known, algorithm* UniCLUB *satisfies:*

$$\mathbb{E}[R(T)] = \widetilde{O}\left(\frac{ud}{\gamma^2\lambda_x} + d\sqrt{mT}\right).$$

Theorem 2. *With Assumption 1,* PhaseUniCLUB *satisfies:*

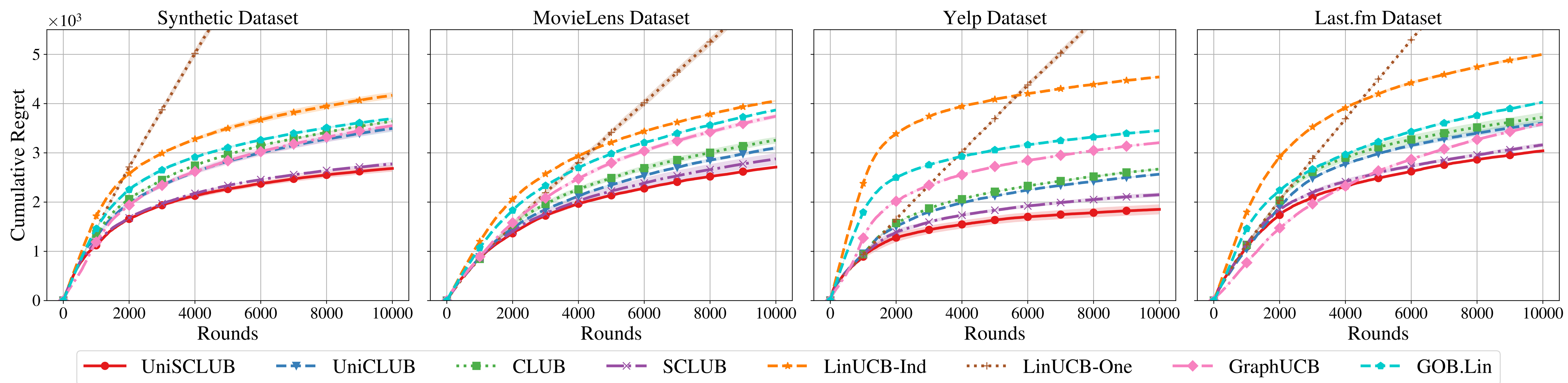
$$\mathbb{E}[R(T)] = \widetilde{O}\left(\frac{ud}{\gamma^5\lambda_x^2} + \left(\frac{ud}{\lambda_x}\right)^{\frac{2}{3}} T^{\frac{1}{3}} + d\sqrt{mT}\right).$$

Theorem 3. *Under the smoothed adversarial context setting, algorithm* SACLUB *satisfies:*

$$\mathbb{E}[R(T)] = \widetilde{O}\left(\frac{ud}{\gamma^2\widetilde{\lambda}_x} + d\sqrt{mT}\right),$$

where $\widetilde{\lambda}_x = \frac{C}{\log K}$ for some constant C.

Evaluations



Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 757–765. 22–24 Jun 2014.

† Zhuohua Li and Maoli Liu contributed equally to this work. Maoli Liu is the corresponding author. * The work of John C.S. Lui is supported in part by the RGC SRFS2122-4S02.