Towards Efficient Conversational Recommendations: Expected Value of Information Meets Bandit Learning

Zhuohua Li, Maoli Liu, Xiangxiang Dai, John C.S. Lui
The Chinese University of Hong Kong

Introduction

- Recommender systems must adapt to user preferences by learning from feedback, such as click rates.
- **Conversational Recommender Systems (CRS)** can also proactively query users to obtain additional feedback.

Conversational Recommendation

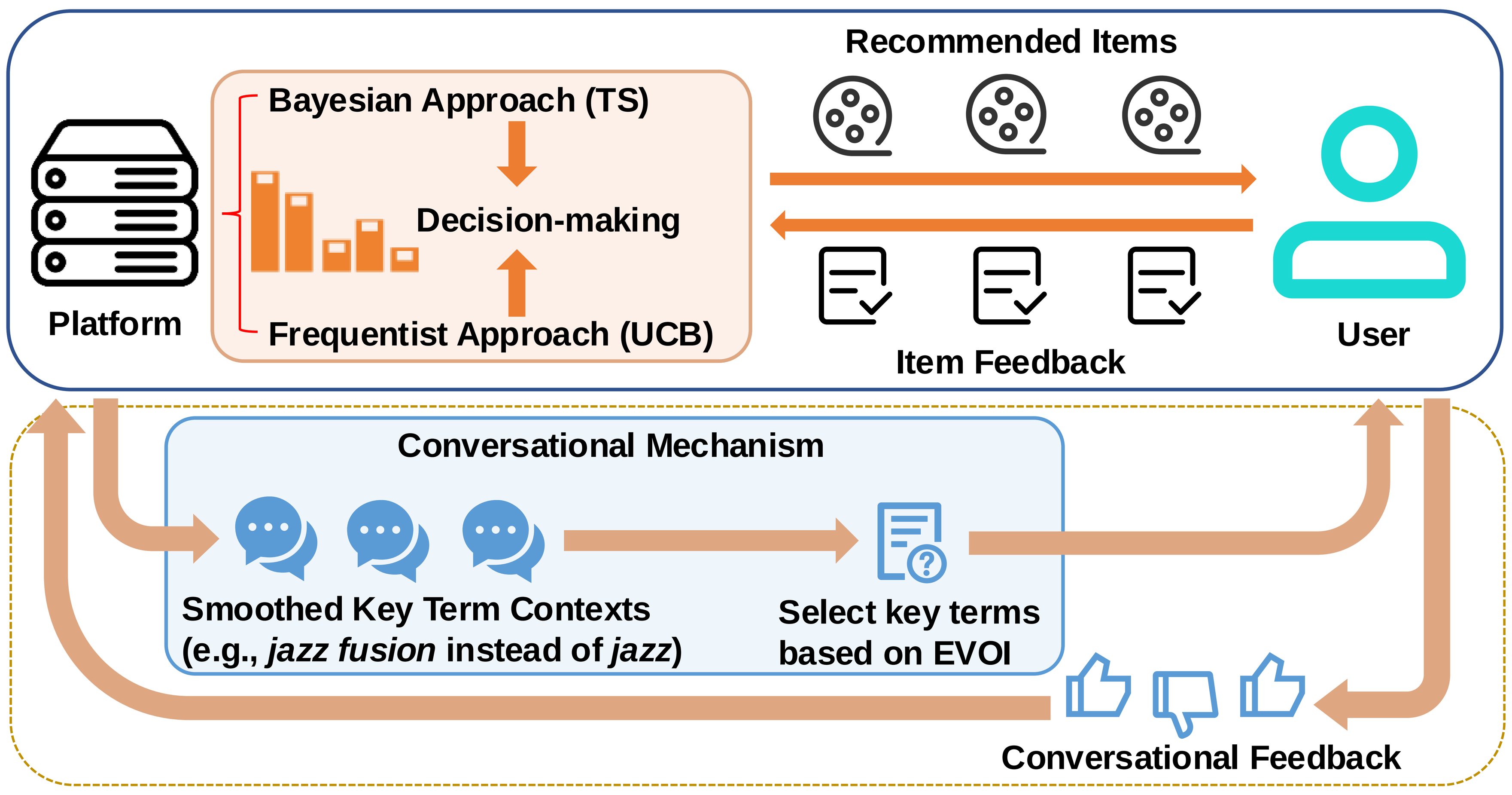

Fig. 1: The workflow of our algorithm design.

Example: CRS can ask users questions to quickly elicit their preferences, e.g., "Do you enjoy jazz music?"

Contributions: EVOI + Conversational Bandits

Two key techniques:

1. *Gradient-based EVOI*: Replaces expensive Bayesian posterior updates with efficient incremental updates using SGD.
2. *Smoothed key term contexts*: Adds random perturbations to queries to uncover finer-grained user preferences.

EVOI provides an effective query selection strategy, while conversational bandits offer long-term performance guarantees.

Two algorithms in Bayesian and frequentist frameworks:

1. ConTS-EVOI: Based on Thompson Sampling (Bayesian).
2. ConUCB-EVOI: Based on LinUCB (frequentist).

Existing Studies and Motivation

Expected Value of Information (EVOI):

- Quantifies the value of a query based on its expected improvement in recommendation quality.

Conversational Bandits:

- Models conversational recommendation as a multi-armed bandit problem to balance exploration and exploitation.

Limitations:

- Traditional EVOI adopts a myopic (greedy) strategy and lacks theoretical guarantees for long-term performance metrics.
- Existing conversational bandit algorithms lack a principled mechanism for selecting informative queries.

Problem Formulation

Interaction Protocol:

- For each time step $t = 1, 2, \dots, T$:
  - The CRS receives a set of arms $\mathcal{A}_t$ (i.e., recommendable items). Each arm $a \in \mathcal{A}_t$ is associated with a feature vector $\boldsymbol{x}_a \in \mathbb{R}^d$.
  - The CRS selects an arm $a_t \in \mathcal{A}_t$ (i.e., recommend an item), and observes a reward $r_t = \boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}^* + \eta_t$ (i.e., whether the user clicks on the item), where $\boldsymbol{\theta}^*$ is the unknown user preference vector, and η_t is noise.
  - The CRS optionally initiates a query $k_t \in \mathcal{K}$ and observes an additional reward $\tilde{r}_t = \boldsymbol{x}_{k_t}^\top \boldsymbol{\theta}^* + \tilde{\eta}_t$, where each query $k \in \mathcal{K}$ is also associated with a feature vector $\boldsymbol{x}_k \in \mathbb{R}^d$.
- The objective is to minimize the cumulative regret:

$$R(T) = \sum_{t=1}^{T} \left(\max_{a \in \mathcal{A}_t} \boldsymbol{x}_a^\top \boldsymbol{\theta}^* - \boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}^* \right),$$

while using as few queries as possible.

Theoretical Results

Theorem 1. *With probability at least $1-\delta$, the cumulative regret of ConTS-EVOI scales in $O\left(d\sqrt{T}\log(T)\right)$.*

Theorem 2. *With probability at least $1-\delta$, the cumulative regret of ConUCB-EVOI scales in $O\left(\sqrt{dT\log(T)} + d\right)$.*

Both algorithms achieve a $\sqrt{d}$ improvement in their dependence on the time horizon T, compared to prior approaches.

Evaluations

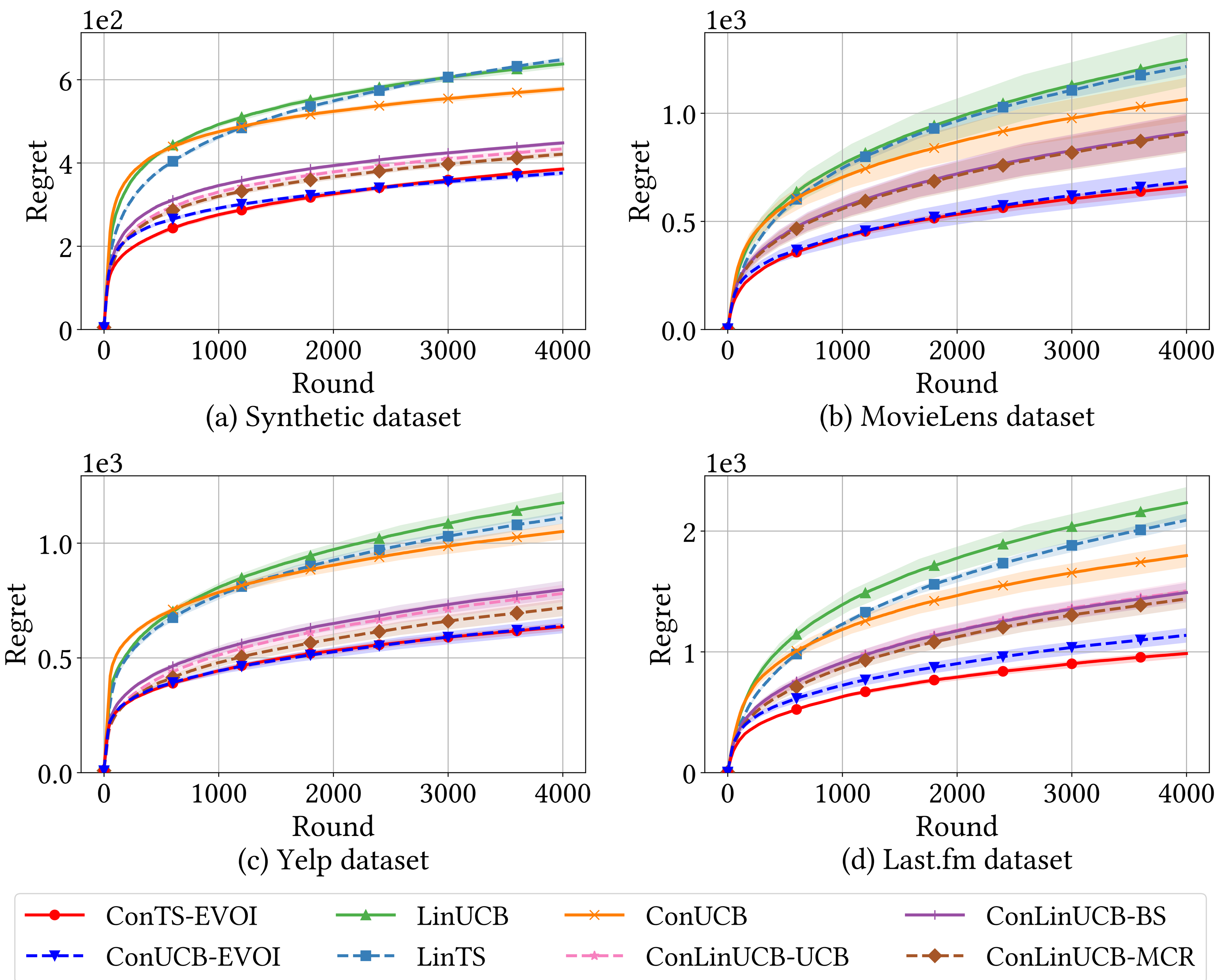

Fig. 2: Comparison of cumulative regret (lower is better).

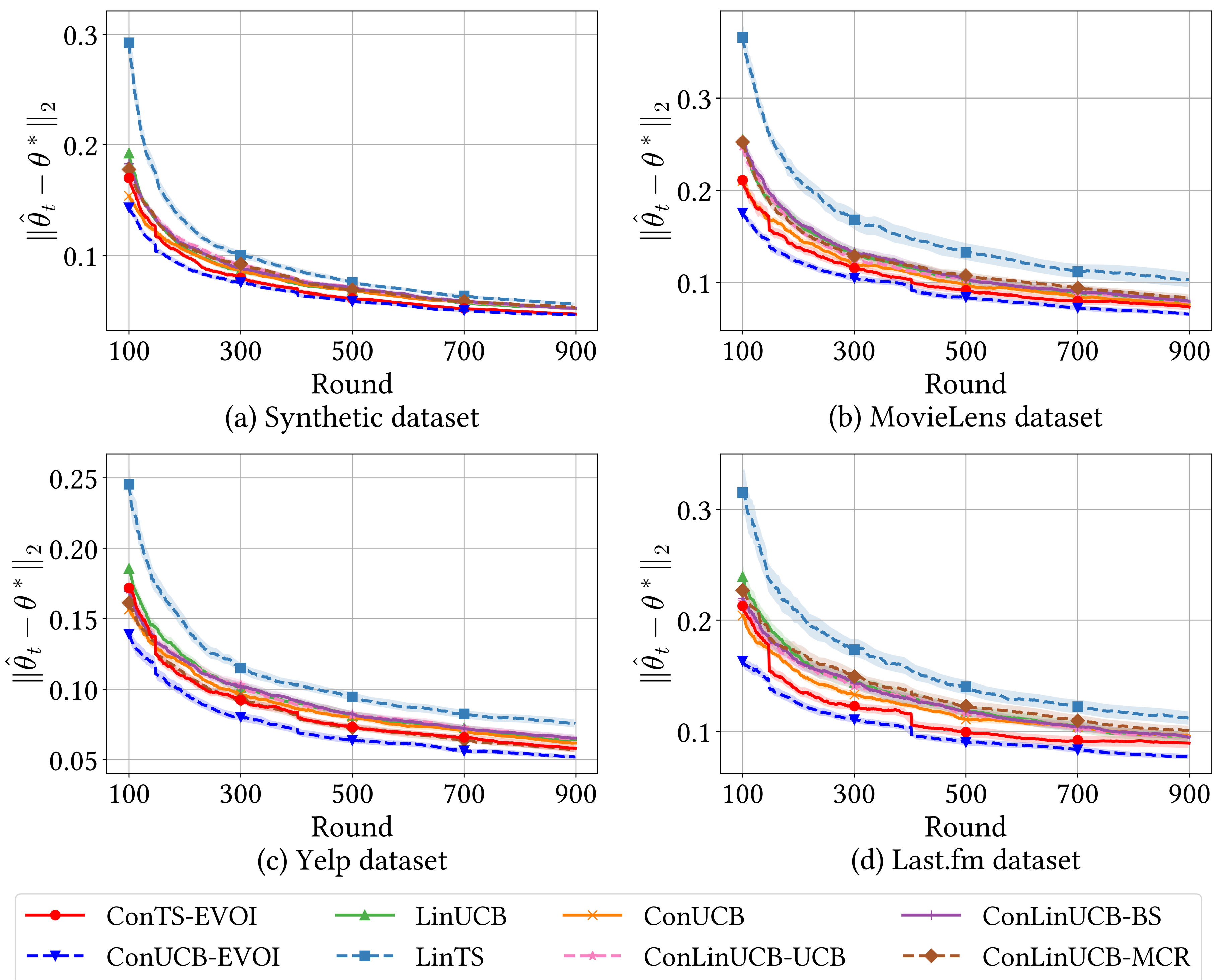

Fig. 3: Comparison of estimation error (lower is better).