Leveraging the Power of Conversations: Optimal Key Term Selection in Conversational Contextual Bandits

Maoli Liu
The Chinese University of Hong Kong
Hong Kong, China
mlliu@cse.cuhk.edu.hk

Zhuohua Li[*] [†]
Guangzhou Institute of Technology, Xidian University
Guangzhou, Guangdong, China
zhli@cse.cuhk.edu.hk

Xiangxiang Dai
The Chinese University of Hong Kong
Hong Kong, China
xxdai23@cse.cuhk.edu.hk

John C.S. Lui
The Chinese University of Hong Kong
Hong Kong, China
cslui@cse.cuhk.edu.hk

Abstract

Conversational recommender systems proactively query users with relevant "*key terms*" and leverage the feedback to elicit users' preferences for personalized recommendations. Conversational contextual bandits, a prevalent approach in this domain, aim to optimize preference learning by balancing exploitation and exploration. However, several limitations hinder their effectiveness in real-world scenarios. First, existing algorithms employ key term selection strategies with insufficient exploration, often failing to thoroughly probe users' preferences and resulting in suboptimal preference estimation. Second, current algorithms typically rely on deterministic rules to initiate conversations, causing unnecessary interactions when preferences are well-understood and missed opportunities when preferences are uncertain. To address these limitations, we propose three novel algorithms: CLiSK, CLiME, and CLiSK-ME. CLiSK introduces *smoothed key term contexts* to enhance exploration in preference learning, CLiME *adaptively initiates conversations* based on preference uncertainty, and CLiSK-ME integrates both techniques. We theoretically prove that all three algorithms achieve a tighter regret upper bound of $O(\sqrt{dT \log T})$ with respect to the time horizon T, improving upon existing methods. Additionally, we provide a matching lower bound $\Omega(\sqrt{dT})$ for conversational bandits, demonstrating that our algorithms are nearly minimax optimal. Extensive evaluations on both synthetic and real-world datasets show that our approaches achieve at least a 14.6% improvement in cumulative regret.

CCS Concepts

• Information systems → Recommender systems; **• Theory of computation → Online learning algorithms**; **Online learning theory**.

Keywords

Conversational Recommendation, Preference Learning, Contextual Bandits, Online Learning

ACM Reference Format:
Maoli Liu, Zhuohua Li, Xiangxiang Dai, and John C.S. Lui. 2025. Leveraging the Power of Conversations: Optimal Key Term Selection in Conversational Contextual Bandits. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3711896.3737025

KDD Availability Link:
The source code of this paper has been made publicly available at https://doi.org/10.5281/zenodo.15490021.

1 Introduction

Recommender systems play a crucial role in applications like movie recommendations, online advertising, and personalized news feeds, where providing relevant and engaging content is essential for user satisfaction. To cater to diverse user interests, recommender systems are designed to interact with users and continuously learn from their feedback. For instance, in product and news recommendations, the system can monitor users' real-time click rates and accordingly refine its recommendations. Modern recommender systems incorporate advanced online learning techniques to adapt in real time and uncover previously unknown user preferences.

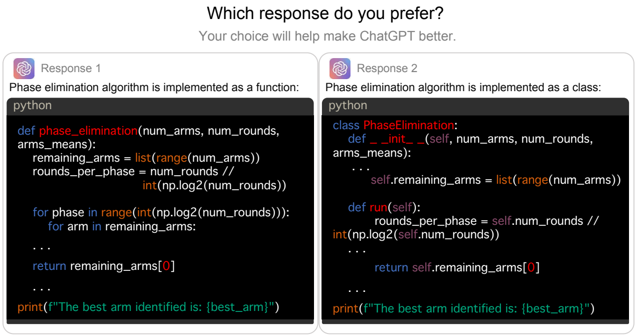
A fundamental challenge in recommender systems is the trade-off between *exploration* (i.e., recommending new items to uncover users' unknown preferences) and *exploitation* (i.e., recommending items that align with users' historical preferences). Contextual bandits [15] address this trade-off by enabling the system to learn from user interactions continuously while optimizing recommendations without compromising the user experience. In this framework, each item to be recommended is treated as an "*arm*", represented by a feature vector. At each round, the agent (i.e., the recommender system) recommends an arm to the user based on historical interactions and the context of each arm, and then receives feedback/rewards (e.g., clicks). The objective of the algorithm executed by the agent is to design an arm recommendation strategy that maximizes cumulative reward (or equivalently, minimizes cumulative regret) over time.

Another major challenge in recommender systems is the "*cold start*" problem, where the system initially lacks sufficient data about

[*] Zhuohua Li is the corresponding author.
[†] Also with The Chinese University of Hong Kong.

KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3737025



Which response do you prefer?

Your choice will help make ChatGPT better.

Response 1

Phase elimination algorithm is implemented as a function:

```python
def phase_elimination(num_arms, num_rounds,
arms_means):
    remaining_arms = list(range(num_arms))
    rounds_per_phase = num_rounds //
                       int(np.log2(num_rounds))

    for phase in range(int(np.log2(num_rounds))):
        for arm in remaining_arms:
    . . .
    return remaining_arms[0]
    . . .
print(f"The best arm identified is: {best_arm}")
```

Response 2

Phase elimination algorithm is implemented as a class:

```python
class PhaseElimination:
    def _ _init_ _(self, num_arms, num_rounds,
arms_means):
     . . .
        self.remaining_arms = list(range(num_arms))

    def run(self):
        rounds_per_phase = self.num_rounds //
int(np.log2(self.num_rounds))
    . . .
        return self.remaining_arms[0]
    . . .
print(f"The best arm identified is: {best_arm}")
```

Figure 1: Illustration of conversational recommendation by ChatGPT, where users select their preferred response from presented options.

new users' preferences, making accurate recommendations difficult. Conversational recommender systems (CRSs) [5, 11, 23, 30] have emerged as a promising solution. Unlike traditional systems that rely solely on feedback from recommended items, CRSs can actively initiate queries with users to collect richer feedback and quickly infer their preferences. For example, as shown in Figure 1, platforms like ChatGPT occasionally present users with multiple response options and allow them to select their preferred one. Through these interactions, ChatGPT can refine its understanding and improve future responses to better align with user preferences. To model these interactions, conversational contextual bandits [29] are proposed as a natural extension of contextual bandits. In this framework, besides recommending items (arms) and observing arm-level feedback, the agent can proactively prompt users with questions about key terms and receive key term-level feedback. The key terms are related to a subset of arms, providing valuable insights into users' preferences and improving recommendation quality.

Despite recent advances in conversational contextual bandits [25, 27, 28], existing approaches still face the following limitations:

- **Insufficient Exploration in Key Term Selection**: Existing studies about conversational bandits fail to sufficiently explore key terms, limiting their effectiveness in preference learning. Zhang et al. [29] introduce the ConUCB algorithm with a regret upper bound of $O(d\sqrt{T}\log T)$, where d is the dimension and T is the number of rounds. However, despite incorporating additional queries about key terms, the method does not yield substantial improvement over non-conversational approaches. Since then, improving regret through conversational interactions has remained an open problem in the field. Wang et al. [25] and Yang et al. [28] introduce an additional assumption that the key term set spans $\mathbb{R}^d$ and propose the ConLinUCB-BS and ConDuel algorithms, respectively. The two algorithms reduce a $\sqrt{\log T}$ term in the regret, but worsen the dependence on d (as discussed in Section 4.4), resulting in a suboptimal regret bound. To achieve optimal regret, more explorative key term selection strategies are needed to efficiently gather informative user feedback and improve learning efficiency.
- **Inflexible Conversation Mechanism**: Existing conversational bandit algorithms [27, 29] often use a deterministic function to control the frequency of conversations. Specifically, the agent can only initiate Q conversations at once per P rounds, where P and Q are fixed integers. However, this rigid approach is impractical and insufficient in real-world scenarios. For example, in a music streaming service, a fixed-frequency approach may cause unnecessary interactions when users' preferences are already well-understood, disrupting the listening experience. Conversely, it may fail to collect feedback when the uncertainty is high, leading to suboptimal recommendations. To address these limitations, a more adaptive conversation mechanism is needed to adjust the interaction frequency based on the preference uncertainty.

Motivated by these observations, we develop three algorithms aimed at improving conversational contextual bandits. To start, we introduce the concept of "*smoothed key term contexts*", inspired by the smoothed analysis for contextual bandits [13], and propose the Conversational LinUCB with Smoothed Key terms (CLiSK) algorithm. Specifically, CLiSK launches conversations at a fixed frequency, similar to Zhang et al. [29], but greedily selects key terms that are slightly perturbed by Gaussian noise. For example, in movie recommendations, instead of asking directly about a genre like "comedy" or "drama", CLiSK blends elements of related genres, such as "comedy-drama" or "dark comedy". This approach helps the system explore users' preferences in a more nuanced manner. We will show that these small perturbations have *strong theoretical implications*, allowing the agent to explore the feature space more effectively and speed up the learning process.

We next develop the Conversational LinUCB with Minimum Eigenvalues (CLiME) algorithm, which introduces an *adaptive conversational mechanism* driven by preference uncertainty. Unlike the fixed-frequency approach of Wang et al. [25], Zhang et al. [29], CLiME assesses preference uncertainty and initiates conversations only when the uncertainty is high, thereby maximizing information gain while avoiding unnecessary interactions. When a conversation is triggered, CLiME selects key terms that target the areas of highest uncertainty within the feature space, rapidly refining user preferences. This adaptive approach not only ensures that conversations are timely and relevant, but also improves the user experience. Additionally, we design a family of *uncertainty checking functions* to determine when to assess the uncertainty, offering greater flexibility and better alignment with diverse applications.

The smoothed key term contexts approach in CLiSK and the adaptive conversation technique in CLiME are orthogonal, allowing them to be applied independently or in combination. Therefore, we further propose the CLiSK-ME algorithm, which integrates both techniques to maximize exploration efficiency and adaptively adjust user interactions. By leveraging the strengths of both methods, CLiSK-ME enhances exploration efficiency and optimizes user interactions for improved preference learning.

Our algorithms introduce advanced key term selection strategies, significantly enhancing the efficiency of conversational contextual bandits. Theoretically, we prove that CLiSK achieves a regret upper bound of $O(\sqrt{dT\log T} + d)$, while CLiME and CLiSK-ME achieve a regret upper bound of $O(\sqrt{dT\log T})$. Notably, all three algorithms reduce the dependence on T by a factor of $\sqrt{d}$ compared to prior studies. To the best of our knowledge, our work is the first to achieve the $\widetilde{O}(\sqrt{dT})$ regret in the conversational bandit literature. In addition, we establish a matching lower bound of $\Omega(\sqrt{dT})$, showing that our algorithms are minimax optimal up to logarithmic factors.

In summary, our contributions are listed as follows.

- We propose three novel conversational bandit algorithms: CLiSK with smoothed key term contexts, CLiME with an adaptive conversation mechanism, and CLiSK-ME, which integrates both for improved preference learning.
- We establish the minimax optimality of our algorithms by proving regret upper bounds of $O(\sqrt{dT \log T} + d)$ for CLiSK and $O(\sqrt{dT \log T})$ for CLiME and CLiSK-ME, along with a matching lower bound of $\Omega(\sqrt{dT})$. These results underscore the theoretical advancements achieved by our methods.
- We conduct extensive evaluations on both synthetic and real-world datasets, showing that our algorithms reduce regret by over 14.6% compared to baselines.

2 Problem Formulation

In conversational contextual bandits, an agent interacts with a user over $T \in \mathbb{N}_+$ rounds. The user's preferences are represented by a fixed but *unknown* vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$, where d is the dimension. The agent's goal is to learn $\boldsymbol{\theta}^*$ to recommend items that align with the user's preferences. There exists a finite arm set denoted by $\mathcal{A}$, where each arm $a \in \mathcal{A}$ represents an item and is associated with a feature vector $\boldsymbol{x}_a \in \mathbb{R}^d$. We denote $[T] = \{1, 2, \dots, T\}$. At each round $t \in [T]$, the agent is given a subset of arms $\mathcal{A}_t \subseteq \mathcal{A}$. The agent then selects an arm $a_t \in \mathcal{A}_t$ and receives a reward $r_{a_t,t}$. The reward is assumed to be linearly related to the preference vector and the feature vector of the arm, i.e., $r_{a_t,t} = \boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}^* + \eta_t$, where η_t is a random noise term.

Let a_t^* be the optimal arm at round t, i.e., $a_t^* = \arg\max_{a \in \mathcal{A}_t} \boldsymbol{x}_a^\top \boldsymbol{\theta}^*$. The agent's objective is to minimize the cumulative regret, which is defined as the total difference between the rewards of the optimal arms and the rewards obtained by the agent, i.e.,

$$\mathrm{R}(T) = \sum_{t=1}^{T} \left(\boldsymbol{x}_{a_t^*}^\top \boldsymbol{\theta}^* - \boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}^* \right).$$

Beyond observing the user's preference information through arm recommendations, the agent can gather additional feedback by launching conversations involving key terms. Specifically, a "key term" represents a category or keyword associated with a subset of arms. For example, in movie recommendations, key terms might include genres like "comedy" or "thriller", and themes such as "romance" or "sci-fi". Let $\mathcal{K}$ denote the finite set of key terms, where each key term $k \in \mathcal{K}$ corresponds to a context vector $\tilde{\boldsymbol{x}}_k \in \mathbb{R}^d$. At round t, if a conversation is initiated, the agent selects a key term $k \in \mathcal{K}$, queries the user, and receives key-term level feedback $\tilde{r}_{k,t}$. We follow the formulation of Wang et al. [25] that the user's preference vector $\boldsymbol{\theta}^*$ remains consistent across both arms and key terms. The relationship between key terms and the user's preference is also linear, i.e., $\tilde{r}_{k,t} = \tilde{\boldsymbol{x}}_k^\top \boldsymbol{\theta}^* + \tilde{\eta}_t$, where $\tilde{\eta}_t$ is a random noise term.

We list and explain our assumptions as follows. Both Assumptions 1 and 2 are consistent with previous works on conversational contextual bandits [25, 29] and linear contextual bandits [1, 15].

Assumption 1. We assume that the feature vectors for both arms and key terms are normalized, i.e., $\|\boldsymbol{x}_a\|_2 = 1$ and $\|\tilde{\boldsymbol{x}}_k\|_2 = 1$ for all $a \in \mathcal{A}$ and $k \in \mathcal{K}$. We also assume the unknown preference vector $\boldsymbol{\theta}^*$ is bounded, i.e., $\|\boldsymbol{\theta}^*\|_2 \leq 1$.

Assumption 2. We assume the noise terms $\eta_t, \tilde{\eta}_t$ are conditionally independent and 1-sub-Gaussian across T rounds.

3 Algorithm Design

In this section, we introduce our proposed algorithms, outlining their key components and implementation details.

3.1 CLiSK Algorithm

To enhance the exploration of users' preferences, we introduce the *smoothed key term contexts* and propose the CLiSK algorithm, detailed in Algorithm 1. The algorithm consists of two main modules: key term selection (Lines 4 to 10) and arm selection (Lines 11 to 16). Specifically, in each round t, the agent first determines whether to initiate a conversation based on a predefined query budget (Lines 2 and 3). If a conversation is initiated, the agent selects a key term k (Line 5) and queries the user about it. Subsequently, the agent updates its estimate of the preference vector $\boldsymbol{\theta}_t$ (Line 11) and selects an arm a_t for recommendation (Line 12). The strategies for key term selection and arm selection are elaborated as follows.

Algorithm 1: CLiSK

```
   Input: $\mathcal{A}, \mathcal{K}, b(t), \lambda, \{\alpha_t\}_{t>0}$
   Initialization: $\boldsymbol{M}_1 = \lambda \boldsymbol{I}_d, \boldsymbol{b}_1 = \boldsymbol{0}_d$
1  for $t = 1, \dots, T$ do
2      $q_t = \lfloor b(t) \rfloor - \lfloor b(t-1) \rfloor$
3      while $q_t > 0$ do
4          Smooth the key term contexts to get $\{\tilde{\tilde{\boldsymbol{x}}}_k\}_{k \in \mathcal{K}}$
5          Select a key term $k = \arg\max_{k \in \mathcal{K}} \tilde{\tilde{\boldsymbol{x}}}_k^\top \boldsymbol{\theta}_t$
6          Query the user's feedback for $k$
7          Receive the key term-level feedback $\tilde{r}_{k,t}$
8          $\boldsymbol{M}_t = \boldsymbol{M}_t + \tilde{\tilde{\boldsymbol{x}}}_{k,t} \tilde{\tilde{\boldsymbol{x}}}_{k,t}^\top$
9          $\boldsymbol{b}_t = \boldsymbol{b}_t + \tilde{r}_{k,t} \tilde{\tilde{\boldsymbol{x}}}_{k,t}$
10         $q_t = q_t - 1$
11     $\boldsymbol{\theta}_t = \boldsymbol{M}_t^{-1} \boldsymbol{b}_t$
12     Select $a_t = \arg\max_{a \in \mathcal{A}_t} \boldsymbol{x}_a^\top \boldsymbol{\theta}_t + \alpha_t \|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}$
13     Ask the user's preference for arm $a_t$
14     Observe the reward $r_{a_t,t}$
15     $\boldsymbol{M}_{t+1} = \boldsymbol{M}_t + \boldsymbol{x}_{a_t} \boldsymbol{x}_{a_t}^\top$
16     $\boldsymbol{b}_{t+1} = \boldsymbol{b}_t + r_{a_t,t} \boldsymbol{x}_{a_t}$
```

3.1.1 Intuition Overview. Building on insights from Kannan et al. [13] and Raghavan et al. [20], we add small perturbations to the key term contexts to deepen the exploration of users' preferences. These perturbations increase data diversity and help uncover preferences that might be overlooked when selecting key terms directly. For instance, instead of using "comedy" alone, variations like "romantic comedy" or "dark comedy" can reveal more specific preferences.

Below is the formal definition of smoothed key term contexts, where the perturbations are modeled as Gaussian noise.

Definition 1 (Smoothed Key Term Contexts)**.** Given a key term set $\mathcal{K}$, the smoothed key term contexts are defined as $\{\tilde{\tilde{\boldsymbol{x}}}_k\}_{k\in\mathcal{K}}$, where $\tilde{\tilde{\boldsymbol{x}}}_k = \tilde{\boldsymbol{x}}_k + \boldsymbol{\varepsilon}_k$ for each $k \in \mathcal{K}$. The noise vector $\boldsymbol{\varepsilon}_k$ is independently drawn from a truncated multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \rho^2 \cdot \boldsymbol{I}_d)$, where $\boldsymbol{I}_d$ is the d-dimensional identity matrix and ρ^2 controls the level of perturbations. Each dimension of $\boldsymbol{\varepsilon}_k$ is truncated within $[-R, R]$ for some $R > 0$, i.e., $|(\boldsymbol{\varepsilon}_k)_j| \le R, \forall j \in [d]$.

3.1.2 Key Term Selection. When initiating conversations, the agent no longer selects key terms directly based on their original contexts. Instead, the agent applies a small random perturbation to each key term's context, as defined in Definition 1 (Line 4). It then greedily selects the key term with the highest value under the perturbed contexts, i.e., $k = \arg\max_{k\in\mathcal{K}} \tilde{\tilde{\boldsymbol{x}}}_k^\top \boldsymbol{\theta}_t$ (Line 5).

Remark 1. Note that the smoothed key term contexts are re-generated for *each conversation*. For notational consistency, we use the same notation $\{\tilde{\tilde{\boldsymbol{x}}}_k\}_{k\in\mathcal{K}}$ to represent the smoothed key term contexts across different conversations.

3.1.3 Conversation Frequency. Following Zhang et al. [29], CLiSK uses a deterministic function $b(t)$ to regulate the frequency of conversation initiation. The function $b(t)$ is monotonically increasing regarding t and satisfies $b(0) = 0$. At round t, the agent initiates $q(t) = \lfloor b(t) \rfloor - \lfloor b(t-1) \rfloor$ conversations if $q(t) > 0$; otherwise, no conversation is conducted.

3.1.4 Arm Selection. CLiSK uses the Upper Confidence Bound (UCB) strategy for arm selection, a prevalent method in linear bandits. At round t, the agent updates its estimated preference vector $\boldsymbol{\theta}_t$ based on both arm-level and key term-level feedback. This estimation follows a ridge regression framework with regularization parameter λ, i.e., $\boldsymbol{\theta}_t = \boldsymbol{M}_t^{-1}\boldsymbol{b}_t$, with $\boldsymbol{M}_t$ and $\boldsymbol{b}_t$ defined as

$$\boldsymbol{M}_t = \sum_{s=1}^{t-1} \boldsymbol{x}_{a_s}\boldsymbol{x}_{a_s}^\top + \sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s} \tilde{\tilde{\boldsymbol{x}}}_k \tilde{\tilde{\boldsymbol{x}}}_k^\top + \lambda \boldsymbol{I}_d,$$

$$\boldsymbol{b}_t = \sum_{s=1}^{t-1} r_{a_s,s}\boldsymbol{x}_{a_s} + \sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s} \tilde{r}_{k,s}\tilde{\tilde{\boldsymbol{x}}}_k,$$

where $\mathcal{K}_s$ is the set of key terms selected at round s. $\boldsymbol{M}_t$ is commonly referred to as the covariance matrix.

After the update, the agent selects the arm with the highest UCB value, i.e., $a_t = \arg\max_{a\in\mathcal{A}_t} \boldsymbol{x}_a^\top \boldsymbol{\theta}_t + \alpha_t \|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}$, where $\|\boldsymbol{x}\|_{\boldsymbol{M}}$ denotes the Mahalanobis norm $\sqrt{\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}}$ and $\{\alpha_t\}_{t>0}$ are parameters designed to balance the exploration-exploitation trade-off.

3.2 CLiME Algorithm

To enable more adaptive and flexible conversation initiation, we introduce the CLiME algorithm, detailed in Algorithm 2. The CLiME adopts the same arm selection strategy as CLiSK, but it introduces key innovations in determining when to initiate conversations and which key terms to select. Unlike CLiSK, which follows a deterministic function $b(t)$ for scheduling conversations, CLiME adaptively determines when to conduct a conversation based on the uncertainty in the preference estimation.

Algorithm 2: CLiME

```
   Input: $\mathcal{A}, \mathcal{K}, \lambda, \alpha, \{\alpha_t\}_{t>0}$
   Initialization: $\boldsymbol{M}_1 = \lambda \boldsymbol{I}_d, \boldsymbol{b}_1 = \boldsymbol{0}_d$
1  for $t = 1, \dots, T$ do
2      if UncertaintyChecking($t$) then
3          Diagonalize $\boldsymbol{M}_t = \sum_{i=1}^{d} \lambda_{\boldsymbol{v}_i}\boldsymbol{v}_i\boldsymbol{v}_i^\top$
4          foreach $\lambda_{\boldsymbol{v}_i} < \alpha t$ do
5              $k = \arg\max_{k\in\mathcal{K}} |\tilde{\boldsymbol{x}}_k^\top \boldsymbol{v}_i|$
6              $n_k = \lceil(\alpha t - \lambda_{\boldsymbol{v}_i})/c_0^2\rceil$
7              Schedule $n_k$ conversations about the key term $k$
                 before next uncertainty checking
8              Update $\boldsymbol{M}_t$ and $\boldsymbol{b}_t$ accordingly
9      $\boldsymbol{\theta}_t = \boldsymbol{M}_t^{-1}\boldsymbol{b}_t$
10     Select $a_t = \arg\max_{a\in\mathcal{A}_t} \boldsymbol{x}_a^\top\boldsymbol{\theta}_t + \alpha_t\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}$
11     Ask the user's preference for arm $a_t$
12     Observe the reward $r_{a_t,t}$
13     $\boldsymbol{M}_{t+1} = \boldsymbol{M}_t + \boldsymbol{x}_{a_t}\boldsymbol{x}_{a_t}^\top$
14     $\boldsymbol{b}_{t+1} = \boldsymbol{b}_t + r_{a_t,t}\boldsymbol{x}_{a_t}$
```

3.2.1 Intuition Overview. The main idea behind CLiME is to adaptively initiate conversations based on the current level of uncertainty in the estimated preference and use key terms to explore the uncertain directions effectively. Specifically, the covariance matrix $\boldsymbol{M}_t$ encodes information about the feature space, where its eigenvectors represent the principal directions within the space, and the corresponding eigenvalues indicate the level of uncertainty along these directions. A smaller eigenvalue indicates a higher uncertainty in the associated direction. Therefore, by guiding the agent to explore such high-uncertainty directions, the agent can reduce uncertainty and improve learning efficiency. If the minimum eigenvalue of $\boldsymbol{M}_t$ remains above a certain value, the agent ensures sufficient exploration of the feature space. To facilitate exploration, we introduce the following assumption.

Assumption 3. We assume that the elements in the key term set $\mathcal{K}$ are sufficiently rich and diverse, such that for any $\boldsymbol{x} \in \mathbb{R}^d$ satisfying $\|\boldsymbol{x}\|_2 = 1$, there exists a key term $k \in \mathcal{K}$ such that $|\tilde{\boldsymbol{x}}_k^\top \boldsymbol{x}| \ge c_0$, where c_0 is some constant close to 1.

This mild assumption ensures that the key term set $\mathcal{K}$ is comprehensive enough to cover all relevant directions in the feature space. In other words, for any direction $\boldsymbol{x}$ that the agent might need to explore, there exists a key term $k \in \mathcal{K}$ whose context $\tilde{\boldsymbol{x}}_k$ aligns sufficiently well with $\boldsymbol{x}$. This diversity allows the agent to effectively reduce uncertainty by exploring underrepresented directions, thereby improving preference learning.

3.2.2 Conversation Initiation and Key Term Selection. In CLiME, conversation initiation and key term selection are designed to maximize the information gained from user interactions. As shown in Algorithm 2, the agent first evaluates the eigenvalues of the covariance matrix $\boldsymbol{M}_t$ (Line 3). If any eigenvalue $\lambda_{\boldsymbol{v}_i}$ falls below a certain threshold (derived from Section 4.2), i.e., $\lambda_{\boldsymbol{v}_i} < \alpha t$ (Line 4), the agent prompts $n_k = \lceil(\alpha t - \lambda_{\boldsymbol{v}_i})/c_0^2\rceil$ conversations by selecting key terms that most closely align with the corresponding eigenvector $\boldsymbol{v}_i$ (Lines 5 to 7). Here, $\alpha \in (0, c_0^2)$ is an exploration control parameter that regulates the exploration level. Note that the agent can distribute these n_k conversations across multiple rounds before re-evaluating the eigenvalues of the covariance matrix.

To further enhance flexibility and accommodate diverse real-world applications, we design an uncertainty checking function `UncertaintyChecking(t)` (Line 2). This function determines when to assess uncertainty and potentially trigger conversations. Examples of such checking functions are given as follows.

- **Continuous Checking**: The agent assesses uncertainty at every round and initiates conversations as needed.
- **Fixed Interval Checking**: The agent assesses uncertainty every P rounds, where P is a fixed integer.
- **Exponential Phase Checking**: The agent evaluates uncertainty at exponentially increasing intervals of 2^i, where $i = 1, 2, \ldots$.

Remark 2. The uncertainty checking functions in CLiME differ fundamentally from the frequency function $b(t)$ in ConUCB [29]. Specifically, these checking functions regulate how often uncertainty is assessed but do not directly dictate conversation initiation. In contrast, $b(t)$ deterministically controls both the timing and number of conversations. CLiME and ConUCB also differ in how they select key terms, further distinguishing the two approaches.

Remark 3. It is worth noting that the smoothed key term contexts approach in CLiSK and the adaptive conversation technique in CLiME are orthogonal. The two strategies can operate independently or be integrated to enhance learning efficiency further. To this end, we introduce the CLiSK-ME algorithm, detailed in Appendix A.1, which integrates both approaches to leverage their complementary strengths.

4 Theoretical Analysis

This section presents the theoretical results of our algorithms, which employ analytical techniques that differ from standard linear bandit methods. Detailed proofs of all lemmas and theorems are provided in the Appendices.

4.1 Regret Analysis of CLiSK Algorithm

Following Zhang et al. [29] and Wang et al. [25], we assume $b(t) = bt$ for some $b \in (0, 1)$. We start with Lemma 1, which bounds the difference between the estimated and true rewards for each arm.

Lemma 1. *Under Assumptions 1 and 2, for CLiSK, for any round $t \in [T]$ and any arm $a \in \mathcal{A}$, with probability at least $1 - \delta$ for some $\delta \in (0, 1)$, we have*

$$\left|\boldsymbol{x}_a^\top \boldsymbol{\theta}_t - \boldsymbol{x}_a^\top \boldsymbol{\theta}^*\right| \leq \alpha_t \|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}},$$

where $\alpha_t = \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{t + \left(1 + \sqrt{d}R\right)bt}{\lambda d}\right)} + \sqrt{\lambda}.$

Next, we examine the smoothed key term contexts and their impact on exploring the feature space.

Lemma 2. *For any round $t \in [T]$, with the smoothed key term contexts in Definition 1, CLiSK has the following lower bound on the minimum eigenvalue of the matrix $\mathbb{E}[\tilde{\tilde{\boldsymbol{x}}}_k \tilde{\tilde{\boldsymbol{x}}}_k^\top]$ for any $k \in \mathcal{K}_t$, i.e.,*

$$\lambda_{\min}\left(\mathbb{E}[\tilde{\tilde{\boldsymbol{x}}}_k \tilde{\tilde{\boldsymbol{x}}}_k^\top]\right) \geq c_1 \frac{\rho^2}{\log|\mathcal{K}|} \triangleq \lambda_{\mathcal{K}},$$

where $c_1 \in (0, 1)$ is some constant.

Lemma 2 provides a lower bound on the minimum eigenvalue of the expected outer product of the selected key term. Intuitively, this implies that under smoothed contexts, the selected key terms exhibit sufficient diversity in the feature space, ensuring that each query contributes meaningful information about the user's preferences.

Lemma 3. *For CLiSK, with probability at least $1 - \delta$ for some $\delta \in (0, 1)$, if $t \geq T_0 \triangleq \frac{8(1+\sqrt{d}R)^2}{b\lambda_{\mathcal{K}}} \log\left(\frac{d}{\delta}\right)$, we have*

$$\lambda_{\min}\left(\sum_{s=1}^{t} \sum_{k \in \mathcal{K}_s} \tilde{\tilde{\boldsymbol{x}}}_k \tilde{\tilde{\boldsymbol{x}}}_k^\top\right) \geq \frac{\lambda_{\mathcal{K}} bt}{2}.$$

Lemma 3 establishes a lower bound on the minimum eigenvalue of the Gram matrix that grows linearly with time t. This guarantees that CLiSK accumulates enough statistical information to effectively estimate the user's preference vector through ridge regression. Following these results, we bound $\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}$ in Lemma 4 and derive a high-probability regret upper bound for CLiSK in Theorem 1.

Lemma 4. *For CLiSK, for any $a \in \mathcal{A}$, if $t \geq T_0 \triangleq \frac{8(1+\sqrt{d}R)^2}{b\lambda_{\mathcal{K}}} \log\left(\frac{d}{\delta}\right)$, with probability at least $1 - \delta$ for some $\delta \in (0, 1)$, $\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}} \leq \sqrt{\frac{2}{\lambda_{\mathcal{K}} bt}}$.*

Theorem 1 (Regret of CLiSK). *With probability at least $1 - \delta$ for some $\delta \in (0, 1)$, the regret upper bound of CLiSK satisfies*

$$\begin{aligned}
\mathrm{R}(T) \leq{} & \frac{8(1 + \sqrt{d}R)^2 \log(|\mathcal{K}|)}{c_1 \rho^2 b} \log\left(\frac{d}{\delta}\right) + 4\sqrt{\frac{2c_1\rho^2 T}{b\log(|\mathcal{K}|)}} \cdot \\
& \left(\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{T + \left(1 + \sqrt{d}R\right)bT}{\lambda d}\right)} + \sqrt{\lambda}\right) \\
={} & O(\sqrt{dT\log(T)} + d),
\end{aligned}$$

where R and ρ^2 are constants in Definition 1.

4.2 Regret Analysis of CLiME Algorithm

We begin with Lemma 5, which closely parallels Lemma 1.

Lemma 5. *Let $\boldsymbol{\theta}_t$ be the estimated preference vector at round t and $\boldsymbol{\theta}^*$ be the true preference vector. Under Assumptions 1, 2 and 3, for CLiME, at round t, for any arm $a \in \mathcal{A}$, with probability at least $1 - \delta$ ($\delta \in (0, 1)$), we have*

$$\left|\boldsymbol{x}_a^\top \boldsymbol{\theta}_t - \boldsymbol{x}_a^\top \boldsymbol{\theta}^*\right| \leq \alpha_t \|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}},$$

where $\alpha_t = \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{t + \alpha dt}{\lambda d c_0^2}\right)} + \sqrt{\lambda}$, *$\alpha$ is an exploration control factor in Algorithm 2, and c_0 is a constant in Assumption 3.*

Since conversations are initiated adaptively in CLiME, the number of conversations conducted up to each round t is not deterministic. A key challenge to prove Lemma 5 is to bound this quantity. Then, we present Lemma 6, which bounds $\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}$.

Lemma 6. *For CLiME, for any arm $a \in \mathcal{A}$, with probability at least $1-\delta$ for some $\delta \in (0,1)$, at round $t \geq 2P$, we have $\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}} \leq \sqrt{\frac{2}{\alpha t}}$, where P is a fixed integer.*

The proof of Lemma 6 relies on establishing a lower bound on the minimum eigenvalue of $\boldsymbol{M}_t$, i.e., $\lambda_{\min}(\boldsymbol{M}_t) \geq \alpha t$, which involves a delicate analysis of covariance matrix eigenvalues. The condition $t \geq 2P$ is introduced to generalize all three checking functions. Building on this, we derive the following theorem for CLiME.

Theorem 2 (Regret of CLiME)**.** *With probability at least $1-\delta$ for some $\delta \in (0,1)$, the regret upper bound of CLiME satisfies*

$$\begin{aligned} \mathrm{R}(T) &\leq 4\sqrt{\frac{2T}{\alpha}}\left(\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{T + \alpha d T}{\lambda d c_0^2}\right)} + \sqrt{\lambda}\right) + 2P \\ &= O\left(\sqrt{dT\log(T)}\right). \end{aligned}$$

Remark 4. Note that Theorem 2 applies to all three uncertainty checking functions discussed in CLiME algorithm, which underscores the generality of our methods.

CLiSK-ME combines the advantages of both smoothed key term contexts and adaptive conversation techniques, ensuring efficient exploration while adaptively adjusting conversation frequency based on uncertainty. As a result, we derive the following corollary.

Corollary 1. *With probability at least $1-\delta$ for some $\delta \in (0,1)$, the regret upper bound of CLiSK-ME satisfies* $\mathrm{R}(T) = O(\sqrt{dT\log(T)})$.

4.3 Lower Bound for Conversational Bandits

We establish a regret lower bound for conversational bandits with *finite* and *time-varying* arm sets. Our result is novel because the well-known lower bound $\Omega(\sqrt{dT})$ by Chu et al. [6] does not consider conversational information and thus cannot be directly applied to our setting. Additionally, the existing lower bound for federated conversational bandits [19] is also inapplicable, as it assumes a *fixed* arm set. The detailed proof is given in Appendix A.11.

Theorem 3 (Regret lower bound)**.** *For any policy that chooses at most one key term per time step, there exists an instance of the conversational bandit problem such that the expected regret is at least $\Omega(\sqrt{dT})$. Furthermore, for any $T = 2^m$ with $m \in [d]$, the regret is at least $\Omega(\sqrt{dT\log(T)})$.*

4.4 Discussion on Optimality

To the best of our knowledge, we are the first to propose algorithms for conversational contextual bandits that achieve the *optimal* regret bound of order $\widetilde{O}(\sqrt{dT})$. We summarize the regret bounds of our proposed algorithms and related algorithms in Table 1 and discuss the theoretical improvements over existing methods.

Table 1: Comparison of theoretical regret bounds.

Algorithm	Conversational	Regret
LinUCB [1]	✗	$O(d\sqrt{T}\log T)$
ConUCB [29], ConLinUCB-MCR [25]	✓	$O(d\sqrt{T}\log T)$
ConLinUCB-BS [25]	✓	At least $O(d\sqrt{T\log T})^*$
CLiSK (Ours, Theorem 1)	✓	$O(\sqrt{dT\log T} + d)$
CLiME (Ours, Theorem 2)	✓	$O(\sqrt{dT\log T})$
CLiSK-ME (Ours, Corollary 1)	✓	$O(\sqrt{dT\log T})$

* The original paper claims a regret of $O(\sqrt{dT\log T})$ but its analysis is flawed.

The regret upper bound of LinUCB [1] is $O(d\sqrt{T}\log T)$, which serves as a standard benchmark in contextual linear bandits. The first algorithm for conversational bandits, ConUCB [29], offers the same regret upper bound as LinUCB, indicating that it does not offer a substantial theoretical improvement over the non-conversational algorithms. Since then, improving regret through conversational interactions has remained an open problem in the field. Under the assumption that the key term set $\mathcal{K}$ spans $\mathbb{R}^d$, ConLinUCB-BS [25] achieves a regret upper bound of $O(\frac{1}{\sqrt{\lambda_{\mathcal{B}}}}\sqrt{dT\log T})$, where $\lambda_{\mathcal{B}} := \lambda_{\min}\left(\mathbb{E}_{k\in\mathrm{unif}(\mathcal{B})}\left[\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top\right]\right)$ and $\mathcal{B}$ is the *barycentric spanner* of $\mathcal{K}$. The authors assume $\lambda_{\mathcal{B}}$ is a constant, leading to a regret bound of $O(\sqrt{dT\log T})$. However, this assumption is incorrect as $\lambda_{\mathcal{B}}$ depends on the dimension d and is not a constant. Specifically, denoting $\boldsymbol{X} := \mathbb{E}_{k\in\mathrm{unif}(\mathcal{B})}\left[\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top\right]$ and $\{\lambda_i\}_{i=1}^d$ as its eigenvalues, we use the fact that $\|\tilde{\boldsymbol{x}}_k\| = 1$ and obtain $\mathrm{Tr}\,(\boldsymbol{X}) = \mathbb{E}_{k\in\mathrm{unif}(\mathcal{B})}\left[\mathrm{Tr}\left(\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top\right)\right] = 1$, thus $\lambda_{\mathcal{B}} \leq \frac{\sum_{i=1}^d \lambda_i}{d} = \frac{\mathrm{Tr}\,(\boldsymbol{X})}{d} = \frac{1}{d}$. Consequently, by plugging this result back into the regret expression, the regret bound of ConLinUCB-BS cannot be better than $O(d\sqrt{T\log T})$. These previous attempts underscore the significance of our work. In contrast, with the smoothed key term context technique and with the adaptive conversation technique, our algorithms achieve a better regret bound of $O(\sqrt{dT\log T} + d)$ and $O(\sqrt{dT\log T})$, respectively. These improvements successfully match the lower bound (Theorem 3) up to logarithmic factors in their dependence on the time horizon T.

5 Evaluation

In this section, we evaluate the performance of our algorithms on both synthetic and real-world datasets. All the experiments were conducted on a machine equipped with a 3.70 GHz Intel Xeon E5-1630 v4 CPU and 32GB RAM.

5.1 Experiment Setups

5.1.1 Datasets. Consistent with existing studies, we generate a synthetic dataset and use three real-world datasets: MovieLens-25M [12], Last.fm [4], and Yelp[1].

For the synthetic dataset, we set the dimension $d = 50$, the number of users $N = 200$, the number of arms $|\mathcal{A}| = 5{,}000$, and the number of key terms $|\mathcal{K}| = 1{,}000$. We generate it following Zhang et al. [29]. First, for each key term $k \in \mathcal{K}$, we sample a pseudo feature vector $\dot{\boldsymbol{x}}_k$ with each dimension drawn from a uniform distribution $\mathcal{U}(-1,1)$. For each arm $i \in \mathcal{A}$, we randomly select an integer $n_i \in \{1,2,\ldots,5\}$ and uniformly sample a subset of key terms $\mathcal{K}_i \subset \mathcal{K}$ with $|\mathcal{K}_i| = n_i$. The weight is defined as $w_{i,k} = 1/n_i$ for each $k \in \mathcal{K}_i$. For each arm i, the feature vector $\boldsymbol{x}_i$ is drawn from a multivariate Gaussian $\mathcal{N}(\sum_{j\in\mathcal{K}_i}\dot{\boldsymbol{x}}_j/n_i, \boldsymbol{I})$. The feature vector for each key term k, denoted by $\tilde{\boldsymbol{x}}_k$, is computed as $\tilde{\boldsymbol{x}}_k = \sum_{i\in\mathcal{A}}\frac{w_{i,k}}{\sum_{j\in\mathcal{A}}w_{j,k}}\boldsymbol{x}_i$. Finally, each user's preference vector

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

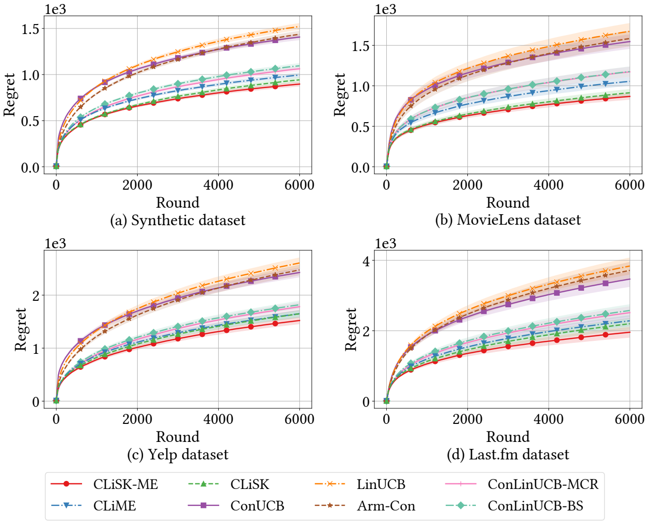
$\boldsymbol{\theta}_u \in \mathbb{R}^d$ is generated by sampling each dimension from $\mathcal{U}(-1, 1)$ and normalizing it to unit length.

For the real-world datasets, we regard movies/artists/businesses as arms. To exclude unrepresentative or insufficiently informative data (such as users who have not submitted any reviews or movies with only a few reviews), we extract a subset of $|\mathcal{A}| = 5,000$ arms with the highest number of user-assigned ratings/tags, and a subset of $N = 200$ users who have assigned the most ratings/tags. Key terms are identified by using the associated movie genres, business categories, or tag IDs in the MovieLens, Yelp, and Last.fm datasets, respectively. For example, each movie is associated with a list of genres, such as “action” or “comedy”, and each business (e.g., restaurant) is categorized by terms such as “Mexican” or “Burgers”. Using the data extracted above, we create a *feedback matrix* $\boldsymbol{R}$ of size $N \times |\mathcal{A}|$, where each element $\boldsymbol{R}_{i,j}$ represents the user i's feedback to arm j. We assume that the user's feedback is binary. For the MovieLens and Yelp datasets, a user's feedback for a movie/business is 1 if the user's rating is higher than 3; otherwise, the feedback is 0. For the Last.fm dataset, a user's feedback for an artist is 1 if the user assigns a tag to the artist. Next, we generate the feature vectors for arms $\boldsymbol{x}_i$ and the preference vectors for users $\boldsymbol{\theta}_u$. Following existing works, we decompose the feedback matrix $\boldsymbol{R}$ using truncated Singular Value Decomposition (SVD) as $\boldsymbol{R} \approx \boldsymbol{\Theta} \boldsymbol{S} \boldsymbol{A}^\top$, where $\boldsymbol{\Theta} \in \mathbb{R}^{N \times d}$ and $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{A}| \times d}$ contain the top-d left and right singular vectors, and $\boldsymbol{S} \in \mathbb{R}^{d \times d}$ is a diagonal matrix with the corresponding top-d singular values. Then each $\boldsymbol{\theta}_u^\top$ corresponds to the u-th row of $\boldsymbol{\Theta S}$ for all $u \in [N]$, and each $\boldsymbol{x}_i^\top$ corresponds to the i-th row of $\boldsymbol{A}$ for all $i \in \mathcal{A}$. The feature vectors for key terms are generated similarly to those in the synthetic dataset, by assigning equal weights for all key terms corresponding to each arm.

5.1.2 *Baseline Algorithms.* We select the following baseline algorithms from existing studies: (1) LinUCB [1]: The standard linear contextual bandit algorithm, which does not consider the conversational setting and only has arm-level feedback. (2) Arm-Con [5]: An extension of LinUCB that initiates conversations directly from arm sets. (3) ConUCB [29]: The first algorithm proposed for conversational contextual bandits that queries key terms when conversations are allowed. (4) ConLinUCB [25]: It consists of three algorithms with different key term selection strategies. ConLinUCB-BS computes the *barycentric spanner* of key terms as an exploration basis. ConLinUCB-MCR selects key terms with the largest confidence radius. ConLinUCB-UCB chooses key terms with the largest upper confidence bounds. Since ConLinUCB-BS and ConLinUCB-MCR demonstrate superior performance, we focus our comparisons on these two variants.

5.2 Evaluation Results

5.2.1 *Cumulative Regret.* First, we compare our algorithms against all baseline algorithms in terms of cumulative regret over $T = 6,000$ rounds. In each round, we randomly select $|\mathcal{A}| = 200$ arms from each dataset. For the baseline algorithms, we adopt the conversation frequency function $b(t) = 5\lfloor\log(T)\rfloor$, as specified in their original papers. We present the results for all three checking functions “Continuous”, “Fixed Interval”, and "Exponential Phase”, for both CLiME and CLiSK-ME. For the “Fixed Interval” function, `UncertaintyChecking` is triggered every 100 rounds, whereas for the "Exponential Phase” it is triggered whenever t is a power of 2. For CLiSK, both the perturbation level ρ^2 and the truncation limit R are set to 1. The results are averaged over 20 trials, and the resulting confidence intervals are included in the figures. Under the “Continuous Checking” function, as shown in Figure 2, our three algorithms consistently achieve the best performance (lowest regret) with an improvement of over 14.6% compared to the best baseline. Similar performance trends hold under the other two checking functions, as illustrated in Figures 3 and 4. These results confirm the validity of our theoretical advancements.



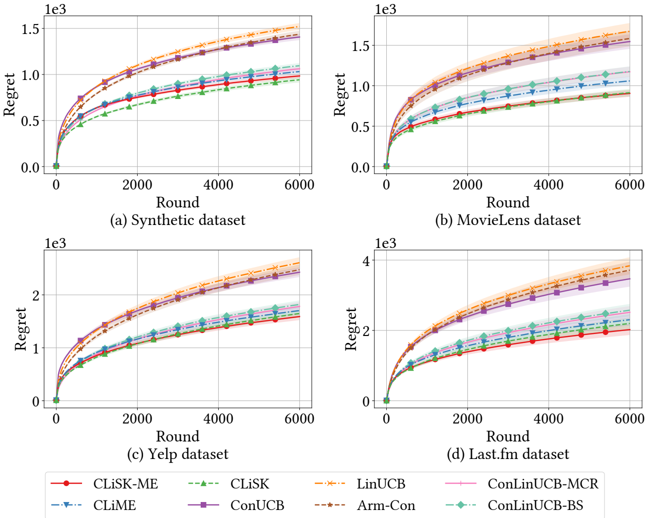
(a) Synthetic dataset

(b) MovieLens dataset

(c) Yelp dataset

(d) Last.fm dataset

Figure 2: Comparison of cumulative regret where CLiME and CLiSK-ME use the “Continuous Checking” function.



(a) Synthetic dataset

(b) MovieLens dataset

(c) Yelp dataset

(d) Last.fm dataset

Figure 3: Comparison of cumulative regret where CLiME and CLiSK-ME use the “Fixed Interval” function.

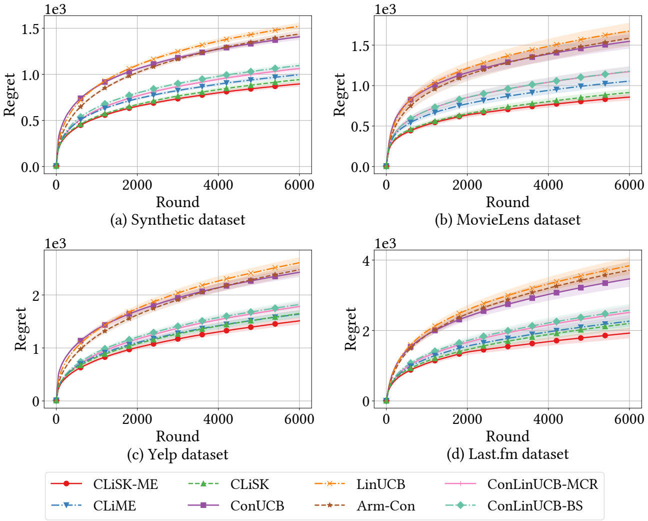


Figure 4: Comparison of cumulative regret where CLiME and CLiSK-ME use the "Exponential Phase" function.

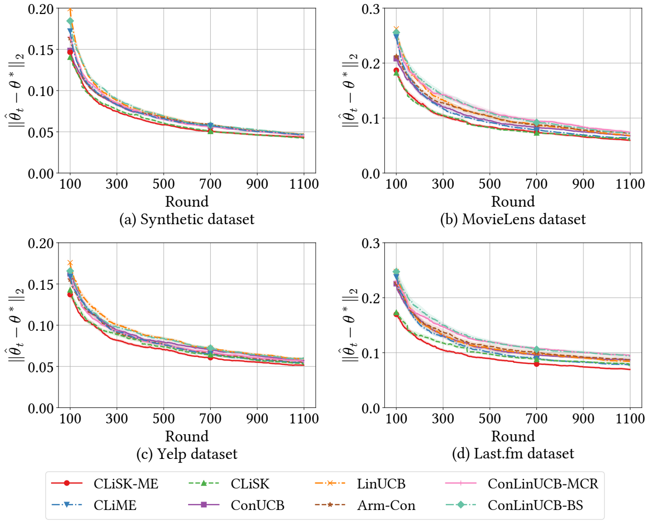


Figure 5: Comparison of estimation precision where CLiME and CLiSK-ME use the "Continuous Checking" function.

5.2.2 *Precision of Estimated Preference Vectors.*

To assess how accurately each algorithm learns the user's preferences over time, we measure the average distance between the estimated vector $\widehat{\boldsymbol{\theta}}_t$ and the ground truth $\boldsymbol{\theta}^*$ for all algorithms over 1000 rounds. We present the results for the "Continuous Checking" function of CLiME and CLiSK-ME, with results for other functions provided in Appendix A.2. As shown in Figure 5, all algorithms exhibit a decreasing estimation error over time. However, our three algorithms consistently achieve the lowest estimation error in all datasets. This is because they leverage our novel conversational mechanism to gather more informative feedback, significantly accelerating the reduction of estimation error. As a result, our algorithms estimate the user's preference vector more quickly and accurately than the baseline methods.

5.2.3 *Number of Conversations.*

Next, we evaluate the number of conversations initiated by CLiME. Since CLiSK and all baseline algorithms initiate conversations based on a deterministic function $b(t)$, their results are consistent across all datasets. Therefore, we plots the scenarios for $b(t) = 5\lfloor\log(t)\rfloor$ and $b(t) = \lfloor t/50 \rfloor$ as in prior studies. It is also important to note that although some existing studies employ a logarithmic $b(t)$ in their experiments, their theoretical results require a linear $b(t)$ to hold. In contrast to the baselines, our algorithm CLiME adaptively initiates conversations depending on the current uncertainty of user preferences, providing greater flexibility and enhancing the user experience. We plot the number of conversations initiated by CLiME with different uncertainty checking functions across 4 datasets. As shown in Figure 6, the number of conversations increases only logarithmically with the number of rounds.

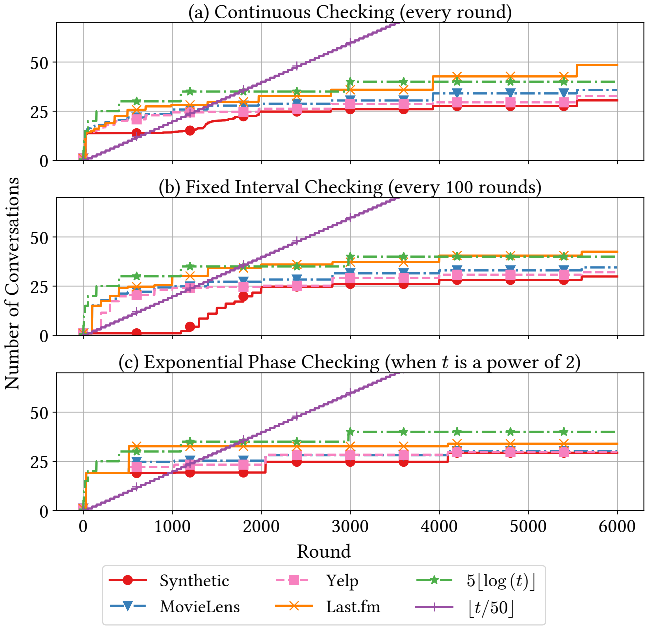


Figure 6: Number of conversations initiated by deterministic approaches and our adaptive approach CLiME with different uncertainty checking functions.

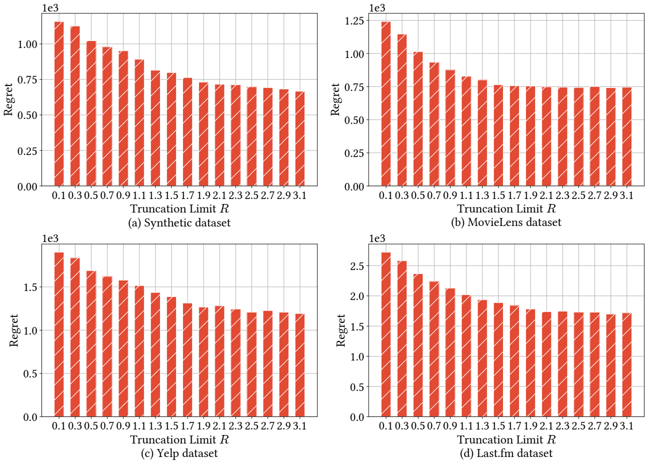
5.2.4 *Running Time.*

To evaluate the computational efficiency, we compare the running times of our algorithms with other conversational methods using the MovieLens dataset across $T = 6,000$ rounds. We separately report the total running times, as well as the times for picking arms and key terms. The results are averaged over 20 runs. As shown in Table 2, our three algorithms show substantial improvements compared to ConUCB and exhibit performance comparable to the ConLinUCB family of algorithms. For CLiME and CLiSK-ME, while matrix operations and eigenvalue computation introduce slight overhead, the algorithms remain efficient, particularly with interval and exponential checking strategies.

Table 2: Comparison of Running Times for Conversational Bandit Algorithms Using the Movielens dataset.

Algorithms		**Running Time (s)**		
		Key terms	Arms	Total
CLiSK-ME	Continuous	1.169	3.443	4.651
	Interval	0.332	3.361	3.723
	Exponential	0.352	3.344	3.724
CLiME	Continuous	0.803	3.371	4.205
	Interval	0.021	3.341	3.390
	Exponential	0.014	3.334	3.375
CLiSK		0.490	3.339	3.857
ConUCB		0.011	8.362	8.403
ConLinUCB	UCB	0.009	3.354	3.392
	MCR	0.007	3.337	3.371
	BS	0.006	3.334	3.366

5.2.5 Ablation Study. We conduct an ablation study evaluating the effect of the truncation limit R. Specifically, we analyze how different values of R affect algorithm performance by comparing the cumulative regrets at round 6,000 across all datasets, as shown in Figure 7. The results indicate that increasing R from 0.1 to 3.1 leads to a decrease in regret, with performance stabilizing when $R > 2$. For the perturbation level ρ^2, we observe that varying it from 0.1 to 3 results in no significant change in regret. Therefore, we do not include a separate figure for this parameter.



(a) Synthetic dataset

(b) MovieLens dataset

(c) Yelp dataset

(d) Last.fm dataset

Figure 7: Effect of the truncation limit R.

6 Related Work

Our research is closely aligned with studies on conversational contextual bandits, particularly focusing on the problem of key term selection within this framework.

Contextual bandits serve as a fundamental framework for online sequential decision-making problems, covering applications like recommender systems [6, 15] and computer networking [10]. Contextual bandit algorithms aim to maximize the cumulative reward in the long run while making the trade-off between exploitation and exploration. Prominent algorithms include LinUCB [1] and Thompson Sampling (TS) [2].

To address the cold start problem, conversational recommender systems (CRSs) [5, 23, 30] are proposed to engage users in conversations to learn their preferences more effectively. Zhang et al. [29] extend the standard contextual bandits to model conversational interactions, and the pioneering ConUCB algorithm with a regret upper bound $O(d\sqrt{T}\log T)$. Following the foundational work of Zhang et al. [29], a branch of research has advanced this field. Li et al. [16] design the first TS-type algorithm ConTS. Wu et al. [26] propose a clustering-based algorithm to automatically generate key terms. Zuo et al. [32] propose Hier-UCB and Hier-LinUCB, leveraging the hierarchical structures between key terms and items. Xie et al. [27] introduce a comparison-based conversation framework and propose RelativeConUCB. Zhao et al. [31] integrate knowledge graphs into conversational bandits. Li et al. [19] investigate federated conversational bandits. Dai et al. [7, 8] study the conversational bandits with misspecified/corrupted models. To enhance learning efficiency, Dai et al. [9] consider multi-agent LLM response identification with a fixed arm set. Wang et al. [25] and Yang et al. [28] investigate the key term selection strategies and propose the ConLinUCB-BS and ConDuel algorithms, respectively. Both algorithms uniformly select key terms from the barycentric spanner of the key term set.

The smoothed analysis for contextual bandits has been widely studied recently [13, 17, 18, 20–22]. The smoothed setting bridges i.i.d. distributional and adversarial contexts. Kannan et al. [13] first introduce the smoothed analysis for linear contextual bandits, showing that small perturbations can lead to sublinear regret with a greedy algorithm. Raghavan et al. [21] and Raghavan et al. [20] show that the greedy algorithm achieves the best possible Bayesian regret in this setting. Sivakumar et al. [22] extend the smoothed analysis to structured linear bandits. Building on these insights, we apply the smoothed key term contexts in conversational contextual bandits.

7 Conclusion

In this paper, we studied key term selection strategies for conversational contextual bandits and introduced three novel algorithms: CLiSK, CLiME, and CLiSK-ME. CLiSK leverages smoothed key term contexts to enhance exploration, while CLiME adaptively initiates conversations with key terms that minimize uncertainty in the feature space. CLiSK-ME integrates both techniques, further improving learning efficiency. We proved that all three algorithms achieve tighter regret bounds than prior studies. Extensive evaluations showed that our algorithms outperform other conversational bandit algorithms.

Acknowledgments

The work of John C.S. Lui was supported in part by the RGC GRF-14202923.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* 24 (2011).

[2] Shipra Agrawal and Navin Goyal. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory.* JMLR Workshop and Conference Proceedings, 39–1.

[3] J. Bretagnolle and C. Huber. 1978. Estimation des densités : Risque minimax. In *Séminaire de Probabilités XII*, C. Dellacherie, P. A. Meyer, and M. Weil (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 342–363.

[4] Ivan Cantador, Peter Brusilovsky, and Tsvi Kuflik. 2011. Second Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec2011). In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11).* 387–388.

[5] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* 815–824.

[6] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. 2011. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics.* JMLR Workshop and Conference Proceedings, 208–214.

[7] Xiangxiang Dai, Zhiyong Wang, Jize Xie, Xutong Liu, and John CS Lui. 2024. Conversational Recommendation with Online Learning and Clustering on Misspecified Users. *IEEE Transactions on Knowledge and Data Engineering* 36, 12 (2024), 7825–7838.

[8] Xiangxiang Dai, Zhiyong Wang, Jize Xie, Tong Yu, and John CS Lui. 2024. Online Learning and Detecting Corrupted Users for Conversational Recommendation Systems. *IEEE Transactions on Knowledge and Data Engineering* 36, 12 (2024), 8939–8953.

[9] Xiangxiang Dai, Yuejin Xie, Maoli Liu, Xuchuang Wang, Zhuohua Li, Huanyu Wang, and John C. S. Lui. 2025. Multi-Agent Conversational Online Learning for Adaptive LLM Response Identification. arXiv:2501.01849 [cs.HC]

[10] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. 2012. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking* 20, 5 (2012), 1466–1478.

[11] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI open* 2 (2021), 100–126.

[12] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (dec 2015), 19 pages.

[13] Sampath Kannan, Jamie H Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. 2018. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in neural information processing systems* 31 (2018).

[14] Tor Lattimore and Csaba Szepesvári. 2020. *Bandit Algorithms.* Cambridge University Press.

[15] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web.* 661–670.

[16] Shijun Li, Wenqiang Lei, Qingyun Wu, Xiangnan He, Peng Jiang, and Tat-Seng Chua. 2021. Seamlessly unifying attributes and items: Conversational recommendation for cold-start users. *ACM Transactions on Information Systems (TOIS)* 39, 4 (2021), 1–29.

[17] Zhuohua Li, Maoli Liu, Xiangxiang Dai, and John C.S. Lui. 2025. Demystifying Online Clustering of Bandits: Enhanced Exploration Under Stochastic and Smoothed Adversarial Contexts. In *The Thirteenth International Conference on Learning Representations.*

[18] Zhuohua Li, Maoli Liu, Xiangxiang Dai, and John C.S. Lui. 2025. Towards Efficient Conversational Recommendations: Expected Value of Information Meets Bandit Learning. In *Proceedings of the ACM on Web Conference 2025* (Sydney NSW, Australia) *(WWW '25).* 4226–4238.

[19] Zhuohua Li, Maoli Liu, and John C. S. Lui. 2024. FedConPE: Efficient Federated Conversational Bandits with Heterogeneous Clients. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24.* 4533–4541.

[20] Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. 2023. Greedy algorithm almost dominates in smoothed contextual bandits. *SIAM J. Comput.* 52, 2 (2023), 487–524.

[21] Manish Raghavan, Aleksandrs Slivkins, Jennifer Vaughan Wortman, and Zhiwei Steven Wu. 2018. The externalities of exploration and how data diversity helps exploitation. In *Conference on Learning Theory.* PMLR, 1724–1738.

[22] Vidyashankar Sivakumar, Steven Wu, and Arindam Banerjee. 2020. Structured linear contextual bandits: A sharp and geometric smoothed analysis. In *International Conference on Machine Learning.* PMLR, 9026–9035.

[23] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval.* 235–244.

[24] Joel A. Tropp. 2011. User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics* 12, 4 (Aug. 2011), 389–434.

[25] Zhiyong Wang, Xutong Liu, Shuai Li, and John CS Lui. 2023. Efficient explorative key-term selection strategies for conversational contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 10288–10295.

[26] Junda Wu, Canzhe Zhao, Tong Yu, Jingyang Li, and Shuai Li. 2021. Clustering of conversational bandits for user preference learning and elicitation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management.* 2129–2139.

[27] Zhihui Xie, Tong Yu, Canzhe Zhao, and Shuai Li. 2021. Comparison-based conversational recommender system with relative bandit feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1400–1409.

[28] Shuhua Yang, Hui Yuan, Xiaoying Zhang, Mengdi Wang, Hong Zhang, and Huazheng Wang. 2024. Conversational Dueling Bandits in Generalized Linear Models. *arXiv preprint arXiv:2407.18488* (2024).

[29] Xiaoying Zhang, Hong Xie, Hang Li, and John CS Lui. 2020. Conversational contextual bandit: Algorithm and application. In *Proceedings of the web conference 2020.* 662–672.

[30] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management.* 177–186.

[31] Canzhe Zhao, Tong Yu, Zhihui Xie, and Shuai Li. 2022. Knowledge-aware conversational preference elicitation with bandit feedback. In *Proceedings of the ACM Web Conference 2022.* 483–492.

[32] Jinhang Zuo, Songwen Hu, Tong Yu, Shuai Li, Handong Zhao, and Carlee Joe-Wong. 2022. Hierarchical conversational preference elicitation with bandit feedback. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management.* 2827–2836.

A Appendix

A.1 CLiSK-ME Algorithm

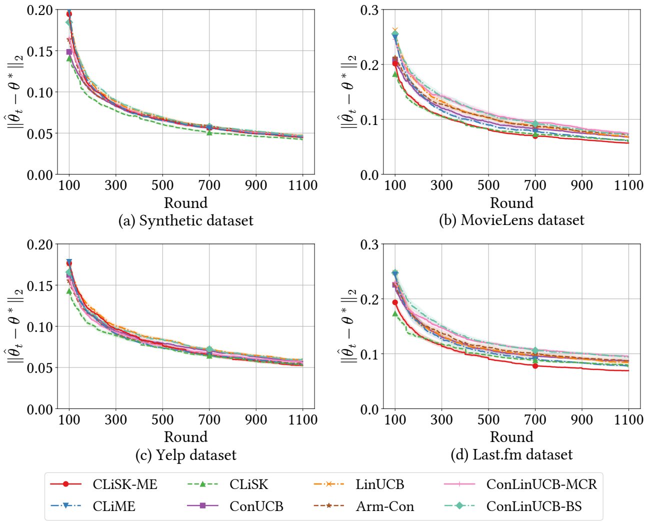
In this section, we present the details of the CLiSK-ME algorithm (Algorithm 3), which integrates the smoothed key term contexts and the adaptive conversation technique.

Algorithm 3: CLiSK-ME

```
Input: $\mathcal{A}, \mathcal{K}, b(t), \lambda, \{\alpha_t\}_{t>0}$
Initialization: $\boldsymbol{M}_1 = \lambda \boldsymbol{I}_d, \boldsymbol{b}_1 = \boldsymbol{0}_d$
for $t = 1, \dots, T$ do
    if UncertaintyChecking($t$) then
        Diagonalize $\boldsymbol{M}_t = \sum_{i=1}^{d} \lambda_{\boldsymbol{v}_i} \boldsymbol{v}_i \boldsymbol{v}_i^\top$
        foreach $\lambda_{\boldsymbol{v}_i} < \alpha t$ do
            $n_{\boldsymbol{v}_i} = \lceil (\alpha t - \lambda_{\boldsymbol{v}_i}) / c_0^2 \rceil$
            for $n_{\boldsymbol{v}_i} > 0$ do
                Smooth the key term contexts to get
                  $\{\tilde{\tilde{\boldsymbol{x}}}_k\}_{k \in \mathcal{K}}$
                $k = \arg\max_{k \in \mathcal{K}} |\tilde{\tilde{\boldsymbol{x}}}_k^\top \boldsymbol{v}_i|$
                Receive the key term-level feedback $\tilde{r}_{k,t}$
                $\boldsymbol{M}_t = \boldsymbol{M}_t + \tilde{\tilde{\boldsymbol{x}}}_{k,t} \tilde{\tilde{\boldsymbol{x}}}_{k,t}^\top$
                $\boldsymbol{b}_t = \boldsymbol{b}_t + \tilde{r}_{k,t} \tilde{\tilde{\boldsymbol{x}}}_{k,t}$
                $n_{\boldsymbol{v}_i} = n_{\boldsymbol{v}_i} - 1$
    $\boldsymbol{\theta}_t = \boldsymbol{M}_t^{-1} \boldsymbol{b}_t$
    Select $a_t = \arg\max_{a \in \mathcal{A}_t} \boldsymbol{x}_a^\top \boldsymbol{\theta}_t + \alpha_t \|\boldsymbol{x}_a\|_{M_t^{-1}}$
    Ask the user's preference for arm $a_t$
    Observe the reward $r_{a_t,t}$
    $\boldsymbol{M}_{t+1} = \boldsymbol{M}_t + \boldsymbol{x}_{a_t} \boldsymbol{x}_{a_t}^\top$
    $\boldsymbol{b}_{t+1} = \boldsymbol{b}_t + r_{a_t,t} \boldsymbol{x}_{a_t}$
```

A.2 Supplementary Experiment Results

We compare the estimation precision for the "Fixed Interval" and "Exponential Phase" uncertainty checking functions of CLiME in Figures 8 and 9. In the former, `UncertaintyChecking` is triggered every 100 rounds while in the latter it is triggered when t is a power of 2. Combined with the results presented in the evaluation results section, the experiments demonstrate that our algorithms consistently outperform the baselines.



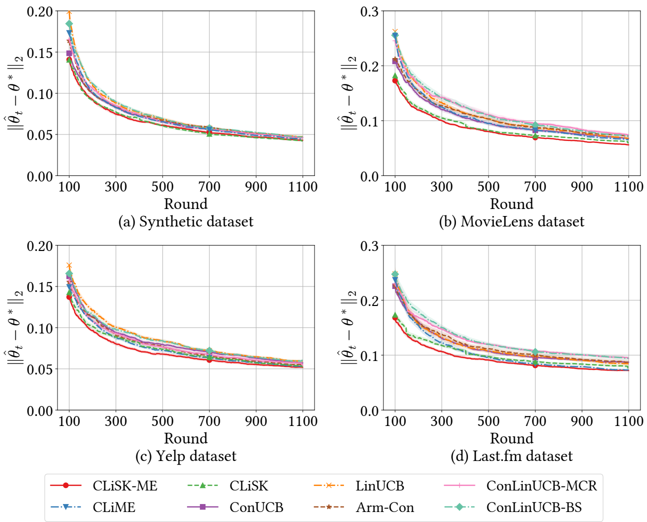
(a) Synthetic dataset

(b) MovieLens dataset

(c) Yelp dataset

(d) Last.fm dataset

Figure 8: Comparison of estimation precision where CLiME and CLiSK-ME use the "Fixed Interval" function.



(a) Synthetic dataset

(b) MovieLens dataset

(c) Yelp dataset

(d) Last.fm dataset

Figure 9: Comparison of estimation precision where CLiME and CLiSK-ME use the "Exponential Phase" function.

A.3 Proof of Lemma 1

Lemma 1. *Under Assumptions 1 and 2, for CLiSK, for any round $t \in [T]$ and any arm $a \in \mathcal{A}$, with probability at least $1 - \delta$ for some $\delta \in (0, 1)$, we have*

$$\left|\boldsymbol{x}_a^\top \boldsymbol{\theta}_t - \boldsymbol{x}_a^\top \boldsymbol{\theta}^*\right| \le \alpha_t \|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}},$$

where $\alpha_t = \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{t + \left(1+\sqrt{d}R\right)bt}{\lambda d}\right)} + \sqrt{\lambda}.$

Proof. For any arm $a \in \mathcal{A}$, from the definition of $\boldsymbol{M}_t$ and $\boldsymbol{b}_t$, and $\boldsymbol{\theta}_t = \boldsymbol{M}_t^{-1}\boldsymbol{b}_t$, we have

$$\begin{aligned}
&\boldsymbol{x}_a^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) = \boldsymbol{x}_a^\top\left(\boldsymbol{M}_t^{-1}\boldsymbol{b}_t - \boldsymbol{\theta}^*\right)\\
=&\boldsymbol{x}_a^\top\left(\boldsymbol{M}_t^{-1}\left(\sum_{s=1}^{t-1} r_{a_s,s}\boldsymbol{x}_{a_s} + \sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{r}_{k,s}\tilde{\tilde{\boldsymbol{x}}}_k\right) - \boldsymbol{\theta}^*\right)\\
=&\boldsymbol{x}_a^\top\left(\boldsymbol{M}_t^{-1}\left(\sum_{s=1}^{t-1}\boldsymbol{x}_{a_s}\left(\boldsymbol{x}_{a_s}^\top\boldsymbol{\theta}^* + \eta_s\right) + \sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\left(\tilde{\tilde{\boldsymbol{x}}}_k^\top\boldsymbol{\theta}^* + \tilde{\eta}_s\right)\right) - \boldsymbol{\theta}^*\right)\\
=&\boldsymbol{x}_a^\top\left(\boldsymbol{M}_t^{-1}\left(\sum_{s=1}^{t-1}\boldsymbol{x}_{a_s}\boldsymbol{x}_{a_s}^\top + \sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top + \lambda\boldsymbol{I}_d - \lambda\boldsymbol{I}_d\right)\boldsymbol{\theta}^* - \boldsymbol{\theta}^*\right)\\
&+\boldsymbol{x}_a^\top\left(\boldsymbol{M}_t^{-1}\left(\sum_{s=1}^{t-1}\boldsymbol{x}_{a_s}\eta_s + \sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\eta}_s\right)\right)\\
=&\lambda\boldsymbol{x}_a^\top\boldsymbol{M}_t^{-1}\boldsymbol{\theta}^* + \boldsymbol{x}_a^\top\left(\boldsymbol{M}_t^{-1}\left(\sum_{s=1}^{t-1}\boldsymbol{x}_{a_s}\eta_s + \sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\eta}_s\right)\right).
\end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\left|\boldsymbol{x}_a^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)\right| \le{}& \lambda\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}\|\boldsymbol{\theta}^*\|_{\boldsymbol{M}_t^{-1}}\\
&+\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}\left\|\sum_{s=1}^{t-1}\boldsymbol{x}_{a_s}\eta_s + \sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\eta}_s\right\|_{\boldsymbol{M}_t^{-1}}. \qquad (1)
\end{aligned}$$

For the first term, by the fact that $\lambda_{\min}(\boldsymbol{M}_t) \ge \lambda$, and by the property of the Rayleigh quotient, we have

$$\frac{\|\boldsymbol{\theta}^*\|_{\boldsymbol{M}_t^{-1}}^2}{\|\boldsymbol{\theta}^*\|_2^2} = \frac{\boldsymbol{\theta}^{*\top}\boldsymbol{M}_t^{-1}\boldsymbol{\theta}^*}{\boldsymbol{\theta}^{*\top}\boldsymbol{\theta}^*} \le \lambda_{\max}(\boldsymbol{M}_t^{-1}) \le \frac{1}{\lambda_{\min}(\boldsymbol{M}_t)} \le \frac{1}{\lambda}.$$

Therefore, we have

$$\begin{aligned}
\lambda\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}\left\|\boldsymbol{\theta}^*\right\|_{\boldsymbol{M}_t^{-1}} &\le \lambda\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}\left\|\boldsymbol{\theta}^*\right\|_2\\
&\le \lambda\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}\sqrt{\frac{1}{\lambda}} = \sqrt{\lambda}\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}. \qquad (2)
\end{aligned}$$

For the second term, from Theorem 1 in Abbasi-Yadkori et al. [1], for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, for all $t \ge 1$, we have

$$\left\|\sum_{s=1}^{t-1}\boldsymbol{x}_{a_s}\eta_s + \sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\eta}_s\right\|_{\boldsymbol{M}_t^{-1}} \le \sqrt{2\log\left(\frac{\det(\boldsymbol{M}_t)^{\frac{1}{2}}\det(\lambda\boldsymbol{I}_d)^{-\frac{1}{2}}}{\delta}\right)}. \qquad (3)$$

By adopting the determinant-trace inequality (Lemma 12), we have

$$\mathrm{Tr}(\boldsymbol{M}_t) \le d\lambda + \sum_{s=1}^{t-1}\mathrm{Tr}(\boldsymbol{x}_{a_s}\boldsymbol{x}_{a_s}^\top) + \sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\mathrm{Tr}(\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top)$$

$$\le d\lambda + t + \left(1+\sqrt{d}R\right)bt,$$

which is obtained because there are at most bt key terms selected by round t and $\|\tilde{\tilde{\boldsymbol{x}}}_k\| \le 1+\sqrt{d}R$ for all $k \in \mathcal{K}$, and therefore,

$$\det(\boldsymbol{M}_t) \le \left(\frac{\operatorname{Tr}(\boldsymbol{M}_t)}{d}\right)^d \le \left(\frac{d\lambda + t + \left(1+\sqrt{d}R\right)bt}{d}\right)^d, \tag{4}$$

where $\operatorname{Tr}(\boldsymbol{X})$ denotes the trace of matrix $\boldsymbol{X}$.

By substituting Equation (4) into Equation (3), we have

$$\left\|\sum_{s=1}^{t-1}\boldsymbol{x}_{a_s}\eta_s + \sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\eta}_s\right\|_{\boldsymbol{M}_t^{-1}} \le \sqrt{2\log\left(\frac{1}{\delta}\right)+\log\left(\frac{\det(\boldsymbol{M}_t)}{\det(\lambda\boldsymbol{I}_d)}\right)}$$

$$\le \sqrt{2\log\left(\frac{1}{\delta}\right)+d\log\left(1+\frac{t+\left(1+\sqrt{d}R\right)bt}{\lambda d}\right)}. \tag{5}$$

Plugging Equation (2) and Equation (5) into Equation (1), we have

$$\left|\boldsymbol{x}_a^\top(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*)\right|$$

$$\le \|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}\left(\sqrt{\lambda}+\sqrt{2\log\left(\frac{1}{\delta}\right)+\log\left(1+\frac{t+(1+\sqrt{d}R)bt}{\lambda d}\right)}\right). \tag{6}$$

which completes the proof. □

A.4 Proof of Lemma 2

Lemma 2. *For any round $t \in [T]$, with the smoothed key term contexts in Definition 1, CLiSK has the following lower bound on the minimum eigenvalue of the matrix $\mathbb{E}[\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top]$ for any $k\in\mathcal{K}_t$, i.e.,*

$$\lambda_{\min}\left(\mathbb{E}[\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top]\right) \ge c_1\frac{\rho^2}{\log|\mathcal{K}|} \triangleq \lambda_{\mathcal{K}},$$

where $c_1 \in (0,1)$ is some constant.

Proof. Fix a time t, and denote the key term selected at this time as k_t. Although multiple key terms may be selected at each time step, they all satisfy the properties of this lemma. Therefore, we do not distinguish between them and use only a single subscript t. Let $\boldsymbol{Q}$ be a unitary matrix that rotates the estimated preference vector $\boldsymbol{\theta}_t$ to align it with the x-axis, maintaining its length but zeroing out all components except the first component, i.e., $\boldsymbol{Q}\boldsymbol{\theta}_t = (\|\boldsymbol{\theta}_t\|, 0, 0, \ldots, 0)$. Note that such $\boldsymbol{Q}$ always exists because it just rotates the space. According to CLiSK's key term selection strategy $\tilde{\tilde{\boldsymbol{x}}}_{k_t} = \arg\max_{k\in\mathcal{K}}\boldsymbol{\theta}_t^\top\tilde{\tilde{\boldsymbol{x}}}_k$, we have

$$\lambda_{\min}\left(\mathbb{E}\left[\tilde{\tilde{\boldsymbol{x}}}_{k_t}\tilde{\tilde{\boldsymbol{x}}}_{k_t}^\top\right]\right) = \lambda_{\min}\left(\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top \;\middle|\; \boldsymbol{x} = \underset{k\in\mathcal{K}}{\arg\max}\,\boldsymbol{\theta}_t^\top\tilde{\tilde{\boldsymbol{x}}}_k\right]\right)$$

$$= \min_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\boldsymbol{w}^\top\,\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top \;\middle|\; \boldsymbol{x} = \underset{k\in\mathcal{K}}{\arg\max}\,\boldsymbol{\theta}_t^\top\tilde{\tilde{\boldsymbol{x}}}_k\right]\boldsymbol{w}$$

$$= \min_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\mathbb{E}\left[(\boldsymbol{w}^\top\boldsymbol{x})^2 \;\middle|\; \boldsymbol{x} = \underset{k\in\mathcal{K}}{\arg\max}\,\boldsymbol{\theta}_t^\top\tilde{\tilde{\boldsymbol{x}}}_k\right]$$

$$\ge \min_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\operatorname{Var}\left[\boldsymbol{w}^\top\boldsymbol{x} \;\middle|\; \boldsymbol{x} = \underset{k\in\mathcal{K}}{\arg\max}\,\boldsymbol{\theta}_t^\top\tilde{\tilde{\boldsymbol{x}}}_k\right]$$

$$= \min_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\operatorname{Var}\left[(\boldsymbol{Q}\boldsymbol{w})^\top\boldsymbol{Q}\boldsymbol{x} \;\middle|\; \boldsymbol{x} = \underset{k\in\mathcal{K}}{\arg\max}(\boldsymbol{Q}\boldsymbol{\theta}_t)^\top\boldsymbol{Q}\tilde{\tilde{\boldsymbol{x}}}_k\right] \tag{7}$$

$$= \min_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\operatorname{Var}\left[\boldsymbol{w}^\top\boldsymbol{Q}\boldsymbol{x} \;\middle|\; \boldsymbol{x} = \underset{k\in\mathcal{K}}{\arg\max}\|\boldsymbol{\theta}_t\|(\boldsymbol{Q}\tilde{\tilde{\boldsymbol{x}}}_k)_1\right] \tag{8}$$

$$= \min_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\operatorname{Var}\left[\boldsymbol{w}^\top\boldsymbol{Q}\boldsymbol{\varepsilon} \;\middle|\; \boldsymbol{\varepsilon} = \underset{\boldsymbol{\varepsilon}_k:k\in\mathcal{K}}{\arg\max}(\boldsymbol{Q}\tilde{\boldsymbol{x}}_k+\boldsymbol{Q}\boldsymbol{\varepsilon}_k)_1\right] \tag{9}$$

$$= \min_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\operatorname{Var}\left[\boldsymbol{w}^\top\boldsymbol{\varepsilon} \;\middle|\; \boldsymbol{\varepsilon} = \underset{\boldsymbol{\varepsilon}_k:k\in\mathcal{K}}{\arg\max}(\boldsymbol{Q}\tilde{\boldsymbol{x}}_k+\boldsymbol{\varepsilon}_k)_1\right] \tag{10}$$

where Equation (7) uses the property of unitary matrices: $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_d$. Equation (8) applies matrix $\boldsymbol{Q}$ so only the first component is non-zero and we use the fact that minimizing over $\boldsymbol{Q}\boldsymbol{w}$ is equivalent to over $\boldsymbol{w}$. Equation (9) follows because each smoothed key term $\tilde{\tilde{\boldsymbol{x}}}_k = \tilde{\boldsymbol{x}}_k + \boldsymbol{\varepsilon}_k$ by definition, and adding a constant a to a random variable does not change its variance. Equation (10) is due to the rotation invariance of symmetrically truncated Gaussian distributions.

Since $\boldsymbol{\varepsilon}_k \sim \mathcal{N}(0, \rho^2\cdot\boldsymbol{I}_d)$ conditioned on $|(\boldsymbol{\varepsilon}_k)_j| \le R, \forall j\in[d]$, by the property of (truncated) multivariate Gaussian distributions, the components of $\boldsymbol{\varepsilon}_{t,i}$ can be equivalently regarded as d independent samples from a (truncated) univariate Gaussian distribution, i.e., $(\boldsymbol{\varepsilon}_k)_j \sim \mathcal{N}(0,\rho^2)$ conditioned on $|(\boldsymbol{\varepsilon}_k)_j| \le R, \forall j \in [d]$. Therefore, we have

$$\operatorname{Var}\left[\boldsymbol{w}^\top\boldsymbol{\varepsilon}\right] = \operatorname{Var}\left[\sum_{i=1}^{d}\boldsymbol{w}_i\boldsymbol{\varepsilon}_i\right] = \sum_{i=1}^{d}\boldsymbol{w}_i^2\operatorname{Var}\left[\boldsymbol{\varepsilon}_i\right],$$

where the exchanging of variance and summation is due to the independence of $\boldsymbol{\varepsilon}_i$. Therefore, we can write

$$\min_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\operatorname{Var}\left[\boldsymbol{w}^\top\boldsymbol{\varepsilon} \;\middle|\; \boldsymbol{\varepsilon} = \underset{\boldsymbol{\varepsilon}_k:k\in\mathcal{K}}{\arg\max}((\boldsymbol{\varepsilon}_k)_1+(\boldsymbol{Q}\tilde{\boldsymbol{x}}_k)_1)\right]$$

$$= \min_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\sum_{j=1}^{d}\boldsymbol{w}_j^2\operatorname{Var}\left[(\boldsymbol{\varepsilon})_j \;\middle|\; \boldsymbol{\varepsilon} = \underset{\boldsymbol{\varepsilon}_k:k\in\mathcal{K}}{\arg\max}((\boldsymbol{\varepsilon}_k)_1+(\boldsymbol{Q}\tilde{\boldsymbol{x}}_k)_1)\right]$$

$$= \min_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\left\{\boldsymbol{w}_1^2\operatorname{Var}\left[(\boldsymbol{\varepsilon})_1 \;\middle|\; \boldsymbol{\varepsilon} = \underset{\boldsymbol{\varepsilon}_k:k\in\mathcal{K}}{\arg\max}((\boldsymbol{\varepsilon}_k)_1+(\boldsymbol{Q}\tilde{\boldsymbol{x}}_k)_1)\right]\right.$$

$$\left.+\sum_{j=2}^{d}\boldsymbol{w}_j^2\operatorname{Var}\left[(\boldsymbol{\varepsilon})_j \;\middle|\; \boldsymbol{\varepsilon} = \underset{\boldsymbol{\varepsilon}_k:k\in\mathcal{K}}{\arg\max}((\boldsymbol{\varepsilon}_k)_1+(\boldsymbol{Q}\tilde{\boldsymbol{x}}_k)_1)\right]\right\}$$

$$= \min_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\left\{\boldsymbol{w}_1^2\operatorname{Var}\left[(\boldsymbol{\varepsilon})_1 \;\middle|\; \boldsymbol{\varepsilon} = \underset{\boldsymbol{\varepsilon}_k:k\in\mathcal{K}}{\arg\max}((\boldsymbol{\varepsilon}_k)_1+(\boldsymbol{Q}\tilde{\boldsymbol{x}}_k)_1)\right]\right.$$

$$\left.+\sum_{j=2}^{d}\boldsymbol{w}_j^2\operatorname{Var}\left[(\boldsymbol{\varepsilon})_j\right]\right\}$$

$$= \min_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\left\{\boldsymbol{w}_1^2\operatorname{Var}\left[(\boldsymbol{\varepsilon})_1 \;\middle|\; \boldsymbol{\varepsilon} = \underset{\boldsymbol{\varepsilon}_k:k\in\mathcal{K}}{\arg\max}((\boldsymbol{\varepsilon}_k)_1+(\boldsymbol{Q}\tilde{\boldsymbol{x}}_k)_1)\right] + (1-\boldsymbol{w}_1^2)\rho^2\right\}$$

$$= \min\left\{\operatorname{Var}\left[(\boldsymbol{\varepsilon})_1 \;\middle|\; \boldsymbol{\varepsilon} = \underset{\boldsymbol{\varepsilon}_k:k\in\mathcal{K}}{\arg\max}((\boldsymbol{\varepsilon}_k)_1+(\boldsymbol{Q}\tilde{\boldsymbol{x}}_k)_1)\right], \rho^2\right\} \ge c_1\frac{\rho^2}{\log|\mathcal{K}|},$$

where in the last inequality, we use Lemma 15 and Lemma 14 in Sivakumar et al. [22] and get

$$\begin{aligned}&\mathrm{Var}\left[(\boldsymbol{\varepsilon})_1 \,\middle|\, \boldsymbol{\varepsilon} = \underset{\boldsymbol{\varepsilon}_k : k\in\mathcal{K}}{\arg\max}((\boldsymbol{\varepsilon}_k)_1 + (\boldsymbol{Q}\tilde{\boldsymbol{x}}_k)_1)\right]\\ \geq\ &\mathrm{Var}\left[(\boldsymbol{\varepsilon})_1 \,\middle|\, \boldsymbol{\varepsilon} = \underset{\boldsymbol{\varepsilon}_k : k\in\mathcal{K}}{\arg\max}(\boldsymbol{\varepsilon}_k)_1\right] \geq c_1 \frac{\rho^2}{\log|\mathcal{K}|}. \qquad \square\end{aligned}$$

A.5 Proof of Lemma 3

Lemma 3. *For CLiSK, with probability at least* $1-\delta$ *for some* $\delta \in (0,1)$*, if* $t \geq T_0 \triangleq \frac{8(1+\sqrt{d}R)^2}{b\lambda_{\mathcal{K}}}\log\left(\frac{d}{\delta}\right)$*, we have*

$$\lambda_{\min}\left(\sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top\right) \geq \frac{\lambda_{\mathcal{K}}bt}{2}.$$

Proof. To apply the matrix Chernoff bound (Lemma 11), we first verify the required two conditions for the self-adjoint matrices $\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top$ for any $k \in \mathcal{K}_s$ and $s \in [t]$. First, $\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top$ is obviously positive semi-definite. Second, by the Courant-Fischer theorem,

$$\begin{aligned}\lambda_{\max}(\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top) &= \max_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\boldsymbol{w}^\top\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top\boldsymbol{w} = \max_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}(\boldsymbol{w}^\top\tilde{\tilde{\boldsymbol{x}}}_k)^2\\ &\leq \max_{\boldsymbol{w}:\|\boldsymbol{w}\|=1}\|\boldsymbol{w}\|^2\|\tilde{\tilde{\boldsymbol{x}}}_k\|^2 \leq (1+\sqrt{d}R)^2.\end{aligned}$$

Next, by Lemma 2 and the super-additivity of the minimum eigenvalue (due to Weyl's inequality), we have

$$\mu_{\min} = \lambda_{\min}\left(\sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\mathbb{E}\left[\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top\right]\right) \geq \sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\lambda_{\min}\left(\mathbb{E}\left[\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top\right]\right) \geq \lambda_{\mathcal{K}}bt,$$

where the last inequality is because there are at most bt key terms selected by round t, so the summation has at most bt terms. So by Lemma 11, we have for any $\varepsilon \in (0,1)$,

$$\begin{aligned}&\Pr\left[\lambda_{\min}\left(\sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top\right) \leq (1-\varepsilon)\lambda_{\mathcal{K}}bt\right]\\ \leq&\Pr\left[\lambda_{\min}\left(\sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top\right) \leq (1-\varepsilon)\mu_{\min}\right]\\ \leq& d\left[\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right]^{\mu_{\min}/(1+\sqrt{d}R)^2}\\ \leq& d\left[\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right]^{\frac{\lambda_{\mathcal{K}}bt}{(1+\sqrt{d}R)^2}},\end{aligned}$$

where the last inequality is because e^{-x} is decreasing. Choosing $\varepsilon = \frac{1}{2}$, we get

$$\Pr\left[\lambda_{\min}\left(\sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top\right) \leq \frac{\lambda_{\mathcal{K}}bt}{2}\right] \leq d\left(\sqrt{2}e^{-\frac{1}{2}}\right)^{\frac{\lambda_{\mathcal{K}}bt}{(1+\sqrt{d}R)^2}}.$$

Letting the RHS be δ, we get $t = \frac{2(1+\sqrt{d}R)^2\log(\frac{d}{\delta})}{\lambda_{\mathcal{K}}b(1-\log(2))} \leq \frac{8(1+\sqrt{d}R)^2}{\lambda_{\mathcal{K}}b}\log\left(\frac{d}{\delta}\right)$. Therefore, $\lambda_{\min}\left(\sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top\right) \geq \frac{\lambda_{\mathcal{K}}bt}{2}$ holds with probability at least $1-\delta$ when $t \geq \frac{8(1+\sqrt{d}R)^2}{\lambda_{\mathcal{K}}b}\log\left(\frac{d}{\delta}\right)$. □

A.6 Proof of Lemma 4

Lemma 4. *For CLiSK, for any* $a \in \mathcal{A}$*, if* $t \geq T_0 \triangleq \frac{8(1+\sqrt{d}R)^2}{b\lambda_{\mathcal{K}}}\log\left(\frac{d}{\delta}\right)$*, with probability at least* $1-\delta$ *for some* $\delta \in (0,1)$*,* $\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}} \leq \sqrt{\frac{2}{\lambda_{\mathcal{K}}bt}}$.

Proof.

$$\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}} = \sqrt{\boldsymbol{x}_a^\top\boldsymbol{M}_t^{-1}\boldsymbol{x}_a} \leq \sqrt{\lambda_{\max}(\boldsymbol{M}_t^{-1})\boldsymbol{x}_a^\top\boldsymbol{x}_a} = \sqrt{\frac{1}{\lambda_{\min}(\boldsymbol{M}_t)}}, \tag{11}$$

where the first inequality is due to the property of the Rayleigh quotient, and the second inequality is due to the fact that $\boldsymbol{x}_a^\top\boldsymbol{x}_a = 1$.

By the definition of $\boldsymbol{M}_t$, we have

$$\begin{aligned}\lambda_{\min}(\boldsymbol{M}_t) &= \lambda_{\min}\left(\sum_{s=1}^{t-1}\boldsymbol{x}_{a_s}\boldsymbol{x}_{a_s}^\top + \sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top + \lambda\boldsymbol{I}_d\right)\\ &\geq \lambda_{\min}\left(\sum_{s=1}^{t}\sum_{k\in\mathcal{K}_s}\tilde{\tilde{\boldsymbol{x}}}_k\tilde{\tilde{\boldsymbol{x}}}_k^\top\right)\\ &\geq \frac{\lambda_{\mathcal{K}}bt}{2},\end{aligned} \tag{12}$$

where the first inequality follows the property of Loewner order that if $\boldsymbol{A} \geq \boldsymbol{B}$ then $\lambda_{\min}(\boldsymbol{A}) \geq \lambda_{\min}(\boldsymbol{B})$, and the last inequality follows from Lemma 3 conditioned on $t \geq T_0$.

Therefore, by plugging Equation (12) into Equation (11), we have $\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}} \leq \sqrt{\frac{2}{\lambda_{\mathcal{K}}bt}}$. □

A.7 Proof of Lemma 5

Lemma 5. *Let* $\boldsymbol{\theta}_t$ *be the estimated preference vector at round* t *and* $\boldsymbol{\theta}^*$ *be the true preference vector. Under Assumptions 1, 2 and 3, for CLiME, at round* t*, for any arm* $a \in \mathcal{A}$*, with probability at least* $1-\delta$ *(*$\delta \in (0,1)$*), we have*

$$\left|\boldsymbol{x}_a^\top\boldsymbol{\theta}_t - \boldsymbol{x}_a^\top\boldsymbol{\theta}^*\right| \leq \alpha_t\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}},$$

where $\alpha_t = \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{t+\alpha dt}{\lambda dc_0^2}\right)} + \sqrt{\lambda}$*,* α *is an exploration control factor in Algorithm 2, and* c_0 *is a constant in Assumption 3.*

Proof. The proof of Lemma 5 is similar to that of Lemma 1. The only difference is the trace and determinant of the matrix $\boldsymbol{M}_t$.

We first show that by round t, at most $\frac{\alpha dt}{c_0^2}$ key terms have been selected since the beginning of the algorithm for all three uncertainty checking functions.

Consider the case where CLiME uses the "Continuous Checking" function, i.e., the agent checks the eigenvalues of the matrix $\boldsymbol{M}_t$ at each round. We first denote the covariance matrix before selecting key terms at round t by $\boldsymbol{M}_t'$, i.e., $\boldsymbol{M}_t = \boldsymbol{M}_t' + \sum_{k\in\mathcal{K}_t}\boldsymbol{x}_k\boldsymbol{x}_k^\top$. For $\boldsymbol{M}_t'$, denote its eigenvectors by $\{\boldsymbol{v}_i\}_{i=1}^d$ and corresponding eigenvalues by $\{\lambda_{\boldsymbol{v}_i}\}_{i=1}^d$. If some key term k is selected at round t, then there must exist an eigenvector $\boldsymbol{v}_i$ such that $\lambda_{\boldsymbol{v}_i} < \alpha t$, and the corresponding key term context $\tilde{\boldsymbol{x}}_k$ is close to $\boldsymbol{v}_i$, i.e., $\tilde{\boldsymbol{x}}_k^\top\boldsymbol{v}_i \geq c_0$.

We can write the vector $\tilde{\boldsymbol{x}}_k = \sum_{i=1}^d \gamma_i\boldsymbol{v}_i$ for some coefficients $\{\gamma_i\}_{i=1}^d$. Then, for $j \in [d]$, Denote $\boldsymbol{z}_j = \sum_{i=1, i\neq j}^d \gamma_i\boldsymbol{v}_i$. For any $j \in [d]$, we have $\tilde{\boldsymbol{x}}_k^\top\boldsymbol{v}_j = \sum_{i=1}^d \gamma_i\boldsymbol{v}_i^\top\boldsymbol{v}_j = \gamma_j \geq c_0$ and $\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top =$

$(\gamma_j \boldsymbol{v}_j + \boldsymbol{z}_j)(\gamma_j \boldsymbol{v}_j + \boldsymbol{z}_j)^\top = \gamma_j^2 \boldsymbol{v}_j \boldsymbol{v}_j^\top + \boldsymbol{z}_j \boldsymbol{z}_j^\top$, and then, we have the following:

$$
\begin{aligned}
&\boldsymbol{M}_t' + \sum_{k \in \mathcal{K}_t} \boldsymbol{x}_k \boldsymbol{x}_k^\top = \boldsymbol{M}_t' + \sum_{i \in [d]: \lambda_{\boldsymbol{v}_i} \le \alpha t} \left\lceil \frac{\alpha t - \lambda_{\boldsymbol{v}_i}}{c_0^2} \right\rceil (\gamma_i^2 \boldsymbol{v}_i \boldsymbol{v}_i^\top + \boldsymbol{z}_i \boldsymbol{z}_i^\top) \\
&\ge \sum_{i=1}^d \lambda_{\boldsymbol{v}_i} \boldsymbol{v}_i \boldsymbol{v}_i^\top + \sum_{i \in [d]: \lambda_{\boldsymbol{v}_i} \le \alpha t} \frac{\alpha t - \lambda_{\boldsymbol{v}_i}}{c_0^2} (\gamma_i^2 \boldsymbol{v}_i \boldsymbol{v}_i^\top + \boldsymbol{z}_i \boldsymbol{z}_i^\top) \\
&\ge \sum_{i=1}^d \lambda_{\boldsymbol{v}_i} \boldsymbol{v}_i \boldsymbol{v}_i^\top + \sum_{i \in [d]: \lambda_{\boldsymbol{v}_i} \le \alpha t} (\alpha t - \lambda_{\boldsymbol{v}_i}) \boldsymbol{v}_i \boldsymbol{v}_i^\top \\
&= \sum_{i \in [d]: \lambda_{\boldsymbol{v}_i} < \alpha t} (\alpha t - \lambda_{\boldsymbol{v}_i} + \lambda_{\boldsymbol{v}_i}) \boldsymbol{v}_i \boldsymbol{v}_i^\top + \sum_{i \in [d]: \lambda_{\boldsymbol{v}_i} > \alpha t} \lambda_{\boldsymbol{v}_i} \boldsymbol{v}_i \boldsymbol{v}_i^\top \\
&\ge \sum_{i=1}^d \alpha t \boldsymbol{v}_i \boldsymbol{v}_i^\top. \qquad (13)
\end{aligned}
$$

Following from Equation (13), we have

$$\lambda_{\min}(\boldsymbol{M}_t) \ge \alpha t. \qquad (14)$$

Denote the number of key terms selected at round t as K_t. We have $K_t = \sum_{i=1}^d \frac{\alpha t - \lambda_{\boldsymbol{v}_i}}{c_0^2}$. Since $\lambda_{\boldsymbol{v}_i} \ge \alpha(t-1), \forall i \in [d]$ according to Equation (14), we have $K_t \le \frac{\alpha d}{c_0^2}$, and then $\sum_{s=1}^t K_s \le \frac{\alpha d t}{c_0^2}$.

For the "Fixed Interval Checking" function, at each uncertainty checking point $t_j = jP$ where $j \in \{1, 2, \dots, \lfloor \frac{T}{P} \rfloor\}$, we have $\lambda_{\min}(\boldsymbol{M}_{t_j}) \ge \alpha t_j$. For the j-th checking, there are $\sum_{i=1}^d \frac{\alpha t_j - \lambda_{\boldsymbol{v}_i}}{c_0^2} \le \sum_{i=1}^d \frac{\alpha t_j - \alpha t_{j-1}}{c_0^2} \le \frac{\alpha d P}{c_0^2}$ conversations to be launched. Thus, by round t, the number of total conversations satisfies $\sum_{j=1}^{\lfloor \frac{t}{P} \rfloor} \frac{\alpha d P}{c_0^2} \le \frac{\alpha d t}{c_0^2}$.

For the "Exponential Phase Checking" function, by round t, there are $\lfloor \log_2(t) \rfloor$ uncertainty checking points. For the j-th checking, there are $\sum_{i=1}^d \frac{\alpha t_j - \lambda_{\boldsymbol{v}_i}}{c_0^2} \le \sum_{i=1}^d \frac{\alpha t_j - \alpha t_{j-1}}{c_0^2} \le \frac{\alpha d 2^{j-1}}{c_0^2}$ conversations to be launched. By round t, the number of total conversations satisfies $\sum_{j=1}^{\lfloor \log_2(t) \rfloor} \frac{\alpha d 2^{j-1}}{c_0^2} \le \frac{\alpha d t}{c_0^2}$.

Therefore, we have

$$\mathrm{Tr}(\boldsymbol{M}_t) \le d\lambda + \sum_{s=1}^{t-1} \mathrm{Tr}(\boldsymbol{x}_{a_s} \boldsymbol{x}_{a_s}^\top) + \sum_{s=1}^{t} \sum_{k \in \mathcal{K}_s} \mathrm{Tr}(\tilde{\boldsymbol{x}}_k \tilde{\boldsymbol{x}}_k^\top) \le d\lambda + t + \frac{\alpha d t}{c_0^2},$$

and

$$\det(\boldsymbol{M}_t) \le \left(\frac{\mathrm{Tr}(\boldsymbol{M}_t)}{d} \right)^d \le \left(\frac{d\lambda + t + \frac{\alpha d t}{c_0^2}}{d} \right)^d \le \left(\lambda + \frac{t + \alpha d t}{c_0^2 d} \right)^d,$$

where the last inequality is obtained by the fact that $c_0 < 1$.

Following the same steps as in the proof of Lemma 1, we can obtain that

$$\left| \boldsymbol{x}_a^\top (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) \right| \le \|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}} \left(\sqrt{\lambda} + \sqrt{2 \log\left(\frac{1}{\delta}\right) + d \log\left(1 + \frac{t + \alpha d t}{\lambda d c_0^2}\right)} \right),$$

which concludes the proof. □

A.8 Proof of Lemma 6

Lemma 6. *For CLiME, for any arm $a \in \mathcal{A}$, with probability at least $1 - \delta$ for some $\delta \in (0, 1)$, at round $t \ge 2P$, we have $\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}} \le \sqrt{\frac{2}{\alpha t}}$, where P is a fixed integer.*

Proof. We first consider the case where CLiME uses the "Continuous Checking" function, i.e., the agent checks the eigenvalues of the matrix $\boldsymbol{M}_t$ at each round. By Equation (14), we have $\lambda_{\min}(\boldsymbol{M}_t) \ge \alpha t$. Then, following from Equation (11) in the proof of Lemma 4, we can obtain that $\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}} \le \sqrt{\frac{1}{\alpha t}}$.

Next, we consider the case for the "Fixed Interval Checking" function. In this case, the agent only checks the eigenvalues of the matrix $\boldsymbol{M}_t$ every P rounds. For the rounds t when the agent checks the uncertainty, we have the same results as $\lambda_{\min}(\boldsymbol{M}_t) \ge \alpha t$; For the rounds t when the agent does not check it, we have $\lambda_{\min}(\boldsymbol{M}_t) \ge \alpha t'$ where t' is the last round that the agent conducts the check and $t - t' \le P$. When $t \ge 2P$, $t' \ge t - P \ge \frac{t}{2}$, we can obtain that $\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}} \le \sqrt{\frac{1}{\alpha t'}} \le \sqrt{\frac{2}{\alpha t}}$.

Finally, we consider the "Exponential Phase Checking" function. At rounds t satisfying $2^i \le t < 2^{i+1}$ for $i = 1, 2, \dots$, the last checking point $t' = 2^i$, then we have $\lambda_{\min}(\boldsymbol{M}_t) \ge \alpha \cdot 2^i$. When $t \ge 2$, we have $\frac{t}{2} \le 2^i$, and then $\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}} \le \sqrt{\frac{1}{\alpha 2^i}} \le \sqrt{\frac{2}{\alpha t}}$.

Therefore, to generalize the bound, we can conclude that when $t \ge 2P$, $\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}} \le \sqrt{\frac{2}{\alpha t}}$ for all three checking functions. □

A.9 Proof of Theorem 1

Theorem 1 (Regret of CLiSK). *With probability at least $1 - \delta$ for some $\delta \in (0, 1)$, the regret upper bound of CLiSK satisfies*

$$
\begin{aligned}
\mathrm{R}(T) \le\ & \frac{8(1 + \sqrt{d}R)^2 \log(|\mathcal{K}|)}{c_1 \rho^2 b} \log\left(\frac{d}{\delta}\right) + 4\sqrt{\frac{2 c_1 \rho^2 T}{b \log(|\mathcal{K}|)}} \cdot \\
& \left(\sqrt{2 \log\left(\frac{1}{\delta}\right) + d \log\left(1 + \frac{T + \left(1 + \sqrt{d}R\right) b T}{\lambda d}\right)} + \sqrt{\lambda} \right) \\
=\ & O(\sqrt{dT \log(T)} + d),
\end{aligned}
$$

where R and ρ^2 are constants in Definition 1.

Proof. Denote the instantaneous regret at round t by reg_t. We first decompose it as follows:

$$
\begin{aligned}
&\mathrm{reg}_t = (\boldsymbol{x}_{a_t^*}^\top \boldsymbol{\theta}^* + \eta_t) - (\boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}^* + \eta_t) \\
=& \boldsymbol{x}_{a_t^*}^\top (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t) + (\boldsymbol{x}_{a_t^*}^\top \boldsymbol{\theta}_t + \alpha_t \|\boldsymbol{x}_{a_t^*}\|_{\boldsymbol{M}_t^{-1}}) - (\boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}_t + \alpha_t \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{M}_t^{-1}}) \\
&+ \boldsymbol{x}_{a_t}^\top (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \alpha_t \|\boldsymbol{x}_{a_t^*}\|_{\boldsymbol{M}_t^{-1}} + \alpha_t \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{M}_t^{-1}} \\
\le& \boldsymbol{x}_{a_t^*}^\top (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t) + \boldsymbol{x}_{a_t}^\top (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \alpha_t \|\boldsymbol{x}_{a_t^*}\|_{\boldsymbol{M}_t^{-1}} + \alpha_t \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{M}_t^{-1}} \qquad (15) \\
\le& \alpha_t \|\boldsymbol{x}_{a_t^*}\|_{\boldsymbol{M}_t^{-1}} + \alpha_t \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{M}_t^{-1}} - \alpha_t \|\boldsymbol{x}_{a_t^*}\|_{\boldsymbol{M}_t^{-1}} + \alpha_t \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{M}_t^{-1}} \qquad (16) \\
\le& 2\alpha_t \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{M}_t^{-1}},
\end{aligned}
$$

where Equation (15) follows from the UCB strategy for arm selection, and Equation (16) follows from Lemma 1. Next, we have

$$\mathrm{R}(T) = \sum_{t=1}^{T_0} \mathrm{reg}_t + \sum_{t=T_0+1}^{T} \mathrm{reg}_t$$

$$\le T_0 + \sum_{t=T_0+1}^{T} 2\alpha_t \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{M}_t^{-1}} \tag{17}$$

$$\le T_0 + 2\sum_{t=T_0+1}^{T} \alpha_t \sqrt{\frac{2}{\lambda_{\mathcal{K}} b t}} \tag{18}$$

$$\le T_0 + 4\alpha_T \sqrt{\frac{2T}{\lambda_{\mathcal{K}} b}} \tag{19}$$

where Equation (17) is because the instantaneous regret $\text{reg}_t \le 1$ by Assumption 1, Equation (18) follows from Lemma 4, and Equation (19) is because α_t is non-decreasing and $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \le 2\sqrt{T}$.

Recall the definition of $T_0 \triangleq \frac{8(1+\sqrt{d}R)^2}{b\lambda_{\mathcal{K}}} \log\left(\frac{d}{\delta}\right)$ in Lemma 3 and the definition of α_t in Lemma 1. Plugging T_0 and α_t into Equation (19), we can obtain the regret bound. □

A.10 Proof of Theorem 2

Theorem 2 (Regret of CLiME). *With probability at least $1-\delta$ for some $\delta \in (0,1)$, the regret upper bound of CLiME satisfies*

$$\mathrm{R}(T) \le 4\sqrt{\frac{2T}{\alpha}}\left(\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{T+\alpha d T}{\lambda d c_0^2}\right)} + \sqrt{\lambda}\right) + 2P$$
$$= O\left(\sqrt{dT\log(T)}\right).$$

Proof. With the same decomposition as in the proof of Theorem 1, we have

$$\mathrm{R}(T) = \sum_{t=1}^{2P} \text{reg}_t + \sum_{t=2P+1}^{T} \text{reg}_t \le 2P + 2\sum_{t=2P+1}^{T} \alpha_t \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{M}_t^{-1}}$$

$$\le 2P + 2\sum_{t=2P+1}^{T} \alpha_t \sqrt{\frac{2}{\alpha t}} \tag{20}$$

$$\le 2P + 4\alpha_T\sqrt{\frac{2T}{\alpha}}. \tag{21}$$

$$= 2P + 4\sqrt{\frac{2T}{\alpha}}\left(\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{T+\alpha d T}{\lambda d c_0^2}\right)} + \sqrt{\lambda}\right), \tag{22}$$

where Equations (20) and (21) follow from Lemma 5 and analogous steps in Theorem 1. Note that $P > 1$ is a given constant for the "Fixed Interval Checking" function. Plugging α_T into the inequality, we can obtain the result and conclude that $\mathrm{R}(T) = O(\sqrt{dT\log(T)})$. □

A.11 Proof of Theorem 3

Since any algorithms for conversational bandits must select both arms and key terms, we model a policy π as a tuple consisting of two components $\pi = (\pi^{\text{arm}}, \pi^{\text{key}})$, where π^{arm} selects arms and π^{key} selects key terms. We assume that at each time step, the policy can select at most one key term; otherwise, the number of key terms could exceed the number of arms, which is impractical. Let $\mathcal{H}_t = \{a_1, x_1, k_1, \widetilde{x}_1, \ldots, a_t, x_t, k_t, \widetilde{x}_t\}$ denote the history of interactions between the policy and the environment up to time t. We note that the presence of key terms at every time step in $\mathcal{H}_t$ is without loss of generality because we allow k_t to be empty if no conversation is initiated at round t. The noise terms associated with both arm-level and key term-level feedback, denoted by η_t and $\widetilde{\eta}_t$, follow the standard Gaussian distribution $\mathcal{N}(0,1)$. We also denote the feature vectors of selected arm and key term as random variables $\boldsymbol{A}_t, \boldsymbol{K}_t \in \mathbb{R}^d$, and the arm-level and key term-level rewards $X_t = \langle \boldsymbol{A}_t, \boldsymbol{\theta}\rangle + \eta_t$ and $\widetilde{X}_t = \langle \boldsymbol{K}_t, \boldsymbol{\theta}\rangle + \widetilde{\eta}_t$, follow $\mathcal{N}(\langle \boldsymbol{A}_t, \boldsymbol{\theta}\rangle, 1)$ and $\mathcal{N}(\langle \boldsymbol{K}_t, \boldsymbol{\theta}\rangle, 1)$, respectively. We denote by $\mathbb{P}_{\boldsymbol{\theta}}$ the probability measure induced by the environment $\boldsymbol{\theta}$ and policy π, and by $\mathbb{E}_{\boldsymbol{\theta}}$ the expectation under $\mathbb{P}_{\boldsymbol{\theta}}$. With these definitions, we present the following lemma.

Lemma 7. *Let $D(P \parallel Q)$ denote the KL divergence between distributions P and Q, and let $\boldsymbol{\theta}, \boldsymbol{\theta}'$ be two environments, then we have*

$$D(\mathbb{P}_{\boldsymbol{\theta}} \parallel \mathbb{P}_{\boldsymbol{\theta}'}) = \frac{1}{2}\sum_{t=1}^{T}\left(\mathbb{E}_{\boldsymbol{\theta}}\left[\left\langle \boldsymbol{A}_t, \boldsymbol{\theta} - \boldsymbol{\theta}'\right\rangle^2\right] + \mathbb{E}_{\boldsymbol{\theta}}\left[\left\langle \boldsymbol{K}_t, \boldsymbol{\theta} - \boldsymbol{\theta}'\right\rangle^2\right]\right).$$

Proof. Given a bandit instance with parameter $\boldsymbol{\theta}$ and a policy π, according to Section 4.6 of Lattimore and Szepesvári [14], we construct the canonical bandit model of our setting as follows. Let $(\Omega, \mathcal{F}, \mathbb{P}_{\boldsymbol{\theta}})$ be a probability space and $\mathcal{A}$ be the set of all possible arms, where $\Omega = (\mathcal{A} \times \mathbb{R})^T$, $\mathcal{F} = \mathcal{B}(\Omega)$, and the density function of the probability measure $\mathbb{P}_{\boldsymbol{\theta}}$ is defined by $p_{\boldsymbol{\theta},\pi} : \Omega \to \mathbb{R}$:

$$p_{\boldsymbol{\theta}}(\mathcal{H}_T) = \prod_{t=1}^{T} \pi_t^{\text{arm}}(a_t \mid \mathcal{H}_{t-1}) p_{a_t}(x_t) \cdot \pi_t^{\text{key}}(k_t \mid \mathcal{H}_{t-1}) \widetilde{p}_{k_t}(\widetilde{x}_t),$$

where p_{a_t} and $\widetilde{p}_{k_t}$ are the density functions of arm-level and key term-level reward distributions P_{a_t} and $\widetilde{P}_{k_t}$, respectively. The definition of $\mathbb{P}_{\boldsymbol{\theta}'}$ is identical except that $p_{a_t}, \widetilde{p}_{k_t}$ are replaced by p'_{a_t}, $\widetilde{p}'_{k_t}$ and $P_{a_t}, \widetilde{P}_{k_t}$ are replaced by $P'_{a_t}, \widetilde{P}'_{k_t}$.

By the definition of KL divergence $D(P \parallel Q) = \int_{\Omega} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P$,

$$D(\mathbb{P}_{\boldsymbol{\theta}} \parallel \mathbb{P}_{\boldsymbol{\theta}'}) = \int_{\Omega} \log\left(\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}'}}\right) \mathrm{d}\mathbb{P}_{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\theta}}\left[\log \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}'}}\right].$$

Note that

$$\log\left(\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}'}}(\mathcal{H}_T)\right) = \log \frac{p_{\boldsymbol{\theta},\pi}(\mathcal{H}_T)}{p_{\boldsymbol{\theta}',\pi}(\mathcal{H}_T)} \tag{23}$$

$$= \log \frac{\prod_{t=1}^{T} \pi_t^{\text{arm}}(a_t \mid \mathcal{H}_{t-1}) p_{a_t}(x_t) \cdot \pi_t^{\text{key}}(k_t \mid \mathcal{H}_{t-1}) \widetilde{p}_{k_t}(\widetilde{x}_t)}{\prod_{t=1}^{T} \pi_t^{\text{arm}}(a_t \mid \mathcal{H}_{t-1}) p'_{a_t}(x_t) \cdot \pi_t^{\text{key}}(k_t \mid \mathcal{H}_{t-1}) \widetilde{p}'_{k_t}(\widetilde{x}_t)}$$

$$= \sum_{t=1}^{T}\left(\log \frac{p_{a_t}(x_t)}{p'_{a_t}(x_t)} + \log \frac{\widetilde{p}_{k_t}(\widetilde{x}_t)}{\widetilde{p}'_{k_t}(\widetilde{x}_t)}\right).$$

where in Equation 23 we used the chain rule for Radon–Nikodym derivatives, and in the last equality, all the terms involving the policy π cancel. Therefore,

$$D(\mathbb{P}_{\boldsymbol{\theta}} \parallel \mathbb{P}_{\boldsymbol{\theta}'})$$
$$= \sum_{t=1}^{T}\left(\mathbb{E}_{\boldsymbol{\theta}}\left[\log \frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)}\right] + \mathbb{E}_{\boldsymbol{\theta}}\left[\log \frac{\widetilde{p}_{K_t}(\widetilde{X}_t)}{\widetilde{p}'_{K_t}(\widetilde{X}_t)}\right]\right)$$
$$= \sum_{t=1}^{T}\left(\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[\log \frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)}\middle| A_t\right]\right] + \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[\log \frac{\widetilde{p}_{K_t}(\widetilde{X}_t)}{\widetilde{p}'_{K_t}(\widetilde{X}_t)}\middle| K_t\right]\right]\right)$$
$$= \sum_{t=1}^{T}\left(\mathbb{E}_{\boldsymbol{\theta}}\left[D(P_{A_t} \parallel P'_{A_t})\right] + \mathbb{E}_{\boldsymbol{\theta}}\left[D(\widetilde{P}_{K_t} \parallel \widetilde{P}'_{K_t})\right]\right)$$

$$=\frac{1}{2}\sum_{t=1}^{T}\Big(\mathbb{E}_{\boldsymbol{\theta}}\Big[\big\langle \boldsymbol{A}_t,\boldsymbol{\theta}-\boldsymbol{\theta}'\big\rangle^2\Big]+\mathbb{E}_{\boldsymbol{\theta}}\Big[\big\langle \boldsymbol{K}_t,\boldsymbol{\theta}-\boldsymbol{\theta}'\big\rangle^2\Big]\Big).$$

where the last equation uses Lemma 10 and the fact that $P_{A_t}\sim\mathcal{N}(\langle A_t,\boldsymbol{\theta}\rangle,1)$, $P'_{A_t}\sim\mathcal{N}(\langle A_t,\boldsymbol{\theta}'\rangle,1)$, $\widetilde{P}_{K_t}\sim\mathcal{N}(\langle K_t,\boldsymbol{\theta}\rangle,1)$, and $\widetilde{P}'_{K_t}\sim\mathcal{N}(\langle K_t,\boldsymbol{\theta}'\rangle,1)$, respectively. □

Next, we present a lower bound for conversational bandits but *without* imposing the constraint that the number of arms is K.

Lemma 8. *Let the arm set and the key term set $\mathcal{A}=\mathcal{K}=[-1,1]^d$ and $\Theta=\left\{\pm\sqrt{\frac{1}{T}}\right\}^d$, then for any policy, there exists an environment $\boldsymbol{\theta}\in\Theta$ such that the expected regret satisfies: $\mathbb{E}_{\boldsymbol{\theta}}[R(T)]\geq\frac{\exp(-4)}{4}d\sqrt{T}$.*

Proof. For any $i\in[d]$ and $\boldsymbol{\theta}\in\Theta$, define $\mathcal{E}_{\boldsymbol{\theta},i}$ as the event that the sign of the i-th coordinate of at least half of $\{\boldsymbol{A}_t\}_{t=1}^T$ does not agree with $\boldsymbol{\theta}$: $\mathcal{E}_{\boldsymbol{\theta},i}=\left\{\sum_{t=1}^T\mathbb{I}\{\operatorname{sign}(\boldsymbol{A}_{ti})\neq\operatorname{sign}(\boldsymbol{\theta}_i)\}\geq\frac{T}{2}\right\}$.

Let $p_{\boldsymbol{\theta},i}=\mathbb{P}_{\boldsymbol{\theta}}\big[\mathcal{E}_{\boldsymbol{\theta},i}\big]$ and $\boldsymbol{\theta}'=(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_{i-1},-\boldsymbol{\theta}_i,\boldsymbol{\theta}_{i+1},\dots,\boldsymbol{\theta}_d)^\top$, i.e., $\boldsymbol{\theta}'$ is the same as $\boldsymbol{\theta}$ except that the i-th coordinate is negated. It is easy to verify that $\mathcal{E}^c_{\boldsymbol{\theta},i}=\mathcal{E}_{\boldsymbol{\theta}',i}$. Thus, applying Lemma 9 and Lemma 7, we obtain

$$\begin{aligned}
&p_{\boldsymbol{\theta},i}+p_{\boldsymbol{\theta}',i}\geq\frac{1}{2}\exp(D(\mathbb{P}_{\boldsymbol{\theta}}\parallel\mathbb{P}_{\boldsymbol{\theta}'}))\\
&=\frac{1}{2}\exp\left(-\frac{1}{2}\sum_{t=1}^{T}\Big(\mathbb{E}_{\boldsymbol{\theta}}\Big[\big\langle \boldsymbol{A}_t,\boldsymbol{\theta}-\boldsymbol{\theta}'\big\rangle^2\Big]+\mathbb{E}_{\boldsymbol{\theta}}\Big[\big\langle \boldsymbol{K}_t,\boldsymbol{\theta}-\boldsymbol{\theta}'\big\rangle^2\Big]\Big)\right)\\
&=\frac{1}{2}\exp(-4),
\end{aligned}$$

where the last equality follows from a straightforward calculation showing that $\langle \boldsymbol{A}_t,\boldsymbol{\theta}-\boldsymbol{\theta}'\rangle=\langle \boldsymbol{A}_t,\boldsymbol{\theta}-\boldsymbol{\theta}'\rangle=4/T$.

Since $|\Theta|=2^d$, we have

$$\begin{aligned}
\sum_{\boldsymbol{\theta}\in\Theta}\frac{1}{|\Theta|}\sum_{i=1}^{d}p_{\boldsymbol{\theta},i}&=\frac{1}{|\Theta|}\sum_{i=1}^{d}\sum_{\boldsymbol{\theta}\in\Theta}p_{\boldsymbol{\theta},i}\\
&=\frac{1}{2^d}\cdot d\cdot\frac{2^d}{2}\cdot\frac{1}{2}\exp(-4)=\frac{d}{4}\exp(-4).
\end{aligned}$$

This implies the existence of some $\boldsymbol{\theta}^*\in\Theta$ such that

$$\sum_{i=1}^{d}p_{\boldsymbol{\theta}^*,i}\geq\frac{d}{4}\exp(-4). \tag{24}$$

Choosing this $\boldsymbol{\theta}^*$ and defining the optimal arm $\boldsymbol{a}^*$ as:

$$\boldsymbol{a}^*=\underset{\boldsymbol{a}\in\mathcal{A}}{\arg\max}\left\langle\boldsymbol{a},\boldsymbol{\theta}^*\right\rangle=\underset{\boldsymbol{a}\in\mathcal{A}}{\arg\max}\sum_{i=1}^{d}\boldsymbol{a}_i^*\boldsymbol{\theta}_i^*.$$

It is easy to verify that to maximize $\sum_{i=1}^d\boldsymbol{a}_i^*\boldsymbol{\theta}_i^*$, we must have $a_i^*=\operatorname{sign}(\boldsymbol{\theta}_i^*)$ for all $i\in[d]$. Therefore, the expected regret is at least

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}^*}[R(T)]&=\mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{t=1}^{T}\left\langle\boldsymbol{a}^*-\boldsymbol{A}_t,\boldsymbol{\theta}^*\right\rangle\right]\\
&=\mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{t=1}^{T}\sum_{i=1}^{d}(\boldsymbol{a}_i^*-\boldsymbol{A}_{ti})\boldsymbol{\theta}_i^*\right]\\
&=\mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{t=1}^{T}\sum_{i=1}^{d}(\operatorname{sign}(\boldsymbol{\theta}_i^*)-\boldsymbol{A}_{ti})\boldsymbol{\theta}_i^*\right]\\
&=\mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{t=1}^{T}\sum_{i=1}^{d}2\mathbb{I}\big\{\operatorname{sign}(\boldsymbol{A}_{ti})\neq\operatorname{sign}(\boldsymbol{\theta}_i^*)\big\}\sqrt{\frac{1}{T}}\right]\\
&=2\sqrt{\frac{1}{T}}\sum_{i=1}^{d}\mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{t=1}^{T}\mathbb{I}\big\{\operatorname{sign}(\boldsymbol{A}_{ti})\neq\operatorname{sign}(\boldsymbol{\theta}_i^*)\big\}\right]\\
&\geq\sqrt{T}\sum_{i=1}^{d}\mathbb{P}_{\boldsymbol{\theta}^*}\left[\sum_{t=1}^{T}\mathbb{I}\big\{\operatorname{sign}(\boldsymbol{A}_{ti})\neq\operatorname{sign}(\boldsymbol{\theta}_i^*)\big\}\geq T/2\right] &(25)\\
&=\sqrt{T}\sum_{i=1}^{d}p_{\boldsymbol{\theta}^*,i}\geq\frac{\exp(-4)}{4}d\sqrt{T},
\end{aligned}$$

where Equation (25) uses Markov's inequality, and the last inequality follows from Equation (24). □

Theorem 3 (Regret lower bound)**.** *For any policy that chooses at most one key term per time step, there exists an instance of the conversational bandit problem such that the expected regret is at least $\Omega(\sqrt{dT})$. Furthermore, for any $T=2^m$ with $m\in[d]$, the regret is at least $\Omega(\sqrt{dT\log(T)})$.*

Proof. Suppose we have $\beta=\frac{d}{m}$ smaller problem instances $I_1.I_2,\dots,I_\beta$, each corresponding to an m-dimensional, K-armed bandit instance with a horizon of T/β and we assume they have preference vectors $\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_\beta\in\mathbb{R}^m$, respectively. We denote the arm set for instance I_j as $\mathcal{A}^{I_j}\subset\mathbb{R}^m$, and the regret incurred by instance I under policy π as $R_I^\pi(T)$. Next, we construct a d-dimensional instance $I=(I_1,I_2,\dots,I_\beta)$ by leting the unknown preference vector for instance I be $\boldsymbol{\theta}=(\boldsymbol{\theta}_1^\top,\dots,\boldsymbol{\theta}_\beta^\top)^\top$, and dividing the time horizon T into β consecutive periods, each of length T/β. For each time step $t\in[T]$, the feature vectors of arms $\mathcal{A}_t$ are constructed from instance I_j, where $j=\lceil t\beta/T\rceil$. Specifically, $\mathcal{A}_t=\big\{(\boldsymbol{0}^\top,\dots,\boldsymbol{x}^\top,\dots,\boldsymbol{0}^\top)^\top\big\}_{\boldsymbol{x}\in\mathcal{A}^{I_j}}$, where the non-zero entry is located at the j-th block. This means that at time t, the learner can only get information about the j-th block of the preference vector $\boldsymbol{\theta}$. Therefore for any policy π, there exists policies $\pi_1,\dots,\pi_\beta$ such that $R_I^\pi(T)=\sum_{j=1}^\beta R_{I_j}^{\pi_j}(\frac{T}{\beta})$. Applying Lemma 8, we can always find instances $I_1,I_2,\dots,I_\beta$ such that

$$\begin{aligned}
R_I^\pi(T)&=\sum_{j=1}^{\beta}R_{I_j}^{\pi_j}(\frac{T}{\beta})\geq\sum_{j=1}^{\beta}\Omega\left(m\sqrt{\frac{T}{\beta}}\right)=\Omega\left(m\sqrt{T\beta}\right)\\
&=\Omega\left(m\sqrt{T\frac{d}{m}}\right)=\Omega\left(\sqrt{dTm}\right)=\Omega\left(\sqrt{dT\log(T)}\right).
\end{aligned}$$ □

A.12 Technical Inequalities

We present the technical inequalities used throughout the proofs. We provide detailed references for readers' convenience.

Lemma 9 (Bretagnolle and Huber [3])**.** *Let P and Q be probability measures on the same measurable space $(\Omega,\mathcal{F})$, and let $A\in\mathcal{F}$ be an arbitrary event. Then,*

$$P(A)+Q(A^c)\geq\frac{1}{2}\exp(-D(P\parallel Q)),$$

where $D(P \parallel Q) = \int_\Omega \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P = \mathbb{E}_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right]$ *is the KL divergence between* P *and* Q*.* $A^c = \Omega \setminus A$ *is the complement of* A*.*

Lemma 10 (KL divergence between Gaussian distributions)**.** *If* $P \sim \mathcal{N}(\mu_1, \sigma^2)$ *and* $Q \sim \mathcal{N}(\mu_2, \sigma^2)$*, then*

$$D(P \parallel Q) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}.$$

Lemma 11 (Matrix Chernoff, Corollary 5.2 in Tropp [24])**.** *Consider a finite sequence* $\{\boldsymbol{X}_k\}$ *of independent, random, self-adjoint matrices with dimension* d*. Assume that each random matrix satisfies*

$$\boldsymbol{X}_k \succeq \boldsymbol{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \leq R \quad \text{almost surely.}$$

Define

$$\boldsymbol{Y} := \sum_k \boldsymbol{X}_k \quad \text{and} \quad \mu_{\min} := \lambda_{\min}(\mathbb{E}[\boldsymbol{Y}]) = \lambda_{\min}\left(\sum_k \mathbb{E}[\boldsymbol{X}_k]\right).$$

Then, for any $\delta \in (0, 1)$*,*

$$\Pr\left[\lambda_{\min}\left(\sum_k \boldsymbol{X}_k\right) \leq (1-\delta)\mu_{\min}\right] \leq d\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/R}.$$

Lemma 12 (Determinant-trace inequality, Lemma 10 in Abbasi-Yadkori et al. [1])**.** *Suppose* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_t \in \mathbb{R}^d$ *and for any* $1 \leq s \leq t$*,* $\|\boldsymbol{X}_s\|_2 \leq L$*. Let* $\overline{\boldsymbol{V}}_t = \lambda \boldsymbol{I} + \sum_{s=1}^t \boldsymbol{X}_s \boldsymbol{X}_s^\top$ *for some* $\lambda > 0$*. Then,*

$$\det(\overline{\boldsymbol{V}}_t) \leq \left(\lambda + \frac{tL^2}{d}\right)^d.$$