Distributed Multi-Task Learning for Stochastic Bandits With Context Distribution and Stage-Wise Constraints

Jiabin Lin and Shana Moothedath

***Abstract*—We present conservative multi-task learning in stochastic linear contextual bandits with *heterogeneous* agents. This extends conservative linear bandits to a distributed setting where M agents tackle *different but related* tasks while adhering to stage-wise performance constraints. The exact context is *unknown*, and only a context distribution is available to the agents as in many practical applications that involve a prediction mechanism to infer context, such as stock market prediction and weather forecast. We propose a distributed upper confidence bound (UCB) algorithm, DiSC-UCB. Our algorithm dynamically constructs a pruned action set for each task in every round, guaranteeing compliance with the constraints. Additionally, it includes synchronized sharing of estimates among agents via a central server using well-structured synchronization steps. For d-dimensional linear bandits, we prove an $\widetilde{O}(d\sqrt{MT})$ regret bound and an $O(M^{1.5}d^3)$ communication bound on the algorithm. We extend the problem to a setting where the agents are unaware of the baseline reward. We provide a modified algorithm, DiSC-UCB-UB, and show that it achieves the same regret and communication bounds. We empirically validated the performance of our algorithm on synthetic data and real-world Movielens-100 K and LastFM data and also compared it with some existing benchmark algorithms.**

***Index Terms*—Distributed learning, online learning, multi-arm bandits, constrained contextual bandits.**

I. Introduction

In CONTEXTUAL Bandits (CB), an agent engages in a series of interactions with an environment over multiple rounds. At the start of each round, the environment presents a context, and in response, the agent selects an action that yields a reward. The agent's objective is to choose actions to maximize cumulative reward over a time horizon of T. CB algorithms find applications in various fields, including robotics, clinical trials, communications, and recommender systems. This paper extends the standard CB problem in three ways.

First, the standard CB model assumes precise context observation, which does not always hold in real-world applications. For instance, contexts can be noisy measurements or predictions, like weather forecasting or stock market analysis [1]. In recommender systems, privacy constraints might limit access to certain user features, but we can infer a distribution over these features [2]. To address context uncertainty, motivated by the prior work [1], we study a scenario where the environment provides a context distribution. The exact context is treated as a sample from this distribution and is hidden. Second, given the increasing demand for safe learning in various real-world systems, particularly those with safety-critical applications, this paper delves into the impact of stage-wise safety constraints on the linear stochastic bandit problem. Drawing inspiration from prior works [3], [4], [5], our approach builds upon the safety constraint introduced by [6] and later studied in [7]. In our scenario, the agent has a baseline policy suggesting actions with guaranteed expected rewards derived from historical data or legacy policies. We enforce a safety constraint that requires the agent's chosen actions to yield expected rewards no less than a predetermined fraction of those recommended by the baseline policy. Third, multi-task learning enables models to simultaneously tackle multiple related tasks, leveraging common patterns and improving overall performance [8], [9], [10], [11], [12]. By sharing knowledge across tasks, multi-task learning can lead to more efficient and effective models, especially when data is limited or expensive to acquire. Multi-task bandit learning has gained interest recently [13], [14], [15], [16], [17], [18], [19], [20]. In this paper, we consider heterogeneous multi-task linear stochastic CB problem with hidden contexts and stage-wise performance constraints. A set of M agents collaborate to solve related but different tasks jointly; while the exact contexts are hidden, only a context distribution is known to the agents, and the agents are subject to performance constraints at every decision round. The problem addressed in this work applies to various bandit learning scenarios, such as recommender systems and personalized medical treatments, where tasks are related but not identical. While movies and TV shows share features, some may be irrelevant across domains. Additionally, context (e.g., user or patient data) can be noisy due to profile errors or shared accounts. Often, a baseline policy guides recommendations based on past data, but platforms may impose constraints, ensuring daily viewership remains above a threshold to maintain revenue stability. These applications can significantly benefit from our approach, as demonstrated in our empirical analysis in Section VI.

Received 13 September 2024; revised 11 February 2025 and 7 April 2025; accepted 17 April 2025. Date of publication 1 May 2025; date of current version 20 June 2025. The associate editor coordinating the review of this article and approving it for publication was Dr. Stefan Vlaski. *(Corresponding author: Jiabin Lin.)*

The authors are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: jiabin@iastate.edu; mshana@iastate.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TSIPN.2025.3566239, provided by the authors.

Digital Object Identifier 10.1109/TSIPN.2025.3566239

A. Our Contributions

The paper makes four key contributions.

1) We formulate the constrained multi-task contextual bandit problem with hidden contexts and propose a distributed UCB algorithm. A key aspect of our approach is the construction of a safe action set at each learning round to filter out actions that fail to meet performance constraints, which is rather challenging since the contexts are unknown. The unsafe action elimination method in existing work [7], which assumes known contexts, is not directly applicable to our setting with unknown contexts. Lemma 3 proves that our safe action set meets the performance constraints, while Lemma 5 proves that the optimal action in each round remains within the pruned set and is not eliminated.
2) We show in Theorem 1 that the regret can be decomposed into three terms: (i) an upper bound for the regret of the distributed linear UCB algorithm, (ii) a term that captures the loss since the contexts are unknown and (iii) a term that accounts for the loss for being conservative to satisfy the performance constraint. A key aspect of our analysis is bounding the number of rounds the learner's actions (N_T) and the conservative actions (N_T^c) are played, which we establish in Theorem 2.
   In Theorem 3, we provide an $\widetilde{O}(d\sqrt{MT})$ regret bound and an $O(M^{1.5}d^3)$ communication bound on the algorithm, where d is the dimension of the feature vector. Our proposed approach significantly outperforms the naive method of solving the M tasks independently, which results in a regret of $\widetilde{O}(dM\sqrt{T})$ [7], [21], thereby demonstrating the effectiveness of multi-task learning.
3) We extend the problem to a setting where the agents are unaware of the baseline reward value. We show that the algorithm can be modified slightly to address this case, and regret and communication bounds remain the same.
4) We conducted numerical experiments using both synthetic and real-world datasets (MovieLens and LastFM) to validate our algorithms. Our methods consistently outperformed benchmark algorithms across all evaluations.

B. Related Work

Multi-task bandit learning: Collaborative bandit learning has been explored in both homogeneous and heterogeneous settings [10], [11], [12], [22], [23], [24], [25], [26]. In the homogeneous setting, all tasks are identical, whereas in the heterogeneous setting, tasks are distinct yet share underlying similarities. In the homogeneous setting, distinct agents solve a common task collaboratively using a shared reward parameter $\theta^\star$ [10], [11], [23], [25], [27]. Our work, on the contrary, introduces a multi-task bandit learning in which heterogeneous agents perform distinct tasks governed by heterogeneous reward parameters $\Theta = \{\theta_i^\star\}_{i\in[M]}$. This model finds relevance in various real-world scenarios where agents, possessing *distinct yet interrelated objectives*, function within a collaborative setting. [12], [22], [24] considered tasks with unique feature vectors. However, their scenario treats each agent as a user/context, resulting in a time-invariant situation. In contrast, our work addresses a time-varying case. Another related line of work is multi-task representation learning of bandits [13], [14], [15], [16], [17], [19], [20] that focuses on learning a shared low-dimensional representation between the tasks. Bandit problems with sparsity constraints are studied in [18], [28]. None of these works consider constrained learning or unobserved contexts. Contextual bandits with partial observations have been studied in [1], [10], [11]. A related line of work is partial monitoring, where the rewards for the chosen actions are unobserved [29], [30], [31]. Instead, the learner receives indirect feedback that correlates with the rewards.

Constrained bandits: Constrained bandits have been well studied, including [32], [33], [34], [35], [36], [37] under various modeling assumptions. The two primary types of constraints are: (i) budget constraints, where each arm incurs a random resource consumption, and (ii) safety/performance constraints, which ensure that the reward in each round meets or exceeds a specified fraction of a baseline performance.

Refs. [3], [4], [37] considered a cumulative performance constraint, which ensures that the total reward over the learning horizon meets or exceeds a specified threshold, in contrast to stage-wise constraints that must be satisfied in every round. [32] considered linear constraint of the form $x^\top B\theta^\star \leqslant C$, where $B, C > 0$ are a known matrix and a positive constant, respectively. The most related works are [6], [7], which also considered stage-wise performance constraints, however, in a single task setting with observable contexts. In the multi-task setting considered in our paper, each task is associated with a distinct constraint, resulting in M constraints. Constrained Markov Decision Process (CMDP) extends traditional MDPs by incorporating a cost function to ensure compliance with constraints [38]. The safety specifications in the majority of the CMDP methods are expected discounted cumulative costs, and the goal is to maximize the expected performance while minimizing the cost incurred from constraints. Constrained restless bandits, as studied by [39] addresses scenarios where resource availability varies over time, necessitating adaptive strategies to optimize decision-making under dynamic constraints.

II. Stochastic Linear Contextual Bandits

In this section, we will first introduce the standard linear bandit, then the stochastic stage-wise constrained setting, and finally, the distributed stochastic stage-wise constrained setting. Let $\mathcal{A}$ be the action set and $\mathcal{C}$ be the context set.

Stochastic linear CBs: In linear bandits, at round $t \in \mathbb{N}$, the agent observes a context $c_t \in \mathcal{C}$ and selects an action $x_t \in \mathcal{A}$. The context-action pair (x_t, c_t) is mapped onto a feature vector $\phi_{x_t,c_t} \in \mathbb{R}^d$. Based on this feature vector ϕ_{x_t,c_t}, the agent receives a reward $y_t \in \mathbb{R}$ from the environment, where $y_t = \phi_{x_t,c_t}^\top \theta^\star + \eta_t$, with $\theta^\star$ representing an unknown reward parameter, and η_t is a $\sigma-$Gaussian noise with zero mean. Thus the expected reward $r(x_t, c_t) = \mathbb{E}[y_t]$. The goal of the linear bandit problem is to maximize the cumulative reward or, equivalently, minimizes the cumulative (pseudo) regret [21], as

$$\min \mathcal{R}_T = \sum_{t=1}^{T} \phi_{x_t^\star, c_t}^\top \theta^\star - \sum_{t=1}^{T} \phi_{x_t, c_t}^\top \theta^\star. \quad (1)$$

Here $x_t^\star$ is the optimal/best action for context c_t, and x_t is the action chosen by the agent for context c_t. [21], [40] showed that the UCB policy achieves sublinear regret $\widetilde{O}(d\sqrt{T})$.

Stochastic linear CBs with stage-wise constraints: Here, the agent is provided with a baseline policy. The baseline action x_{b_t} has an expected reward $r_{b_t} = \phi_{x_{b_t}, c_t}^\top \theta^\star$. The agent's actions are subject to the stage-wise conservative constraint $r_t = \phi_{x_t, c_t}^\top \theta^\star \geqslant (1-\alpha) r_{b_t}$, where $\alpha \in [0,1]$, such that the agent's action will be chosen only if it satisfies the constraint, otherwise, the baseline action will be performed [6], [7]. This guarantees that the expected reward for the agent at any round t remains at least a pre-defined fraction $(1-\alpha)$ of the baseline reward. The goal is to maximize the cumulative reward while satisfying the stage-wise constraint. Formally, minimize the cumulative (pseudo) regret while meeting the constraint

$$\min \mathcal{R}_T \text{ such that } \phi_{x_t, c_t}^\top \theta^\star \geqslant (1-\alpha) r_{b_t}, \text{for all } t \in [T]. \quad (2)$$

For this setting, [6] presented an algorithm, referred to as SEGE with regret bound $\widetilde{O}(d\sqrt{T})$ when the actions are constrained to be from an ellipsoid, and [7] presented a UCB algorithm, referred to as SCLUCB, and a Thompson sampling algorithm, referred to as SCLTS, with regret bounds $\widetilde{O}(d\sqrt{T})$.

III. Problem Formulation and Notation

Notations: The norm of a vector $z \in \mathbb{R}^d$ with respect to a matrix $V \in \mathbb{R}^{d \times d}$ is defined as $\|z\|_V := \sqrt{z^\top V z}$. Further, $\top$ denotes matrix or vector transpose. $\mathcal{F}_t = (\mathcal{F}_1, \sigma(x_1, \eta_1, \cdots x_t, \eta_t))$ be the filtration ($\sigma-$algebra) that represents the information up to round t. For an integer Z, we denote $[Z] = \{1, 2, \ldots, Z\}$.

Problem Formulation: In this work, we study multi-task stochastic linear CBs with stage-wise constraints and context distribution. We consider a set of M heterogeneous agents performing *different but related tasks* such that their reward parameters $\theta_i^\star \in \mathbb{R}^{d_i}$, for $i \in [M]$, satisfy a sparse structural constraint. Our problem models a joint multi-tasking bandit problem, where the set $\Theta = \{\theta_1^\star, \theta_2^\star, \ldots, \theta_M^\star\}$ satisfies the sparse structural constraints described below.

Constraint 1: (Sparse structural constraint on Θ) We define an index set for tasks $\mathcal{I} = \{I_1, I_2, \ldots, I_M\}$, where $|I_i| = d_i$. For $i \in \{1, \ldots, M\}$, I_i is the index set of features that are relevant to agent i. Without loss of generality, we assume $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_M$. Thus, $I_i = \{k \in \{1, 2, \ldots, d_1\} : k^{\text{th}} \text{ feature of } \theta_1^\star \in \theta_i^\star\}$, where $\theta_i^\star := \theta_1^\star|_{I_i}$.

Further, we consider that at round t, the context c_t is unobservable and only a distribution of the context denoted as the agents observe μ_t. At round t, the environment chooses a distribution $\mu_t \in \mathcal{P}(\mathcal{C})$ over the context set and samples a context realization $c_t \sim \mu_t$. The agents observe only μ_t and not c_t and each agent selects an action, say action chosen by agent i is $x_{t,i}$, and receive reward $y_{t,i}$, where $y_{t,i} = \phi_{x_{t,i}, c_t}^\top \theta_i^\star + \eta_{t,i}$. Moreover, each agent possesses a baseline policy, $\pi_{b_t,i}$, shaped by their past experiences and domain expertise. This baseline policy guides the derivation of a baseline action $x_{b_t,i}$ aligned with the specific context c_t. This interplays with the performance constraint, wherein each agent i is bound to select an action meeting the condition defined $\pi_{b_t,i}$ as described below.

Constraint 2: (Performance constraint on agent i w.r.t $\pi_{b_t,i}$) Given a context c_t and the baseline policy $\pi_{b_t,i}$, agent i can select an action $x_{t,i}$ only if $\phi_{x_{t,i}, c_t}^\top \theta_i^\star \geqslant (1-\alpha) r_{b_t,i}$, where $r_{b_t,i}$ is the baseline action derived from $\pi_{b_t,i}$ for c_t.

Our aim is to maximize the cumulative reward, $\sum_{i=1}^{M} \sum_{t=1}^{T} y_{t,i}$, while simultaneously satisfying constraints 1 and 2. Formally, our aim is to minimize the cumulative regret

$$\mathcal{R}_T = \sum_{i=1}^{M} \sum_{t=1}^{T} (\phi_{x_{t,i}^\star, c_t} - \phi_{x_{t,i}, c_t})^\top \theta_i^\star \text{ s.t constraints1 \& 2 hold.} \quad (3)$$

Here $x_{t,i}^\star = \arg\max_{x_{t,i} \in \mathcal{A}} \mathbb{E}_{c \sim \mu_t}[r_{x_{t,i}, c}]$ is the best action provided we know μ_t, but not c_t, and T is the total number of rounds, and $(1-\alpha) \in (0,1)$ is the maximum fraction of loss in the performance compared to the baseline policy the decision maker is willing to accept during learning. Let $\kappa_{b_t,i} = \phi_{x_{t,i}^\star, c_t}^\top \theta_i^\star - r_{b_t,i}$ be the difference between the expected reward of the optimal action $x_{t,i}^\star$ and the baseline action $x_{b_t,i}$ at round t.

To solve the multi-task problem, we first transform the problem as follows. We transform the distributed multi-task problem with feature vector $\phi_{x_{t,i}, c_t} \in \mathbb{R}^{d_i}$ and heterogeneous reward parameters $\theta_i^\star \in \mathbb{R}^{d_i}$ for each agent into a distributed linear bandit model with the heterogeneous feature vectors $\phi_i(x_{t,i}, c_t) \in \mathbb{R}^d$ and shared reward parameter $\theta^\star$. We perform the mapping of the feature vector $\phi_{x_{t,i}, c_t} \in \mathbb{R}^{d_i}$ to a new feature vector $\phi_i(x_{t,i}, c_t) \in \mathbb{R}^{d_1}$. Henceforth for notational brevity, we define $d := d_1$ after dropping the subscript. This is achieved by retaining the features corresponding to the index set I_i and setting the value to zero for features not in the index set I_i resulting a d-dimensional feature vector $\phi_i(x_{t,i}, c_t)$. We elaborate on this process in the example given in Section A. Henceforth, our analysis use heterogeneous features and shared $\theta^\star$.

Assumption 1: Each element $\eta_{t,i}$ of the noise sequence $\{\eta_{t,i}\}_{t=1,i=1}^{\infty,M}$ is conditionally $\sigma-$subGaussian, i.e., $\mathbb{E}[e^{\lambda \eta_{t,i}} | \mathcal{F}_{t-1}] \geqslant exp(\frac{\lambda^2 \sigma^2}{2})$, for all $\lambda \in \mathbb{R}$.

Assumption 2: For simplicity, we assume $\phi_{x_{t,i}, c_t}^\top \theta_i^\star \in [0,1]$, $\|\phi_{x_{t,i}, c_t}\|_2 \leqslant 1$, and $\|\theta_i^\star\|_2 \leqslant 1$, for all $i \in [M]$ and all $x \in \mathcal{A}$.

Assumption 3: There exist constants $\kappa_l \geqslant \kappa_h \geqslant 0, r_h \geqslant r_l \geqslant 0$ such that at each round t, $\kappa_l \leqslant \kappa_{b_t,i} \leqslant \kappa_h$, $r_l \leqslant r_{b_t,i} \leqslant r_h$.

IV. Distributed Stage-Wise Contextual Bandits With Context Distribution

A. Proposed Algorithm

In this section, we introduce our proposed UCB algorithm, referred to as *distributed stage-wise contextual bandits with context distribution* algorithm (DiSC-UCB). We present the pseudocode in Algorithm 1.

Given the distribution μ_t, as in [1] we first construct the heterogeneous feature vectors $\Psi_{t,i} = \{\psi_i(x, \mu_t) : x \in \mathcal{A}\}$, where

$\{\psi_i(x,\mu_t) := \mathbb{E}_{c\sim\mu_t}[\phi_i(x,c_t)]\}$ is the expected feature vector of action x under μ_t. We use $\Psi_{t,i}$ as the feature set at round t. DiSC-UCB is based on the optimization in the face of the uncertainty principle. In each round $t \in [T]$, each agent $i \in [M]$ maintains confidence set $\mathcal{B}_{t,i} \subseteq \mathbb{R}^d$ that contains the unknown reward parameter $\theta^\star$ with high probability and constructs a pruned action set $\mathcal{X}_{t,i}$ by eliminating a subset of the actions that violate the constraints. After $\mathcal{X}_{t,i}$ is determined, agents derive the corresponding pruned feature set $\Xi_{t,i} = \{\psi_i(x,\mu_t) : x \in \mathcal{X}_{t,i}\}$, which is a subset of $\Psi_{t,i}$. Each agent then chooses an optimistic estimate $\tilde{\theta}_{t,i} \in \arg\max_{\theta\in\mathcal{B}_{t,i}}(\max_{x\in\mathcal{X}_{t,i}} \psi_i(x,\mu_t)^\top\theta))$ and chooses an action $x'_{t,i} \in \arg\max_{x\in\mathcal{X}_{t,i}} \psi_i(x,\mu_t)^\top\theta$. Equivalently, the agent chooses $(x'_{t,i},\tilde{\theta}_{t,i}) \in \arg\max_{(x,\theta)\in\mathcal{X}_{t,i}\times\mathcal{B}_{t,i}} \psi_i(x,\mu_t)^\top\theta$.

If the optimization is feasible, the agents choose their respective optimistic action $x'_{t,i}$ and get the feature vector $\psi_i(x'_{t,i},\mu_t)$ under a certain condition; otherwise, the agents choose their baseline action $x_{b_t,i}$ and get the conservative feature vector $\psi_i(x_{t,i},\mu_t) = (1-\rho)\psi_i(x_{b_t,i},\mu_t) + \rho\zeta_{t,i}$. This approach, introduced in [6] and later applied in [7], [41], ensures the agents learn even when not satisfying the performance constraint in a round. We assume the agents know the baseline action $x_{b_t,i}$ and its corresponding expected reward $r_{b_t,i}$ [6], [7], [41]. Here, α is a known parameter, similar to [3] and [4]. After receiving their reward $y_{t,i}$, agents update their local parameters, which are then used as the basis for updating their respective confidence set and pruned action sets.

In the synchronization phase, at predetermined time intervals, agents exchange all the latest gathered estimates. We refer to the rounds between these synchronization points as *epochs*. Such a synchronization method was introduced in [23], designed based on the observation in [21] that the change in the determinant of $\bar{V}_{t,i} = \lambda I + \psi_i(x_{t,i},\mu_t)\psi_i(x_{t,i},\mu_t)^\top$ is a good indicator of the learning progress. Specifically, synchronization happens only when agent i recognizes that the log-determinant of $\bar{V}_{t,i}$ has changed by more than a constant factor since the last synchronization. This method reduces the communication cost of the algorithm. Next, we will explain the construction of the confidence set and the pruned action set.

Construction of the Confidence Set $\mathcal{B}_{t,i}$: After obtaining the estimate $\hat{\theta}_{t,i}$ of the unknown parameter $\theta^\star$, we construct the confidence set $\mathcal{B}_{t,i}$ as follows.

$$\mathcal{B}_{t,i} = \left\{\theta \in \mathbb{R}^d : \|\hat{\theta}_{t,i} - \theta\|_{\bar{V}_{t,i}} \leqslant \beta_{t,i}\right\}, \text{ where}$$
$$\beta_{t,i} = \beta_{t,i}(\sigma,\delta) = \sigma\sqrt{2\log\left(\frac{\det(\bar{V}_{t,i})^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2}$$
$$\text{and } \bar{V}_{t,i} = \lambda I + W_{t,i}, \hat{\theta}_{t,i} = \bar{V}_{t,i}^{-1}U_{t,i}. \tag{4}$$

In Lemma 1 we show that by setting $\beta_{t,i} = \beta_{t,i}(\sqrt{1+\sigma^2},\delta/2)$, we can construct the confidence set $\mathcal{B}_{t,i}$ such that the reward parameter $\theta^\star$ will always be contained within the confidence set $\mathcal{B}_{t,i}$ with a high probability.

Construction of Pruned Action Set $\mathcal{X}_{t,i}$: In standard constrained bandit settings [7], verifying whether an action satisfies the constraints is straightforward because the context is known. As a result, the safe action set is typically constructed by directly checking each action against the constraints and eliminating those that fail to meet them. However, in our setting, the context is unknown, making such a direct approach to constructing the safe action set infeasible. Since the agents only observe the context distribution and the exact contexts are unknown,

Algorithm 1: Distributed Stage-wise Contextual Bandits with Context Distribution (DiSC-UCB).

```
Initialization: B = (T log MT / dM), λ = 1, W_syn = 0, U_syn = 0, W_{t,i} = 0, U_{t,i} = 0, t_last = 0, V_last = λI
for t = 1, 2, ..., T do
  Nature chooses μ_t ∈ P(C) and agent observes μ_t
  Set Ψ_{t,i} = {ψ_i(x, μ_t) : x ∈ A} where
  {ψ_i(x, μ_t) := E_{c~μ_t}[φ_i(x, c_t)]} for each agent i
  for Agent i = 1, 2, ..., M, do
    V̄_{t,i} = λI + W_syn + W_{t,i}, θ̂_{t,i} = V̄_{t,i}^{-1}(U_syn + U_{t,i})
    Construct confidence set B_{t,i} using V̄_{t,i}, θ̂_{t,i}
    Compute pruned action set X_{t,i} using ψ_i(x, μ_t), θ̂_{t,i}
    Construct feature set Ξ_{t,i} = {ψ_i(x, μ_t) : x ∈ X_{t,i}}
    if the following optimization is feasible:
    (x'_{t,i}, θ̃_{t,i}) = argmax_{(x,θ)∈X_{t,i}×B_{t,i}} ⟨ψ_i(x, μ_t), θ⟩ then
      Set F = 1, else F = 0
    end if
    if F = 1 and λ_min(V̄_{t,i}) ⩾ (2β_{t,i} / (α r_{b_t,i}))^2 then
      Choose x_{t,i} = x'_{t,i}, get the feature vector ψ_i(x'_{t,i}, μ_t), and receive the reward y_{t,i}
    else
      Choose x_{t,i} = x_{b_t,i}, get conservative feature vector
      ψ_i(x_{t,i}, μ_t) = (1 − ρ)ψ_i(x_{b_t,i}, μ_t) + ρζ_{t,i} and y_{b_t,i}
    end if
    Update U_{t,i} = U_{t,i} + ψ_i(x_{t,i}, μ_t)y_{t,i},
    W_{t,i} = W_{t,i} + ψ_i(x_{t,i}, μ_t)ψ_i(x_{t,i}, μ_t)^⊤,
    V_{t,i} = λI + W_syn + W_{t,i}
    if log(det(V_{t,i})/det(V_last)) · (t − t_last) ⩾ B then
      Send a synchronization signal to the server to start a communication round
    end if
    Synchronization round:
    if a communication round is started then
      All agents i ∈ [M] send W_{t,i}, U_{t,i} to server
      Server computes W_syn = W_syn + Σ_{i=1}^M W_{t,i}, U_syn = U_syn + Σ_{i=1}^M U_{t,i}
      All agents receive W_syn, U_syn from server
      Set W_{t,i} = U_{t,i} = 0, t_last = t, for all i,
      V_last = λI + W_syn
    end if
  end for
end for
```

we need to use estimated feature maps to construct a feasible action set. We now present our approach to tackling this issue by constructing a pruned action set, $\mathcal{X}_{t,i}$, to eliminate unsafe actions.

In every iteration, each agent refines its action set by excluding actions that fail to satisfy the baseline condition. This is further complicated by the unknown nature of the actual context, rendering the utilization of the feature vector $\phi_i(x_{t,i}, c_t)$ impractical for constructing the pruned action set. Our approach utilizes $\hat{\theta}_{t,i}$, $\bar{V}_{t,i}$, and $\psi_i(x_{t,i}, \mu_t)$. The pruned action set aims to eliminate actions violating the baseline constraint, necessitating criteria for excluding unsafe actions. While constructing the pruned action set, we analyze two cases. 1) $\phi_i(x_{t,i}, c_t)^\top \hat{\theta}_{t,i} \geqslant \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i}$ and 2) $\phi_i(x_{t,i}, c_t)^\top \hat{\theta}_{t,i} \leqslant \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i}$. We address case 1) first and explain the process of constructing a subset of actions that satisfy the constraint for all $v \in \mathcal{B}_{t,i}$. We provide alternate sufficient conditions for deriving the safe action set. We use the symbol '$\Leftarrow$' to show that a certain condition implies another. Define

$$\mathcal{X}^1_{t,i} := \{x_{t,i} \in \mathcal{A} : \phi_i(x_{t,i}, c_t)^\top v \geqslant (1-\alpha) r_{b_t,i}, \forall\, v \in \mathcal{B}_{t,i}\} \tag{5}$$

$$\begin{aligned}&\Leftarrow \{x_{t,i} \in \mathcal{A} : \phi_i(x_{t,i}, c_t)^\top (v - \hat{\theta}_{t,i}) + \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i}\\ &+(\phi_i(x_{t,i}, c_t) - \psi_i(x_{t,i}, \mu_t))^\top \hat{\theta}_{t,i} \geqslant (1-\alpha) r_{b_t,i}, \forall v \in \mathcal{B}_{t,i}\}\\ &\Leftarrow \{x_{t,i} \in \mathcal{A} : \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i} \geqslant \frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}} + (1-\alpha) r_{b_t,i}\}\end{aligned} \tag{6}$$

where the last step follows from $\phi_i(x_{t,i}, c_t)^\top \hat{\theta}_{t,i} \geqslant \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i}$ and $\phi_i(x_{t,i}, c_t)^\top (v - \hat{\theta}_{t,i}) \geqslant -\frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}}$ from Lemma 2. All actions that meet the conditions in (6) also fulfill the requirements of (5), thus ensuring safety.

Now we consider case 2), where $\phi_i(x_{t,i}, c_t)^\top \hat{\theta}_{t,i} \leqslant \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i}$. In this case, our approach is to first identify actions that violate the baseline constraint, $\bar{\mathcal{X}}^2_{t,i}$, and then eliminate those actions from the action set $\mathcal{A}$.

$$\bar{\mathcal{X}}^2_{t,i} := \{x_{t,i} \in \mathcal{A} : \phi_i(x_{t,i}, c_t)^\top v \leqslant (1-\alpha) r_{b_t,i}, \forall v \in \mathcal{B}_{t,i}\} \tag{7}$$

$$\begin{aligned}&\Leftarrow \{x_{t,i} \in \mathcal{A} : \phi_i(x_{t,i}, c_t)^\top (v - \hat{\theta}_{t,i})\\ &+\phi_i(x_{t,i}, c_t)^\top \hat{\theta}_{t,i} \leqslant (1-\alpha) r_{b_t,i}, \forall v \in \mathcal{B}_{t,i}\}\\ &\Leftarrow \{x_{t,i} \in \mathcal{A} : \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i} \leqslant \frac{-\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}} + (1-\alpha) r_{b_t,i}\}\end{aligned} \tag{8}$$

where the last step follows from $\phi_i(x_{t,i}, c_t)^\top \hat{\theta}_{t,i} \leqslant \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i}$ and $\phi_i(x_{t,i}, c_t)^\top (v - \hat{\theta}_{t,i}) \leqslant \frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}}$ from Lemma 2. Note that, all actions that meet the conditions in (8) also fulfill the requirements of 7, consequently rendering them unsafe. By taking the difference between $\mathcal{A}$ and $\bar{\mathcal{X}}^2_{t,i}$, we determine $\mathcal{X}^2_{t,i} = \mathcal{A} \setminus \bar{\mathcal{X}}^2_{t,i}$

$$= \{x_{t,i} \in \mathcal{A} : \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i} \geqslant \frac{-\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}} + (1-\alpha) r_{b_t,i}\}.$$

Given $\mathcal{X}^1_{t,i}$ and $\mathcal{X}^2_{t,i}$, we obtain the pruned action set by taking the intersection between $\mathcal{X}^1_{t,i}$ and $\mathcal{X}^2_{t,i}$, given by

$$\mathcal{X}_{t,i} = \left\{ x_{t,i} \in \mathcal{A} : \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i} \geqslant \frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}} + (1-\alpha) r_{b_t,i} \right\}.$$

At round t, each agent chooses a pair $(x'_{t,i}, \tilde{\theta}_{t,i})$, where $x'_{t,i} \in \mathcal{X}_{t,i}$ and $\tilde{\theta}_{t,i} \in \mathcal{B}_{t,i}$ that jointly maximizes the current reward while ensuring that the baseline constraint is met. That is,

$$(x'_{t,i}, \tilde{\theta}_{t,i}) = \underset{(x,\theta) \in \mathcal{X}_{t,i} \times \mathcal{B}_{t,i}}{\arg\max} \langle \psi_i(x, \mu_t), \theta \rangle .$$

The pruned action set $\mathcal{X}_{t,i}$ is constructed by considering all $\theta \in \mathcal{B}_{t,i}$, not just $\theta^\star$. If the pruned action set is non-empty and satisfies a constraint $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_t,i}})^2$, the agent chooses $x'_{t,i}$ and get the feature vector $\psi_i(x'_{t,i}, \mu_t)$; otherwise, the agent chooses the baseline action $x_{b_t,i}$ and get the conservative feature vector $\psi_i(x_{t,i}, \mu_t) = (1-\rho)\psi_i(x_{b_t,i}, \mu_t) + \rho \zeta_{t,i}$, which is detailed below. We consider the constraint on $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_t,i}})^2$, as it ensures that, with high probability, the best action $x^\star_{t,i}$ always belongs to the pruned action set. We will provide detailed proof for this claim later in Lemma 5. A similar condition on the smallest eigenvalue of the Gram matrix is used in [7], yet our proof approach is different from [7] due to the reliance on known feature vectors (contexts), which contrasts with our scenario. We note that since $\bar{\mathcal{X}}^2_{t,i}$ is a subset of the actions that are unsafe, $\mathcal{X}^2_{t,i}$ may contain some unsafe actions. We will demonstrate in Lemma 3 that when the agent's action is played, the chosen action from the pruned set will not violate the baseline constraint with high probability.

Conservative feature vector: In our problem, each agent i is assigned a baseline policy, and in each round t, a baseline action $x_{b_t,i}$ is recommended based on that policy. The agent's objective is to carry out explorations, ensuring that the rewards achieved from exploratory action $x_{t,i}$ remain reasonably comparable to the rewards from the baseline action. Our approach draws inspiration from [6] and [7], where the conservative feature vectors for the baseline actions are combined with random exploration while maintaining adherence to stage-wise safety constraints. We construct a conservative feature vector $\psi_i(x_{t,i}, \mu_t)$ to be a convex combination of the baseline action's feature vector $\psi_i(x_{b_t,i}, \mu_t)$ and a random noise vector $\zeta_{t,i}$, expressed as $\psi_i(x_{t,i}, \mu_t) = (1-\rho)\psi_i(x_{b_t,i}, \mu_t) + \rho \zeta_{t,i}$, where $\zeta_{t,i}$ is a sequence of independent random vectors with zero means, and we assume that $\|\zeta_{t,i}\|_2 = 1$. ρ is a constant value in the range $(0, \frac{\alpha r_l}{1+r_h}]$, and we can ensure that when $\rho \in (0, \frac{\alpha r_l}{1+r_h}]$, the conservative feature vector is always in the feature vector set $\Xi_{t,i}$ (Lemma 4). By identifying the conservative feature vector through this convex combination, we can guarantee that the agent continues to learn in every round.

B. Theoretical Analysis on Safety Guarantees

In this section, we establish safety guarantees for the DiSC-UCB algorithm. We begin with two preliminary results in Lemmas 1 and 2. Then, in Lemmas 3, 4, and 5, we prove that the pruned safe action set satisfies the performance constraints while ensuring the optimal actions are preserved.

Lemma 1: For any $\delta > 0$, with a probability of $1 - M\delta$, $\theta^\star$ will always exist inside the confidence set $\mathcal{B}_{t,i}$ defined by (4) where $\beta_{t,i} = \beta_{t,i}(\sqrt{1+\sigma^2}, \delta/2)$ for all value of t and i.

Proof: Drawing inspiration from Theorem 1 in [1], we implement a similar approach to analyze the reward $y_{t,i}$, In particular, we observe that the reward $y_{t,i} = \phi_i(x_{t,i}, c_t)^\top \theta^\star + \eta_{t,i}$ can be alternatively represented as

$$y_{t,i} = \psi_i(x_{t,i}, \mu_t)^\top \theta^\star + \xi_{t,i} + \eta_{t,i},$$

where $\xi_{t,i} = (\phi_i(x_{t,i}, c_t) - \psi_i(x_{t,i}, \mu_t))^\top \theta^\star$. In this representation, $(\xi_{t,i} + \eta_{t,i})$ serves as the noise component associated with $\psi_i(x_{t,i}, \mu_t)^\top \theta^\star$. Given that $|\xi_{t,i}| \leqslant 1$, $\xi_{t,i}$ is a 1-subgaussian. Therefore, $y_{t,i}$ can be viewed as an observation of reward obtained from $\psi_i(x_{t,i}, \mu_t)^\top \theta^\star$ with the presence of noise parameterized by $\sqrt{1+\sigma^2}$. Thus, we can define the confidence bound for the least squares estimator of $\psi_i(x_{t,i}, \mu_t)$ while guaranteeing that $\theta^\star$ is always present within it with probability $1-\delta$, using updated parameters. The parameter ρ for $\beta_{t,i}$ is given by $\sqrt{1+\sigma^2}$. Further, when considering the upper bound of cumulative regret, we use the Azuma-Hoeffding inequality to determine the upper bound of the regret gap term from the context distribution with probability $1-\delta$. By using the union bound, the probabilities can be combined, providing a resulting probability of $1 - 2\delta$. We propose a substitution such that $\delta' = 2\delta$. Therefore, the parameter δ for our $\beta_{t,i}$ is changed to $\frac{\delta'}{2}$. Finally, our proof is completed by considering the presence of M agents and using the union bound.

According to Lemma 1, with a probability of $1 - M\delta$, $\theta^\star$ is always included inside the confidence set $\mathcal{B}_{t,i}$ for all values of t and i. Thus, Lemma 1 guarantees that for each round t, $\theta^\star \in \mathcal{B}_{t,i}$ holds for every agent i with probability at least $1 - M\delta$. We have the following result for Algorithm 1.

Lemma 2: In the DiSC-UCB algorithm, Algorithm 1, all values of $v \in \mathcal{B}_{t,i}$, satisfy the following two inequalities.

$$-\frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}} \leqslant \phi_i(x_{t,i}, c_t)^\top (\hat{\theta}_{t,i} - v) \leqslant \frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}} \text{ and}$$

$$-\frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}} \leqslant \psi_i(x_{t,i}, \mu_t)^\top (\hat{\theta}_{t,i} - v) \leqslant \frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}}.$$

Proof: We begin by noticing that $\bar{V}_{t,i}$ is a symmetric semi-positive definite matrix thus, it can be decomposed by eigenvalues. From this, we can derive the inequality as follows.

$$\|\phi_i(x_{t,i}, c_t)\|_{\bar{V}_{t,i}^{-1}} = \sqrt{\phi_i(x_{t,i}, c_t)^\top \bar{V}_{t,i}^{-1} \phi_i(x_{t,i}, c_t)}$$
$$\leqslant \sqrt{\phi_i(x_{t,i}, c_t)^\top (Q\Sigma Q^\top)^{-1} \phi_i(x_{t,i}, c_t)}$$
$$\leqslant \sqrt{\frac{\phi_i(x_{t,i}, c_t)^\top Q Q^\top \phi_i(x_{t,i}, c_t)}{\lambda_{\min}(\bar{V}_{t,i})}}$$
$$= \sqrt{\frac{\phi_i(x_{t,i}, c_t)^\top \phi_i(x_{t,i}, c_t)}{\lambda_{\min}(\bar{V}_{t,i})}} = \frac{\|\phi_i(x_{t,i}, c_t)\|_2}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}} \leqslant \frac{1}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}}.$$

Then, by using the Cauchy–Schwarz inequality, we can write

$$\phi_i(x_{t,i}, c_t)^\top (\hat{\theta}_{t,i} - v)$$
$$\leqslant \|\phi_i(x_{t,i}, c_t)\| \|\hat{\theta}_{t,i} - v\|$$
$$= \|\bar{V}_{t,i}^{1/2} \phi_i(x_{t,i}, c_t)\|_{\bar{V}_{t,i}^{-1}} \|\bar{V}_{t,i}^{-1/2} (\hat{\theta}_{t,i} - v)\|_{\bar{V}_{t,i}}$$
$$\leqslant \|\bar{V}_{t,i}^{1/2}\|_{\bar{V}_{t,i}^{-1}} \|\phi_i(x_{t,i}, c_t)\|_{\bar{V}_{t,i}^{-1}} \|\bar{V}_{t,i}^{-1/2}\|_{\bar{V}_{t,i}} \|(\hat{\theta}_{t,i} - v)\|_{\bar{V}_{t,i}}$$
$$= \|\phi_i(x_{t,i}, c_t)\|_{\bar{V}_{t,i}^{-1}} \|(\hat{\theta}_{t,i} - v)\|_{\bar{V}_{t,i}} \leqslant \frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}}$$

Similarly, we derive the second condition. □

In the next lemma, we show that an action chosen by the learning agent in line 14 of Algorithm 1 satisfies the baseline constraint. Let us define $\bar{x}_{t,i}^\star = \arg\max_{\bar{x}_{t,i}^\star \in \mathcal{A}} r_{x_{t,i},c}$.

Lemma 3: In the DiSC-UCB algorithm, Algorithm 1, with probability $1 - M\delta$, any action chosen by the agent from the pruned action set $\mathcal{X}_{t,i}$ satisfies the performance constraint if $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_{t,i}}})^2$.

Proof: Let $\bar{x}$ denote an arbitrary action in the pruned action set $\mathcal{X}_{t,i}$ that does not meet the performance constraint. Our goal is to show that when the agent's action is played, an action in $\mathcal{X}_{t,i}$ that satisfies the performance constraint will be selected, i.e., $\bar{x}$ is not selected. Based on the definition of $\bar{x}_{t,i}^\star$, it can be observed that $\bar{x}_{t,i}^\star$ is always contained within the pruned action set $\mathcal{X}_{t,i}$ and meets the performance constraint. To this end, if we show

$$\max_{\theta_1 \in \mathcal{B}_{t,i}} \psi_i(\bar{x}_{t,i}^\star, \mu_t)^\top \theta_1 > \max_{\theta_2 \in \mathcal{B}_{t,i}} \psi_i(\bar{x}, \mu_t)^\top \theta_2,$$

it ensures that actions within $\mathcal{X}_{t,i}$ that violate the baseline constraint are never selected. Since $\bar{x}$ does not satisfy the baseline constraint, we know $\phi_i(\bar{x}, c_t)^\top \theta^\star < (1-\alpha) r_{b_t,i}$ which leads to

$$\psi_i(\bar{x}, \mu_t)^\top \theta^\star = \mathbb{E}[\phi_i(\bar{x}, c_t)^\top \theta^\star] < (1-\alpha) r_{b_t,i}. \quad (9)$$

Moreover, we have

$$\psi_i(\bar{x}_{t,i}^\star, \mu_t)^\top \theta^\star - r_{b_t,i} = \mathbb{E}[\phi_i(\bar{x}_{t,i}^\star, c_t)^\top \theta^\star - r_{b_t,i}] \geqslant 0. \quad (10)$$

To show $\max_{\theta_1 \in \mathcal{B}_{t,i}} \psi_i(\bar{x}_{t,i}^\star, \mu_t)^\top \theta_1 > \max_{\theta_2 \in \mathcal{B}_{t,i}} \psi_i(\bar{x}, \mu_t)^\top \theta_2$, based on Lemma 1, it is sufficient to demonstrate that with probability $1 - M\delta$, $\psi_i(\bar{x}_{t,i}^\star, \mu_t)^\top \theta^\star > \max_{\theta_2 \in \mathcal{B}_{t,i}} [\psi_i(\bar{x}, \mu_t)^\top \theta^\star + \psi_i(\bar{x}, \mu_t)^\top (\theta_2 - \hat{\theta}_{t,i}) + \psi_i(\bar{x}, \mu_t)^\top (\hat{\theta}_{t,i} - \theta^\star)]$. By using (9), (10) and Lemma 2, we prove the aforementioned inequality always holds if $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_{t,i}}})^2$. Therefore, we conclude that with probability $1 - M\delta$, any action chosen by Algorithm 1 from the pruned action set $\mathcal{X}_{t,i}$ satisfies the performance constraint if $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_{t,i}}})^2$.

Lemma 4: At each round t, given the fraction α, for any $\rho \in (0, \bar{\rho}]$, where $\bar{\rho} = \frac{\alpha r_l}{1+r_h}$, the conservative feature vector $\psi_i(x_{t,i}, \mu_t) = (1-\rho)\psi_i(x_{b_t,i}, \mu_t) + \rho\zeta_{t,i}$ is safe.

Proof: To demonstrate the safety of the conservative feature vector $\psi_i(x_{t,i}, \mu_t) = (1-\rho)\psi_i(x_{b_t,i}, \mu_t) + \rho\zeta_{t,i}$, we need to show that $((1-\rho)\phi_i(x_{b_t,i}, c_t) + \rho\zeta_{t,i})^\top\theta^\star \geqslant (1-\alpha)r_{b_t,i}$ always holds. This can be shown by verifying the following condition

$$r_{b_t,i} - \rho r_{b_t,i} + \rho\zeta_{t,i}^\top\theta^\star \geqslant (1-\alpha)r_{b_t,i},$$

which is equivalent to $\rho(r_{b_t,i} - \zeta_{t,i}^\top\theta^\star) \leqslant \alpha r_{b_t,i}$ By applying Cauchy Schwarz inequality, we deduce

$$\rho \leqslant \frac{\alpha r_{b_t,i}}{1 + r_{b_t,i}}. \tag{11}$$

Consequently, by setting a lower bound for the right-hand side of (11) with the assumption $r_l \leqslant r_{b_t,i} \leqslant r_h, \rho \leqslant \frac{\alpha r_l}{1+r_h}$. Therefore, for any $\rho \leqslant \frac{\alpha r_l}{1+r_h}$, the conservative feature vector $\psi_i(x_{t,i}, \mu_t) = (1-\rho)\psi_i(x_{b_t,i}, \mu_t) + \rho\zeta_{t,i}$ is safe. □

Remark 1: Lemma 3 and Lemma 4 thus jointly prove that all the actions chosen by the proposed DiSC-UCB algorithm guarantees safety constraints.

Next, we show optimal action $x_{t,i}^\star$ always exists within the pruned action set when $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_t,i}})^2$. We extend the approach in Lemma C.1 in [7] to the unknown context case.

Lemma 5: Let $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_t,i}})^2$. Then, with probability $1 - M\delta$, the optimal action $x_{t,i}^\star$ lies in the pruned action set $\mathcal{X}_{t,i}$ for all M agent, i.e., $x_{t,i}^\star \in \mathcal{X}_{t,i}$.

Proof: To prove the optimal action $x_{t,i}^\star$ always exists in the pruned action set under the condition on the smallest eigenvalue of the Gram matrix, we need to show

$$\psi_i(x_{t,i}^\star, \mu_t)^\top\hat{\theta}_{t,i} \geqslant \frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}} + (1-\alpha)r_{b_t,i},$$

which is equivalent to demonstrating $\psi_i(x_{t,i}^\star, \mu_t)^\top(\hat{\theta}_{t,i} - \theta^\star) + \psi_i(x_{t,i}^\star, \mu_t)^\top\theta^\star \geqslant \frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}} + (1-\alpha)r_{b_t,i}$. By using Lemma 1, it can be determined that with probability $1 - M\delta$, $\theta^\star$ lies within the confidence set $\mathcal{B}_{t,i}$. Subsequently, applying Lemma 2, we recognize that with probability $1 - M\delta$, $\psi_i(x_{t,i}^\star, \mu_t)^\top(\hat{\theta}_{t,i} - \theta^\star) \geqslant -\frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}}$, and it is also known that $\psi_i(x_{t,i}^\star, \mu_t)^\top\theta^\star - r_{b_t,i} \geqslant \psi_i(\bar{x}_{t,i}^\star, \mu_t)^\top\theta^\star - r_{b_t,i} = \mathbb{E}[\phi_i(\bar{x}_{t,i}^\star, c_t)^\top\theta^\star - r_{b_t,i}] \geqslant 0$. Hence, the sufficient condition for our result is $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_t,i}})^2$. Thus, the above analysis verifies that with probability $1 - M\delta$, the optimal action $x_{t,i}^\star$ is always present in the pruned action set with probability $1 - M\delta$ under the condition $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_t,i}})^2$. □

C. Regret Analysis

In this section, we derive regret and communication bounds for DiSC-UCB. Let $|N_{t-1}|$ denote the set of rounds $j < t$ where our algorithm chooses the safe action, and similarly, $|N_{t-1}^c| = \{1, \ldots, t-1\} - |N_{t-1}|$ represents the set of rounds $j < t$ where our algorithm selects conservative actions.

In Theorem 1, we decompose our cumulative regret into three terms. The first two terms reflect the regret of choosing the agents' suggested actions. Term 2 is due to the agent's limitation of only observing the context distribution μ_t without accessing the exact context c_t. To quantify this term, we reduce it to a martingale difference sequence and then apply the Azuma-Hoeffding inequality. Term 3 results from selecting the conservative feature vector whenever the actions suggested by the agents do not meet the constraints.

Theorem 1: The regret of the DiSC-UCB algorithm, Algorithm 1, can be decomposed into three terms as follows

$$\mathcal{R}_T \leqslant \underbrace{4\beta_T\sqrt{M|N_T|d\log(M|N_T|)}(1+\log(M|N_T|))}_{Term\ 1} + \underbrace{4\sqrt{2M|N_T|\log\left(\frac{2}{\delta}\right)}}_{Term\ 2} + \underbrace{\sum_{i=1}^{M}\sum_{t\in|N_T^c|}(\kappa_h + \rho r_h + \rho)}_{Term\ 3}.$$

Proof: Let τ be the last round in which Algorithm 1 plays the agent's action, $\tau = \max\{1 \leqslant t \leqslant T \mid x_{t,i} = x'_{t,i}\}$. By the definition of cumulative regret, we have

$$\begin{aligned}
\mathcal{R}_T &= \sum_{i=1}^{M}\sum_{t=1}^{T}(\phi_i(x_{t,i}^\star, c_t)^\top\theta^\star - \phi_i(x_{t,i}, c_t)^\top\theta^\star) \\
&= \sum_{i=1}^{M}\sum_{t\in|N_T|}(\phi_i(x_{t,i}^\star, c_t)^\top\theta^\star - \phi_i(x_{t,i}, c_t)^\top\theta^\star) \\
&\quad + \sum_{i=1}^{M}\sum_{t\in|N_T^c|}(\phi_i(x_{t,i}^\star, c_t)^\top\theta^\star - \phi_i(x_{t,i}, c_t)^\top\theta^\star) \\
&= \sum_{i=1}^{M}\sum_{t\in|N_T|}(\phi_i(x_{t,i}^\star, c_t)^\top\theta^\star - \phi_i(x_{t,i}, c_t)^\top\theta^\star) \\
&\quad + \sum_{i=1}^{M}\sum_{t\in|N_T^c|}(\phi_i(x_{t,i}^\star, c_t)^\top\theta^\star - (1-\rho)\phi_i(x_{b_t,i}, c_t)^\top \\
&\quad \times \theta^\star - \rho\zeta_{t,i}^\top\theta^\star) \\
&\leqslant \underbrace{\sum_{i=1}^{M}\sum_{t\in|N_T|}(\phi_i(x_{t,i}^\star, c_t)^\top\theta^\star - \phi_i(x_{t,i}, c_t)^\top\theta^\star)}_{Agents'\ term} \\
&\quad + \sum_{i=1}^{M}\sum_{t\in|N_T^c|}(\kappa_{b_t,i} + \rho r_{b_t,i} + \rho) \qquad (12) \\
&\leqslant \underbrace{4\beta_T\sqrt{M|N_T|d\log(MT)} + 4\beta_T\sqrt{MTd\log MT}\log(MT)}_{Term\ 1} \\
&\quad + \underbrace{\sqrt{2M|N_T|\log\frac{2}{\delta}}}_{Term\ 2} + \underbrace{\sum_{i=1}^{M}\sum_{t\in|N_T^c|}(\kappa_h + \rho r_h + \rho)}_{Term\ 3}, \qquad (13)
\end{aligned}$$

where (12) follows from definition of $\kappa_{b_t,i} := \phi_i(x^\star_{t,i}, c_t)^\top\theta^\star - r_{b_t,i}$, and $-\zeta_{t,i}^\top\theta^\star \leqslant |\zeta_{t,i}^\top\theta^\star| \leqslant \|\zeta_{t,i}\|\|\theta^\star\| \leqslant 1$. Equation (13) is derived through the analysis of the *Agents' term* 'in (12) and $\kappa_{b_t,i} \leqslant \kappa_h$, $r_{b_t,i} \leqslant r_h$. The Agents' term captures the unconstrained case studied in our earlier work [10] (see Theorem 4.1, [10], presented in Proposition 3 in Appendix˜B in arXiv version.). Given that $|N_T| = T - |N_T^c|$, the remaining section of our proof focuses on determining the upper bound and lower bound for $|N_T^c|$. □

Consider any round t during which the agent plays the agent's action, i.e., at round t, both condition $F = 1$ is met and $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_t,i}})^2$ is satisfied. By Lemma 5, if $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_t,i}})^2$, it is guaranteed that $x^\star_{t,i} \in \mathcal{X}_{t,i}$. Consequently, $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_t,i}})^2$ is sufficient to guarantee that $\mathcal{X}_{t,i}$ is non-empty. To this end, our analysis of Algorithm 1 will henceforth only focuses on the condition $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_t,i}})^2$.

Lemma 6: The smallest eigenvalue of the Gram matrix $\lambda_{\min}(\bar{V}_{t,i})$ satisfies $\lambda_{\min}(\bar{V}_{t,i})$ is upper bounded by

$$(\lambda + MT) + 2M\rho(\rho-1)|N_T^c| + \sqrt{32M\rho^2(1-\rho)^2|N_T^c|\log\left(\frac{d}{\delta}\right)}.$$

Proof: We start with the definitions of $\bar{V}_{t,i}, W_{\text{syn}}, W_{t,i}$

$$\begin{aligned}\bar{V}_{t,i} &= \lambda I + W_{\text{syn}} + W_{t,i}\\ &= \lambda I + \sum_{i=1}^{M}\sum_{t=1}^{t_{last}} \psi_i(x_{t,i},\mu_t)\psi_i(x_{t,i},\mu_t)^\top\\ &\quad + \sum_{t=t_{last}}^{T} \psi_i(x_{t,i},\mu_t)\psi_i(x_{t,i},\mu_t)^\top\\ &\preceq \lambda I + \sum_{i=1}^{M}\sum_{t=1}^{T} \psi_i(x_{t,i},\mu_t)\psi_i(x_{t,i},\mu_t)^\top\\ &\preceq \lambda I + \sum_{i=1}^{M}\sum_{t\in|N_T|} \psi_i(x_{t,i},\mu_t)\psi_i(x_{t,i},\mu_t)^\top\\ &\quad + \sum_{i=1}^{M}\sum_{t\in|N_T^c|} ((1-\rho)\psi_i(x_{b_t,i},\mu_t) + \rho\zeta_{t,i})\\ &\quad \times ((1-\rho)\psi_i(x_{b_t,i},\mu_t) + \rho\zeta_{t,i})^\top\\ &\preceq \lambda I + \sum_{i=1}^{M}\sum_{t\in|N_T|} I + \sum_{i=1}^{M}\sum_{t\in|N_T^c|} ((1-\rho)^2 I\\ &\quad + \rho(1-\rho)\psi_i(x_{b_t,i},\mu_t)\zeta_{t,i}^\top\\ &\quad + \rho(1-\rho)\zeta_{t,i}\psi_i(x_{b_t,i},\mu_t)^\top + \rho^2 I), \qquad (14)\end{aligned}$$

where (14) follows by substituting the feature vectors corresponding to the rounds in which the learner's chosen action is played and those in which the conservative action is played. The last step follows since $\psi_i(x_{t,i},\mu_t)\psi_i(x_{t,i},\mu_t)^\top \preceq I$ are rank-1 matrices with $\|\psi_i(x_{t,i},\mu_t)\|_2 \leqslant 1$, and thus their eigenvalues are either 1 or 0. By considering the above relationships and using $|N_T| = T - |N_T^c|$, we obtain

$$\bar{V}_{t,i} \preceq \lambda I + M(T - |N_T^c|)I + M|N_T^c|(2\rho^2 - 2\rho + 1)I + \sum_{i=1}^{M}\sum_{t\in|N_T^c|} U_{t,i},$$

where $U_{t,i} = \rho(1-\rho)\psi_i(x_{b_t,i},\mu_t)\zeta_{t,i}^\top + \rho(1-\rho)\zeta_{t,i}\psi_i(x_{b_t,i},\mu_t)^\top$. By applying Weyl's inequality,

$$\lambda_{\min}(\bar{V}_{t,i}) \leqslant (\lambda + MT) + 2M\rho(\rho-1)|N_T^c| + \lambda_{\max}\left(\sum_{i=1}^{M}\sum_{t\in|N_T^c|} U_{t,i}\right). \qquad (15)$$

Next, we use the matrix Azuma inequality to determine the upper bound of $\lambda_{\max}(\sum_{i=1}^{M}\sum_{t\in|N_T^c|} U_{t,i})$. From the definition of $U_{t,i}$, it follows that $\mathbb{E}[U_{t,i}|F_{s-1}] = 0$ and

$$\begin{aligned}\max_{\|u\|_2=\|v\|_2=1} u^\top U_{t,i} v &= \rho(1-\rho)(u^\top\psi_i(x_{b_t,i},\mu_t))(v^\top\zeta_{t,i})^\top\\ &\quad + \rho(1-\rho)(u^\top\zeta_{t,i})(v^\top\psi_i(x_{b_t,i},\mu_t))^\top \qquad (16)\\ &\leqslant \rho(1-\rho)\|\psi_i(x_{b_t,i},\mu_t)\|\|\zeta_{t,i}\|\\ &\quad + \rho(1-\rho)\|\zeta_{t,i}\|\|\psi_i(x_{b_t,i},\mu_t)\|\\ &\leqslant 2\rho(1-\rho), \qquad (17)\end{aligned}$$

where (16) follows from Cauchy-Schwarz inequality, (17) follows from $\|\psi_i(x_{b_t,i},\mu_t)\| \leqslant 1$ and $\|\zeta_{t,i}\| = 1$. Further, we utilize the property that for any matrix A, we have $A^2 \preceq \|A\|_2^2 I$, where $\|A\|_2$ is the maximum singular value of A given by $\sigma_{\max}(A) = \max_{\|u\|_2=\|v\|_2=1} u^\top A v$. Thus

$$U_{t,i}^2 \preceq \sigma_{\max}(U_{t,i})^2 I \preceq 4\rho^2(1-\rho)^2 I.$$

Moreover, by using the triangular inequality, we can express

$$\|\sum_{i=1}^{M}\sum_{t\in|N_T^c|} U_{t,i}^2\| \leqslant \sum_{i=1}^{M}\sum_{t\in|N_T^c|} \|U_{t,i}^2\| \leqslant 4M\rho^2(1-\rho)^2|N_T^c|.$$

Now, by applying the matrix Azuma inequality, for any $c \geqslant 0$,

$$\mathbb{P}\left(\lambda_{\max}(\sum_{i=1}^{M}\sum_{t\in|N_T^c|} U_{t,i}) \geqslant c\right) \leqslant d\exp\left(-\frac{c^2}{32M\rho^2(1-\rho)^2|N_T^c|}\right).$$

Thus we have, with probability $1-\delta$,

$$\lambda_{\max}\left(\sum_{i=1}^{M}\sum_{t\in|N_T^c|} U_{t,i}\right) \leqslant \sqrt{32M\rho^2(1-\rho)^2|N_T^c|\log\left(\frac{d}{\delta}\right)}.$$

Combining the above result with (15), we obtain

$$\lambda_{\min}(\bar{V}_{t,i}) \leqslant (\lambda + MT) + 2M\rho(\rho-1)|N_T^c|$$

$$+\sqrt{32M\rho^2(1-\rho)^2|N_T^c|\log\left(\frac{d}{\delta}\right)}. \quad (18)$$

This concludes the proof of Lemma 6, which demonstrates the upper bound for $\lambda_{\min}(\bar{V}_{t,i})$ based on the stated inequalities. □

Lemma 7: For any, $a, b, c > 0$, if $ax + c \leqslant \sqrt{bx}$, then the following holds for $x \geqslant 0$

$$\frac{b-2ac-\sqrt{b^2-4abc}}{2a^2} \leqslant x \leqslant \frac{b-2ac+\sqrt{b^2-4abc}}{2a^2}.$$

Proof: Let $a, b, c > 0$, and consider the inequality $ax + c \leqslant \sqrt{bx}$. We can square both sides and rearrange them to obtain the quadratic inequality $a^2x^2 - (b-2ac)x + c^2 \leqslant 0$. Since $a^2 > 0$, we can apply the solution formula for quadratic inequalities to express the solution for $x \frac{b-2ac-\sqrt{b^2-4abc}}{2a^2} \leqslant x \leqslant \frac{b-2ac+\sqrt{b^2-4abc}}{2a^2}$. □

To determine the upper bound of the cumulative regret in Theorem 1, we determine the upper bounds of $|N_T|$ and $|N_T^c|$. Given that $|N_T^c| = T - |N_T|$, we compute the upper and lower bounds for $|N_T^c|$. This is given in Theorem 2.

Theorem 2: In the DiSC-UCB algorithm, Algorithm 1, the upper bound and lower bound of $|N_T^c|$ are given by

$$|N_T^c| \geqslant \frac{4}{M}\log\left(\frac{d}{\delta}\right) - \frac{\left(\frac{2\beta_{0,i}}{\alpha r_l}\right)^2 - (\lambda + MT)}{2M\rho(1-\rho)}$$
$$- \sqrt{\frac{16}{M^2}\log^2\left(\frac{d}{\delta}\right) - \frac{4\left(\left(\frac{2\beta_{0,i}}{\alpha r_l}\right)^2 - (\lambda+MT)\right)\log\left(\frac{d}{\delta}\right)}{M^2\rho(1-\rho)}}$$
$$|N_T^c| \leqslant$$
$$\sqrt{\left(\frac{4\sqrt{6}}{M}\log\left(\frac{d}{\delta}\right)\right)^2 - \frac{16\left(\left(\frac{2\beta_{0,i}}{\alpha r_l}\right)^2 - (\lambda+MT)\right)\log\left(\frac{d}{\delta}\right)}{M^2\rho(1-\rho)}},$$

where $\beta_{0,i} = \sigma\sqrt{2\log\frac{1}{\delta}} + \lambda^{\frac{1}{2}}$ and $T \leqslant \frac{1}{M}\left[\left(\frac{2\beta_{\tau,i}}{\alpha r_l}\right)^2 - \lambda\right]$.

Proof: Recall that τ is the last round in which Algorithm 1 plays the agent's action. Given any round t that the agent's action is played, we have $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_{b_t,i}})^2$. By using $r_{b_t,i} \geqslant r_l$, it follows that $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2\beta_{t,i}}{\alpha r_l})^2$. From (18), we derive

$$\left(\frac{2\beta_{\tau,i}}{\alpha r_l}\right)^2 \leqslant (\lambda + M\tau) + 2M\rho(\rho-1)|N_\tau^c|$$
$$+\sqrt{32M\rho^2(1-\rho)^2|N_\tau^c|\log\left(\frac{d}{\delta}\right)}$$

By setting the gradient of $2\rho(\rho-1)$ to zero, $2\rho(\rho-1) \in [0, \frac{1}{2}]$. Moreover, it is obvious that $|N_T^c| - |N_\tau^c| = T - \tau$. Consequently, we can obtain the inequalities $(\frac{2\beta_{\tau,i}}{\alpha r_l})^2 \leqslant (\lambda + MT) + 2M\rho(\rho-1)|N_T^c| + \sqrt{32M\rho^2(1-\rho)^2|N_T^c|\log(\frac{d}{\delta})}2M\rho(1-\rho)|N_T^c| + (\frac{2\beta_{\tau,i}}{\alpha r_l})^2 - (\lambda + MT)$ which is

$$\leqslant \sqrt{32M\rho^2(1-\rho)^2|N_T^c|\log\left(\frac{d}{\delta}\right)}. \quad (19)$$

By using the substitution $|N_T^c| = T - |N_T|$, we have the following equivalent inequality

$$(\frac{2\beta_{\tau,i}}{\alpha r_l})^2 \leqslant \lambda + MT(2\rho^2 - 2\rho + 1) + 2M\rho(1-\rho)|N_T|$$
$$+\sqrt{32M\rho^2(1-\rho)^2(T-|N_T|)\log\left(\frac{d}{\delta}\right)}.$$

By rearranging the terms of these inequalities we get,

$$\sqrt{32M\rho^2(1-\rho)^2(T-|N_T|)\log\left(\frac{d}{\delta}\right)}$$
$$\geqslant -2M\rho(1-\rho)|N_T| + \left(\frac{2\beta_{\tau,i}}{\alpha r_l}\right)^2$$
$$-(\lambda + MT(2\rho^2 - 2\rho + 1)). \quad (20)$$

If $|N_T^c| = 0$, it implies that the agent's action will be played in every iteration and the baseline constraint is not active. In this case, cumulative regret follows with the findings in Proposition 3 in Appendix˜B in arXiv version (Theorem 4.1, [10]), making this a trivial case. Hence, the focus of this paper is when $|N_T^c| \neq 0$. Recognizing that (20) is not always true for every round, we must have $(\frac{2\beta_{\tau,i}}{\alpha r_l})^2 - (\lambda + MT(2\rho^2 - 2\rho + 1)) \geqslant 0$. Given $\frac{1}{2} \leqslant 2\rho^2 - 2\rho + 1 \leqslant 1$, it follows that $(\frac{2\beta_{\tau,i}}{\alpha r_l})^2 - (\lambda + MT) \geqslant 0$. Recall (19). We know

$$a := 2M\rho(1-\rho) \geqslant 0 \text{ and } b := 32M\rho^2(1-\rho)^2\log(\frac{d}{\delta}) \geqslant 0.$$

From Theorem 2, we know $c := ((\frac{2\beta_{\tau,i}}{\alpha r_l})^2 - (\lambda + MT)) \geqslant 0$. Using this (19) can be rewritten as $a|N_T^c| + c \leqslant \sqrt{b|N_T^c|}$, which gives $a^2|N_T^c|^2 - (b-2ac)|N_T^c| + c^2 \leqslant 0$. Recognizing the existence of a solution for this inequality and since $|N_T^c| \geqslant 0$, it implies that $b - 2ac \geqslant 0$. By using Lemma 7, we get $\frac{b-2ac-\sqrt{b^2-4abc}}{2a^2} \leqslant |N_T^c| \leqslant \frac{b-2ac+\sqrt{b^2-4abc}}{2a^2}$.

First, we analyze the upper bound. Given the fact that $b - 2ac \geqslant 0$ and using the identity $\sqrt{x} + \sqrt{y} \leqslant \sqrt{2x+2y}$ for any $x, y \geqslant 0$, it follows that $b - 2ac + \sqrt{b^2-4abc} = \sqrt{(b-2ac)^2} + \sqrt{b^2-4abc} \leqslant \sqrt{4b^2 - 16abc + 8a^2c^2} \leqslant \sqrt{6b^2 - 16abc}$. Thus the upper bound can be simplified as

$$\frac{b-2ac+\sqrt{b^2-4abc}}{2a^2} \leqslant \sqrt{\frac{3}{2}\left(\frac{b}{a^2}\right)^2 - \frac{4bc}{a^3}}.$$

Subsequently, when considering the partial derivative with regard to c for $\frac{b-2ac-\sqrt{b^2-4abc}}{2a^2}$ and $\sqrt{\frac{3}{2}(\frac{b}{a^2})^2 - \frac{4bc}{a^3}}$, the gradient of the former is non-negative, while the gradient of the latter is non-positive. Given the fact that $\beta_{t,i}$ is an increasing sequence, we choose to replace $\beta_{\tau,i}$ with $\beta_{0,i}$ in both the upper and lower bounds. Based on the above analysis and the given definition of

$\beta_{0,i} = \sigma\sqrt{2\log\frac{1}{\delta}} + \lambda^{\frac{1}{2}}$, we obtain the lower and upper bounds for $|N_T^c|$ as given in Theorem 2.

Remark 2: We note that in the proof of Theorem 2 we get $T \leqslant \frac{1}{M}[(\frac{2\beta_{\tau,i}}{\alpha r_l})^2 - \lambda]$. This ensures that the lower bound is always smaller than the upper bound in Theorem 2.

Combining Theorem 1 and Theorem 2, we get the following bound for the cumulative regret for Alg. 1 (DiSC-UCB).

Theorem 3: The cumulative regret of DiSC-UCB algorithm, Algorithm 1, with $\beta_{t,i} = \beta_{t,i}(\sqrt{1+\sigma^2}, \delta/2)$ is bounded at round T with probability at least $1 - M\delta$ by

$$\mathcal{R}_T \leqslant 4\beta_T\sqrt{dc'\log(MT)} + 4\beta_T\sqrt{MTd\log MT}\log(MT) + \sqrt{2c'\log\left(\frac{2}{\delta}\right)} + c''(\kappa_h + \rho r_h + \rho),$$

where $c', c'' > 0$ are given by $c' = O(MT)$ and $c'' = O(\sqrt{MT})$. Further, for $\delta = \frac{1}{M^2 T}$, Algorithm 1 achieves a regret of $O(d\sqrt{MT}\log^2 T)$ with $O(M^{1.5}d^3)$ communication cost.

Proof: From 13, we know cumulative regret is bounded

$$\begin{aligned}\mathcal{R}_T &\leqslant 4\beta_T\sqrt{M|N_T|d\log(MT)} \\ &\quad + 4\beta_T\sqrt{MTd\log MT}\log(MT) \\ &\quad + \sqrt{2M|N_T|\log\frac{2}{\delta}} + M|N_T^c|(\kappa_h + \rho r_{b_t,i} + \rho).\end{aligned}$$

From Theorem 2, we get the upper and lower bound for $|N_T^c|$. Therefore, the cumulative regret is bound by

$$\mathcal{R}_T \leqslant 4\beta_T\sqrt{dc'\log(MT)} + 4\beta_T\sqrt{MTd\log MT}\log(MT) + \sqrt{2c'\log\left(\frac{2}{\delta}\right)} + c''(\kappa_h + \rho r_h + \rho),$$

where c' and c'' are as given in [1]. Furthermore, given that $c' = O(MT)$, $c'' = O(\sqrt{MT})$, and $\beta_T = O(\sqrt{d\log\frac{T}{\delta}}) = O(\sqrt{d\log MT})$, for $\delta = \frac{1}{M^2 T}$, the cumulative regret $\mathcal{R}_T$ can be bounded as

$$\begin{aligned}\mathcal{R}_T &\leqslant 4\beta_T\sqrt{dc'\log(MT)} + 4\beta_T\sqrt{MTd\log MT}\log(MT) \\ &\quad + \sqrt{2c'\log\left(\frac{2}{\delta}\right)} + c''(\kappa_h + \rho r_h + \rho) \\ &= O(d\sqrt{MT}\log MT) + O(d\sqrt{MT}\log^2 MT) \\ &\quad + O(\sqrt{MT\log M^2T}) + O(\sqrt{MT}) \\ &= O(d\sqrt{MT}\log^2 MT) = \widetilde{O}(d\sqrt{MT}),\end{aligned}$$

where the last step is followed by the condition $T > M$.

Communication: The communication cost for our algorithm follows the approach in [23]. □

Remark 3: A naive adaptation of solving the M tasks separately would result in $\widetilde{O}(dM\sqrt{T})$ regret. In contrast, our proposed method achieves a sublinear regret of $\widetilde{O}(d\sqrt{MT})$ which validates the effectiveness of the proposed approach.

V. Unknown Baseline Reward

This section considers the unknown baseline reward setting. Such a setting was first studied in [3] for conservative CB with cumulative performance constraint and in [7] for CB with stage-wise constraints. We extend the model in [7] to the distributed setting with unknown contexts. We assume the agents know the lower bound r_ℓ in Assumption 3. We describe the modifications to the DiSC-UCB algorithm to handle the unknown baseline (DiSC-UCB-UB). Then, we prove that the regret and communication bounds for DiSC-UCB-UB are in the same order as DiSC-UCB. The approach is primarily based on the observation that $\theta^\star$ lies in $\mathcal{B}_{t,i}$ with high probability. Hence we use the upper bound to replace the value of $r_{b_t,i}$

$$\begin{aligned}\max_{v\in\mathcal{B}_{t,i}} \psi_i(x_{b_t,i}, \mu_t)^\top v &\geqslant \psi_i(x_{b_t,i}, \mu_t)^\top \theta^\star \\ &= \mathbb{E}[\phi_i(x_{b_t,i}, c_t)^\top \theta^\star] = r_{b_t,i}.\end{aligned}$$

Hence, the safety constraint can be formulated as

$$\min_{v\in\mathcal{B}_{t,i}} \phi_i(x_{t,i}, c_t)^\top v \geqslant (1-\alpha)\max_{v\in\mathcal{B}_{t,i}} \psi_i(x_{b_t,i}, \mu_t)^\top v.$$

Given the agent's lack of knowledge regarding the reward provided by the baseline policy and its only knowledge of the lower bound r_l we can construct the pruned action set as

$$\mathcal{Z}_{t,i} = \left\{ x_{t,i} \in \mathcal{A} : \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i} \geqslant \frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}} + (1-\alpha)\max_{v\in\mathcal{B}_{t,i}} \psi_i(x_{b_t,i}, \mu_t)^\top v \right\}. \tag{21}$$

It can be observed that when the condition

$$\lambda_{\min}(\bar{V}_{t,i}) \geqslant \left(\frac{2(2-\alpha)\beta_{t,i}}{\alpha r_l}\right)^2 \tag{22}$$

is satisfied, the optimal action $x_{t,i}^\star$ is contained within the pruned action set $\mathcal{Z}_{t,i}$ with high probability. The details of these two derivations are presented below. The necessary modifications to the DiSC-UCB algorithm can be made by updating lines 8 and 13 with Eqs. (21) and (22), respectively.

Construction of the Pruned Action Set $\mathcal{Z}_{t,i}$: We analyze two cases. 1) $\phi_i(x_{t,i}, c_t)^\top \hat{\theta}_{t,i} \geqslant \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i}$ and 2) $\phi_i(x_{t,i}, c_t)^\top \hat{\theta}_{t,i} \leqslant \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i}$. Let us first address case 1), and explain the process of constructing a subset of actions

[1] The constants $c', c'' > 0$ is given by

$$c' = MT - 4\log(\frac{d}{\delta}) + \frac{(\frac{2}{\alpha r_l}(\sigma\sqrt{2\log\frac{1}{\delta}} + \lambda^{\frac{1}{2}}))^2 - (\lambda + MT)}{2\rho(1-\rho)} + \sqrt{16\log^2(\frac{d}{\delta}) - \frac{4((\frac{2}{\alpha r_l}(\sigma\sqrt{2\log\frac{1}{\delta}}+\lambda^{\frac{1}{2}}))^2 - (\lambda+MT))\log(\frac{d}{\delta})}{\rho(1-\rho)}}$$

$$c'' = \sqrt{(4\sqrt{6}\log(\frac{d}{\delta}))^2 - \frac{16((\frac{2}{\alpha r_l}(\sigma\sqrt{2\log\frac{1}{\delta}} + \lambda^{\frac{1}{2}}))^2 - (\lambda + MT))\log(\frac{d}{\delta})}{\rho(1-\rho)}}.$$

that satisfy the constraint for all $v \in \mathcal{B}_{t,i}$. Define $\mathcal{Z}^1_{t,i}$

$$:= \{x_{t,i} \in \mathcal{A} : \min_{v\in\mathcal{B}_{t,i}} \phi_i(x_{t,i}, c_t)^\top v \geqslant (1-\alpha) \max_{v\in\mathcal{B}_{t,i}} \psi_i(x_{b_t,i}, \mu_t)^\top v\} \quad (23)$$

$$\Leftarrow \Big\{x_{t,i} \in \mathcal{A} : \min_{v\in\mathcal{B}_{t,i}} \phi_i(x_{t,i}, c_t)^\top (v - \hat{\theta}_{t,i}) + \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i} + (\phi_i(x_{t,i}, c_t) - \psi_i(x_{t,i}, \mu_t))^\top \hat{\theta}_{t,i} \geqslant (1-\alpha) \max_{v\in\mathcal{B}_{t,i}} \psi_i(x_{b_t,i}, \mu_t)^\top v\Big\}$$

$$\Leftarrow \Big\{x_{t,i} \in \mathcal{A} : \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i} \geqslant \frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}} + (1-\alpha) \max_{v\in\mathcal{B}_{t,i}} \psi_i(x_{b_t,i}, \mu_t)^\top v\Big\} \quad (24)$$

where the last step follows from $\phi_i(x_{t,i}, c_t)^\top \hat{\theta}_{t,i} \geqslant \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i}$ and $\phi_i(x_{t,i}, c_t)^\top (v - \hat{\theta}_{t,i}) \geqslant -\frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}}$ from Lemma 2. All actions that meet the conditions in (24) also fulfill the requirements of (23), thus ensuring safety. Now we consider case 2), where $\phi_i(x_{t,i}, c_t)^\top \hat{\theta}_{t,i} \leqslant \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i}$. In this case, our approach is to first identify actions that violate the baseline constraint, $\bar{\mathcal{Z}}^2_{t,i}$, and then eliminate those actions from the action set $\mathcal{A}$.

$$\bar{\mathcal{Z}}^2_{t,i} := \Big\{x_{t,i} \in \mathcal{A} : \min_{v\in\mathcal{B}_{t,i}} \phi_i(x_{t,i}, c_t)^\top v \leqslant (1-\alpha) \max_{v\in\mathcal{B}_{t,i}} \psi_i(x_{b_t,i}, \mu_t)^\top v\Big\} \quad (25)$$

$$\Leftarrow \Big\{x_{t,i} \in \mathcal{A} : \min_{v\in\mathcal{B}_{t,i}} \phi_i(x_{t,i}, c_t)^\top (v - \hat{\theta}_{t,i}) + \phi_i(x_{t,i}, c_t)^\top \hat{\theta}_{t,i} \leqslant (1-\alpha) \max_{v\in\mathcal{B}_{t,i}} \psi_i(x_{b_t,i}, \mu_t)^\top v\Big\}$$

$$\Leftarrow \Big\{x_{t,i} \in \mathcal{A} : \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i} \leqslant (1-\alpha) \max_{v\in\mathcal{B}_{t,i}} \psi_i(x_{b_t,i}, \mu_t)^\top v\Big\} \quad (26)$$

where the last step follows from $\phi_i(x_{t,i}, c_t)^\top \hat{\theta}_{t,i} \leqslant \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i}$. Note that all actions that meet the conditions in (26) also fulfill the requirements of (25), consequently rendering them unsafe. By taking the difference between $\mathcal{A}$ and $\bar{\mathcal{Z}}^2_{t,i}$, we determine $\mathcal{Z}^2_{t,i} = \mathcal{A} \setminus \bar{\mathcal{Z}}^2_{t,i}$

$$= \Big\{x_{t,i} \in \mathcal{A} : \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i} \geqslant (1-\alpha) \max_{v\in\mathcal{B}_{t,i}} \psi_i(x_{b_t,i}, \mu_t)^\top v\Big\}.$$

Given $\mathcal{Z}^1_{t,i}$ and $\mathcal{Z}^2_{t,i}$, we obtain the pruned action set by taking the intersection between $\mathcal{Z}^1_{t,i}$ and $\mathcal{Z}^2_{t,i}$, given by

$$\mathcal{Z}_{t,i} = \Big\{x_{t,i} \in \mathcal{A} : \psi_i(x_{t,i}, \mu_t)^\top \hat{\theta}_{t,i} \geqslant \frac{\beta_{t,i}}{\sqrt{\lambda_{\min}(\bar{V}_{t,i})}} + (1-\alpha) \max_{v\in\mathcal{B}_{t,i}} \psi_i(x_{b_t,i}, \mu_t)^\top v\Big\}.$$

We describe the modifications to the DiSC-UCB algorithm to handle the unknown baseline case. We refer to the modified algorithm for the unknown case as DiSC-UCB-UB. DiSC-UCB-UB differs from DiSC-UCB only in two lines, mainly in the pruned action set construction. In the modified algorithm, DiSC-UCB-UB, we replace line 8 of Algorithm 1 with the new pruned action set $\mathcal{Z}_{t,i}$ and line 13 as **If** $F = 1$ and $\lambda_{\min}(\bar{V}_{t,i}) \geqslant (\frac{2(2-\alpha)\beta_{t,i}}{\alpha r_l})^2$. Then, we prove that the regret and the communication bounds for DiSC-UCB-UB are in the same order as those of DiSC-UCB.

Next, we present the key result that bounds the cumulative regret and communication cost for DiSC-UCB-UB.

Theorem 4: The cumulative regret of DiSC-UCB-UB (Unknown baseline setting) with $\beta_{t,i} = \beta_{t,i}(\sqrt{1+\sigma^2}, \delta/2)$ is bounded at round T with probability at least $1 - M\delta$ by

$$\mathcal{R}_T \leqslant 4\beta_T \sqrt{d\bar{c}' \log(MT)} + 4\beta_T \sqrt{MTd \log MT} \log(MT) + \sqrt{2\bar{c}' \log\left(\frac{2}{\delta}\right)} + \bar{c}''(2\rho + 1 - r_l),$$

where $\bar{c}', \bar{c}'' > 0$ are given by $\bar{c}' = O(MT)$ and $\bar{c}'' = O(\sqrt{MT})$. Further, for $\delta = \frac{1}{M^2 T}$, DiSC-UCB-UB achieves a regret of $O(d\sqrt{MT} \log^2 T)$ with $O(M^{1.5} d^3)$ communication cost.

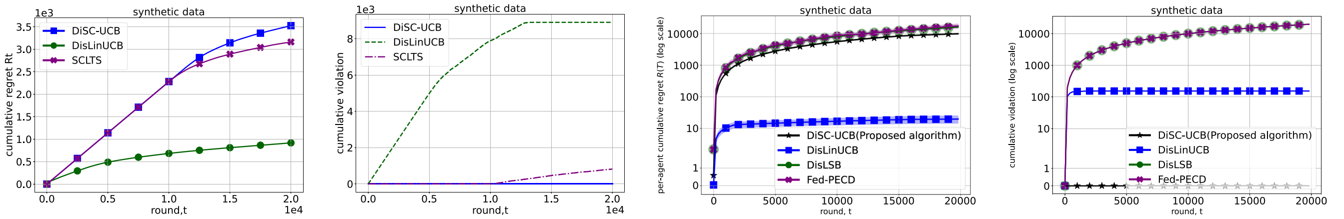
The proofs of DiSC-UCB-UB results follow a similar approach to DiSC-UCB. We present all the proofs for DiSC-UCB-UB in Appendix D in arXiv version (supplementary material).

VI. Numerical Experiments

In this section, we validate the performance of our DiSC-UCB algorithm on synthetic and real-world movielens and LastFM datasets and compare it with the SCLTS algorithm proposed in [7] and with the unconstrained distributed algorithm DisLinUCB in [23], DisLSB [22], [23], and Fed-PE [12], [24]. While SCLTS implements the Thompson algorithm for stage-wise conservative linear CB, DisLinUCB, DisLSB, and Fed-PE considers unconstrained distributed linear CB problem. We note that all the baselines assume contexts are known and $M = 1$, single agent.

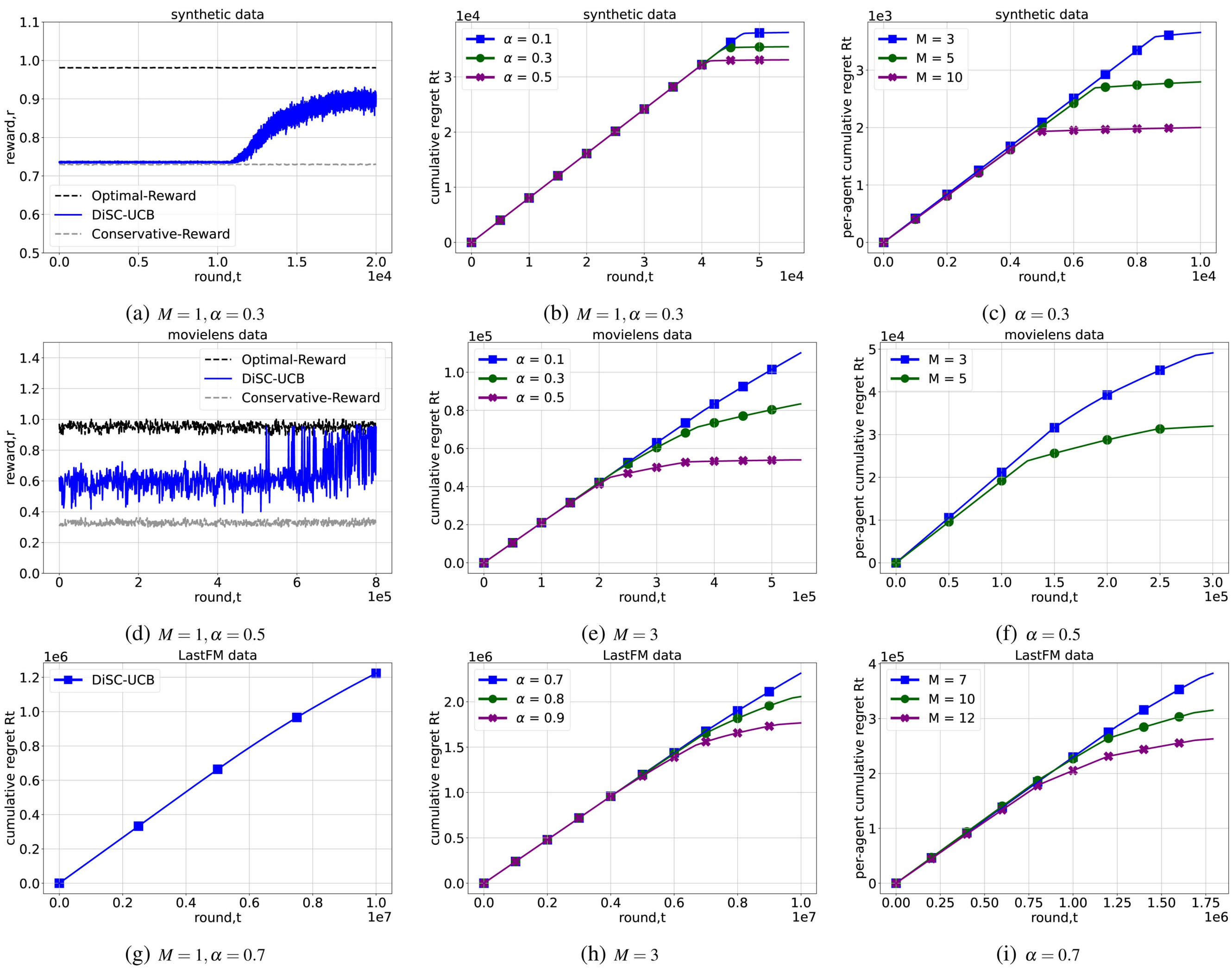
A. Datasets

Movielens data: We used movielens-100 K data [42] to evaluate the performance of our algorithm. We first get the rating matrix $r_{x,c} \in \mathbb{R}^{943\times1682}$. The entries of the rating are between 0 and 5, which we normalized to be in [0,1]. We randomly choose a set of 50 movies and a set of 100 users for our analysis, i.e., $|\mathcal{C}| = 100$ and $K = 50$. We then performed a non-negative matrix factorization of $r_{x,c} = WH$, where $W \in \mathbb{R}^{100\times3}$, $H \in \mathbb{R}^{3\times50}$. To construct the feature vector ϕ for the g^{th} user and j^{th} movie, we choose the g^{th} column of matrix



(a) $M = 1, \alpha = 0.3$ (b) $M = 1, \alpha = 0.3$ (c) $M = 3, \alpha = 0.25$ (d) $M = 3, \alpha = 0.25$

Fig. 1. Comparison of cumulative regret and cumulative violation of DiSC-UCB with SCLTS [7], DisLinUCB [23], DisLSB [22], and Fed-PE [24] modified for unknown context Fed-PECD using [12]. Synthetic data: In (a), (b) we set the parameters as $\lambda = 1$, $d = 2$, $R = 1$, $K = 40$, $M = 1$, $\theta^\star = [0.9, 0.4]$, and noise variance $= 2.5 \times 10^{-3}$, and the baseline action is set by the 10^{th} best action. In (c), (d), we set the parameters as: $\lambda = 0.1$, $R = 0.1$, $d = 2$, number of contexts $|\mathcal{C}| = 100$, number of actions $K = 10$, and number of agents $M = 3$. We considered a noise with a mean of 0 and a variance of 0.01 to obtain ψ from ϕ. The true parameters are $\theta_1^\star = [0.9, 0.4]$, $\theta_2^\star = [0.9, 0]$, and $\theta_3^\star = [0, 0.4]$. The baseline is the 2nd best action, and $\alpha = 0.25$. All plots were averaged over 100 independent trials.



(a) $M = 1, \alpha = 0.3$ (b) $M = 1, \alpha = 0.3$ (c) $\alpha = 0.3$

(d) $M = 1, \alpha = 0.5$ (e) $M = 3$ (f) $\alpha = 0.5$

(g) $M = 1, \alpha = 0.7$ (h) $M = 3$ (i) $\alpha = 0.7$

Fig. 2. Synthetic data: In (a), we set the parameters as $\lambda = 1$, $d = 2$, $R = 1$, $K = 40$, $M = 1$, $\theta^\star = [0.9, 0.4]$, and noise variance $= 2.5 \times 10^{-3}$, and the baseline action is set by the 10^{th} best action. In (b), the parameters are set as $R = 0.1$, $K = 90$, $M = 3$, $\theta^\star = [1, 1]$, noise variance $= 10^{-4}$, and the baseline action of a particular round is set as the 80^{th} best action of that round. The α values are varied as $\alpha = \{0.1, 0.3, 0.5\}$. In (c), $R = 1$, $K = 90$, $M = \{3, 5, 10\}$, $\theta^\star = [1, 1]$, the reward parameters for the different tasks $\theta_i^\star \in \Theta = \{[1, 1], [1, 0], [0, 1]\}$, noise variance $= 10^{-2}$, and baseline is the 30^{th} best action. Movielens data: In (d), (e), and (f), $R = 0.1$, $K = 50$, noise variance$= 10^{-2}$, $\theta^\star = \frac{1}{\sqrt{3}}[1, 0, 0, 0, 1, 0, 0, 0, 1]$, and the baseline action is set as the 40^{th} best action. LastFM data: In (g), (h), (i), $R = \lambda = 0.05$, $K = 50$, noise variance$= 10^{-3}$, $\theta^\star = \frac{1}{\sqrt{3}}[1, 0, 0, 0, 1, 0, 0, 0, 1]$, and the baseline action is set as the 5^{th} best action.

$W, W_g \in \mathbb{R}^3$ and the j^{th} row of H, $H_j^\top \in \mathbb{R}^{1\times 3}$. We perform the outer product $W_g H_j^\top$ to obtain a 3×3 matrix. We vectorized this matrix to obtain $\phi \in \mathbb{R}^9$. We considered a noise with a mean of 0 and a variance of 0.01 to obtain ψ from ϕ. We set the number of agents as $M = 3$, and the $\theta_1^\star = (1/\sqrt{3})[1, 0, 0, 0, 1, 0, 0, 0, 1]$, $\theta_3^\star = (1/\sqrt{3})[1, 0, 0, 0, 0, 0, 0, 0, 1]$, and $\theta_3^\star = (1/\sqrt{3})$ $[1, 0, 0, 0, 1, 0, 0, 0, 0]$. We transformed the multi-task problem with heterogeneous reward parameter $\Theta = \{\theta_1^\star, \theta_2^\star, \theta_3^\star\}$ to a distributed CB problem with common reward parameter $\theta^\star = \theta_1^\star$ and heterogeneous feature vectors ϕ_i for agent i by setting the respective feature in ϕ to zero.

LastFM data: The LastFM dataset is derived from the online music streaming service Last.fm [43] and includes data from 1892 users and 17632 artists. We consider each artist as an individual action. The reward is 1 if the user has listened to an artist at least once, and 0 otherwise. We keep only those users who have received at least 30 interaction rewards. Consequently, we get a rating matrix $r_{x,c} \in \mathbb{R}^{741\times 538}$. Following that, we performed non-negative matrix factorization on the rating matrix, denoted as $r_{x,c} = WH$, where $W \in \mathbb{R}^{741\times 3}$ and $H \in \mathbb{R}^{3\times 538}$. To construct the feature vector ϕ for the g^{th} user and j^{th} artist, we choose the g^{th} column of matrix W, $W_g \in \mathbb{R}^3$ and the j^{th} row of H, $H_j^\top \in \mathbb{R}^{1\times 3}$. We perform the outer product $W_g H_j^\top$ to obtain a 3×3 matrix. We vectorized this matrix to obtain $\phi \in \mathbb{R}^9$. We considered a noise with a mean of 0 and a variance of 10^{-3} to obtain ψ from ϕ. The baseline action is set as the 5^{th} best action and $\lambda = R = 0.05$.

B. Comparison of DiSC-UCB With Existing Constrained and Distributed Approaches

In Figs. 1(a), 1(b), we set $M = 1$ and compare the performance of DiSC-UCB with SCLTS and DisLinUCB. SCLTS studied the constrained bandit problem with a single-agent and known contexts. To this end, we set $M = 1$ in this comparison. We report the cumulative expected regret and the cumulative number of constraint violations. SCLTS is the constrained approach developed in [7] for single-bandit problems with a known context. Since the contexts are unknown SCLTS algorithm uses the noisy feature vectors for constructing the safe action set, which will lead to constraint violations as demonstrated in the toy example in Section A and in Fig. 1(b). DisLinUCB does not cater to constraints and, hence, will have more violations. On the other hand, SCLTS and DisLinUCB present smaller regrets given that they are loosely constrained and unconstrained, respectively, and hence perform more explorations in the initial stage. DiSC-UCB plays conservative actions in the initial rounds, resulting in a comparatively larger regret, however, zero violations as expected as shown in Fig. 1(a) and 1(b). In Fig. 1(c) and 1(d), we compared our proposed DiSC-UCB algorithm with three other benchmark distributed/federated approaches. Figures show that the three benchmarks, which are unconstrained, result in large constraint violations, and hence, these approaches are unsuitable for the hard-constrained problem considered in this paper.

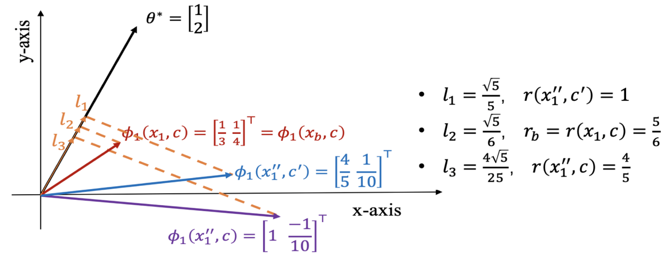


Fig. 3. An example demonstrating how an unsafe action x_1'' under true context appears to be safe under noisy context observation leading to incorrect conclusions about safe actions.

C. Regret Versus System Parameters

In addition to the results reported in the main paper, we plot the regret by varying α and the reward at every round of learning to ensure that the constraints are met. We present the plots showing the variation of the cumulative regret with respect to round t for different values of α for the synthetic data in Fig. 2(b). We observe that as the value of α increases, the cumulative regret decreases, which is expected since a larger value of α implies less strict performance constraint. In Fig. 2(a), we represent the reward plot. From the plot, we observe that the per-step reward is always larger than the baseline reward shown in the dotted line, which validates that our proposed algorithm DiSC-UCB satisfies the performance constraint at every stage of learning. It also shows the improvement in the reward as the learning round progresses. Fig. 2(c) demonstrates how agent collaboration improves the per-agent cumulative regret. For movielens data, Fig. 2(d) shows the reward plot at each round verifying constraints are met at every round. Fig. 2(e) and 2(f) provide the cumulative regret as M and α are varied. For the LastFM data, Fig. 2(g) presents the reward plot and Fig. 2(h) and 2(i) rovide the cumulative regret as M and α are varied.

VII. Conclusion

We studied the multi-task stochastic linear CB problem with stage-wise constraints when the agents observe only the context distribution and the exact contexts are unknown. We proposed a UCB algorithm, referred to as DiSC-UCB. For d-dimensional linear bandits, we prove an $O(d\sqrt{MT}\log^2 T)$ regret bound and an $O(M^{1.5}d^3)$ communication bound on the algorithm. We extended to the setting where the baseline rewards are unknown and showed that the same bounds hold for regret and communication. We empirically validated the performance of our algorithm on synthetic data and on real-world movielens-100 K and LastFM data and compared with benchmarks, SCLTS and DisLinUCB. As part of the future work, we plan to investigate budget constraints where the goal is to minimize the regret before using the total budget.

Appendix
Toy Example

We present a toy example in Fig. 3. Consider $M = 3$ agents with reward parameters $\theta_1^\star = [1, 2]^\top$, $\theta_2^\star = [1]$, and

$\theta_3^\star = [2]$, and $\alpha \in [0,1]$. The corresponding index set is $\mathcal{I} = \{\{1,2\},\{1\},\{2\}\}$. That is, the number of features is 2, and we consider 3 tasks. While all three features are relevant to task 1, only the first feature is relevant to task 2, and only the second feature is relevant to task 3. For a context $c \in \mathcal{C}$, let each agent i has three actions, denoted by $\mathcal{A} = \{x_i, x_i', x_i''\}$. Let the feature vector set Φ_i for agent i is

$$\Phi_1 = \begin{array}{c} \begin{matrix} \phi_{x_1,c} & \phi_{x_1',c} & \phi_{x_1'',c} \end{matrix} \\ \begin{bmatrix} \frac{1}{3} & 1 & 1 \\ \frac{1}{4} & 0 & -\frac{1}{10} \end{bmatrix} \end{array}, \Phi_2 = \begin{array}{c} \begin{matrix} \phi_{x_2,c} & \phi_{x_2',c} & \phi_{x_2'',c} \end{matrix} \\ \begin{bmatrix} \frac{1}{3} & 1 & 1 \end{bmatrix} \end{array},$$

$$\Phi_3 = \begin{array}{c} \begin{matrix} \phi_{x_3,c} & \phi_{x_3',c} & \phi_{x_3'',c} \end{matrix} \\ \begin{bmatrix} \frac{1}{4} & 0 & -\frac{1}{10} \end{bmatrix} \end{array}.$$

Using the index set $\mathcal{I}$ and appending zero elements, we map the reward parameter $\theta_i^\star$ and feature vector sets Φ_i to a common reward parameter $\theta^\star$ and heterogeneous feature vector sets Φ_i' as given below, where $\theta^\star = \theta_1^\star = [1\ 2]^\top$.

$$\Phi_1' = \begin{array}{c} \begin{matrix} \phi_1(x_1,c) & \phi_1(x_1',c) & \phi_1(x_1'',c) \end{matrix} \\ \begin{bmatrix} \frac{1}{3} & 1 & 1 \\ \frac{1}{4} & 0 & -\frac{1}{10} \end{bmatrix} \end{array},$$

$$\Phi_2' = \begin{array}{c} \begin{matrix} \phi_2(x_2,c) & \phi_2(x_2',c) & \phi_2(x_2'',c) \end{matrix} \\ \begin{bmatrix} \frac{1}{3} & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \end{array},$$

$$\Phi_3' = \begin{array}{c} \begin{matrix} \phi_3(x_3,c) & \phi_3(x_3',c) & \phi_3(x_3'',c) \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{4} & 0 & -\frac{1}{10} \end{bmatrix} \end{array}$$

For context c, action x_1' is the optimal action for agent 1. Consider a scenario, such as weather prediction or stock market prediction, where agents only observe a context distribution μ and the true context c is unknown. Let $c' := \mathbb{E}_\mu[\mathcal{C}]$ and let feature vectors for c' for agent 1 is

$$\Phi_1'' = \begin{array}{c} \begin{matrix} \phi_1(x_1,c') & \phi_1(x_1',c') & \phi_1(x_1'',c') \end{matrix} \\ \begin{bmatrix} \frac{1}{4} & \frac{6}{5} & \frac{4}{5} \\ \frac{1}{3} & -\frac{1}{5} & \frac{1}{10} \end{bmatrix} \end{array}.$$

Let action x_1 be the baseline action, i.e., $x_1 = x_b$. x_1'' do not satisfy the performance constraint for context c. Similarly, x_1' does not meet the performance constraint for c'. Consequently, the feasible action set for c is $\mathcal{A}_c = \{x_1, x_1'\}$, whereas for context c' it is $\mathcal{A}_{c'} = \{x_1, x_1''\}$. Given that performance constraints are considered in the exact context c, the selection should have been made from $\mathcal{A}_c$ and not $\mathcal{A}_{c'}$. As x_1'' is not within the feasible action set $\mathcal{A}_c$, the agent will choose the baseline action x_b to ensure it meets the performance constraint. Although the feasible action set $\mathcal{A}_c$ includes x_1', it is not present in $\mathcal{A}_{c'}$, hence limiting the agent's exploration capability and preventing the selection of the optimal action x_1'. This violates Lemma C.1 proposed in [7], which claims that the optimal action is always present in the feasible action set. This demonstrates the significance of a new approach when the agent can only observe the context distribution.

REFERENCES

[1] J. Kirschner and A. Krause, "Stochastic bandits with context distributions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 14113–14122.

[2] S. Lamprier, T. Gisselbrecht, and P. Gallinari, "Profile-based bandit with unknown profiles," *J. Mach. Learn. Res.*, vol. 19, no. 1, pp. 2060–2099, 2018.

[3] A. Kazerouni, M. Ghavamzadeh, Y. Abbasi Yadkori, and B. Van Roy, "Conservative contextual linear bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 3913–3922.

[4] Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvári, "Conservative bandits," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1254–1262.

[5] J. Yang and S. Ren, "Robust bandit learning with imperfect context," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 12, pp. 10594–10602.

[6] K. Khezeli and E. Bitar, "Safe linear stochastic bandits," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 10202–10209.

[7] A. Moradipari, C. Thrampoulidis, and M. Alizadeh, "Stage-wise conservative linear bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 11191–11201.

[8] Y. Zhang and Q. Yang, "An overview of multi-task learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 30–43, 2018.

[9] J. Wang, M. Kolar, and N. Srerbo, "Distributed multi-task learning," in *Proc. Artif. Intell. Statist.*, 2016, pp. 751–760.

[10] J. Lin and S. Moothedath, "Distributed stochastic bandits with hidden contexts," in *Proc. 2023 Eur. Control Conf.*, 2023, pp. 1–6.

[11] J. Lin and S. Moothedath, "Distributed stochastic bandit learning with delayed context observation," in *Proc. 2023 Eur. Control Conf.*, 2023, pp. 1–6.

[12] J. Lin and S. Moothedath, "Federated learning for heterogeneous bandits with unobserved contexts," in *Proc. IEEE Int. Symp. Inf. Theory*, 2024, pp. 2086–2091.

[13] A. A. Deshmukh, U. Dogan, and C. Scott, "Multi-task learning for contextual bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 4848–4856.

[14] M. Fang and D. Tao, "Active multi-task learning via bandits," in *Proc. 2015 SIAM Int. Conf. Data Mining*, 2015, pp. 505–513.

[15] L. Cella, K. Lounici, G. Pacreau, and M. Pontil, "Multi-task representation learning with stochastic linear bandits," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2023, pp. 4822–4847.

[16] J. Hu, X. Chen, C. Jin, L. Li, and L. Wang, "Near-optimal representation learning for linear bandits and linear RL," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4349–4358.

[17] J. Yang, W. Hu, J. D. Lee, and S. S. Du, "Impact of representation learning in bandits. In Proceedings of the 9th International Conference on Learning Representations," 2021.

[18] L. Cella and M. Pontil, "Multi-task and meta-learning with sparse linear bandits," in *Proc. Uncertainty Artif. Intell.*, 2021, pp. 1692–1702.

[19] J. Lin, S. Moothedath, and N. Vaswani, "Fast and sample efficient multi-task representation learning in stochastic contextual bandits," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 30227–30251.

[20] R. Lu, A. Zhao, S. S. Du, and G. Huang, "Provable general function class representation learning in multitask bandits and MDP," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 11507–11519.

[21] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, vol. 24, pp. 2312–2320.

[22] J. Lin and S. Moothedath, "Feature selection in distributed stochastic linear bandits," in *Proc. Amer. Control Conf.*, 2023, pp. 3939–3944.

[23] Y. Wang, J. Hu, X. Chen, and L. Wang, "Distributed bandit learning: Near-optimal regret with efficient communication," in *Proc. of Int. Conf. on Learn. Representations*, 2019, pp. 1–31.

[24] R. Huang, W. Wu, J. Yang, and C. Shen, "Federated linear contextual bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 27057–27068.

[25] N. Korda, B. Szorenyi, and S. Li, "Distributed clustering of linear bandits in peer to peer networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1301–1309.

[26] J. Zhu, A. Koppel, A. Velasquez, and J. Liu, "Byzantine-resilient decentralized multi-armed bandits," *Trans. on Mach. Learn. Res.*, 2024.

[27] J. Zhu, E. Mulle, C. S. Smith, A. Koppel, and J. Liu, "Decentralized upper confidence bound algorithms for homogeneous multi-agent multi-armed bandits," *IEEE Trans. Autom. Control*, early access, Jan. 2, 2025, doi: 10.1109/TAC.2024.3525417.

[28] B. Hao, T. Lattimore, and M. Wang, "High-dimensional sparse linear bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 10753–10763.
[29] T. Lin, B. Abrahao, R. Kleinberg, J. Lui, and W. Chen, "Combinatorial partial monitoring game with linear feedback and its applications," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 901–909.
[30] J. Kirschner, T. Lattimore, and A. Krause, "Information directed sampling for linear partial monitoring," in *Proc. Conf. Learn. Theory*, 2020, pp. 2328–2369.
[31] J. Kirschner, T. Lattimore, and A. Krause, "Linear partial monitoring for sequential decision making: Algorithms, regret bounds and applications," *J. Mach. Learn. Res.*, vol. 24, no. 1, pp. 16606–16650, 2023.
[32] S. Amani, M. Alizadeh, and C. Thrampoulidis, "Linear stochastic bandits under safety constraints ochastic Taylor derivative estimator: Efficient amortization for arbitrary differential operators," *In Adv. in Neural Inf. Process. Syst.*, 2019.
[33] Y. Mansour, A. Slivkins, and V. Syrgkanis, "Bayesian incentive-compatible bandit exploration," in *Proc. ACM Conf. Econ. Comput.*, 2015, pp. 565–582.
[34] S. Katariya, B. Kveton, Z. Wen, and V. K. Potluru, "Conservative exploration using interleaving," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2019, pp. 954–963.
[35] Y. Sui, A. Gotovos, J. Burdick, and A. Krause, "Safe exploration for optimization with Gaussian processes," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 997–1005.
[36] K. N. Varma, S. Lale, and A. Anandkumar, "Stochastic linear bandits with unknown safety constraints and local feedback," in *Proc. Int. Conf. Mach. Learn. Workshop New Front. Learn., Control, Dynamical Syst.*, 2023.
[37] J. Lin, X. Y. Lee, T. Jubery, S. Moothedath, S. Sarkar, and B. Ganapathysubramanian, "Stochastic conservative contextual linear bandits," in *Proc. IEEE Conf. Decis. Control*, 2022, pp. 7321–7326.
[38] E. Altman, *Constrained Markov Decision Processes*, France, Routledge, 2021.
[39] K. R. Kaza, R. H. Meshram, V. Mehta, and S. N. Merchant, "Constrained restless bandits for dynamic scheduling in cyber-physical systems," *IEEE Access*, vol. 12, pp. 182274–182295, 2024.
[40] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Proc. Annu. Conf. Learn. Theory*, 2008, pp. 355–366.
[41] S. Chaudhary and D. Kalathil, "Safe online convex optimization with unknown linear safety constraints," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 6, pp. 6175–6182.
[42] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interactive Intell. Syst.*, vol. 5, no. 4, pp. 1–19, 2015.
[43] I. Cantador, P. Brusilovsky and T. Kuflik, "2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011). In Proceedings of the 5th ACM conference on Recommender systems," *RecSys 2011*, New York, NY, USA: ACM, 2011.

Jiabin Lin received the B.S. degree in computer science and communication Engineering from the Guangxi University of Science and Technology, Liuzhou, China, in 2019, and the M.S. degree in electrical engineering from the University of South Florida, Tampa, FL, USA, in 2021. He is currently a graduate student with the Department of Electrical and Computer Engineering with Iowa State University, Ames, IA, USA . His research interests include bandit learning, multi-task learning, and representation learning.

Shana Moothedath received the Ph.D. degree in electrical engineering from Indian Institute of Technology Bombay (IITB), Mumbai, Maharashtra, India, in 2018. She was a Postdoctoral Scholar in electrical and computer engineering with the University of Washington, Seattle (WA), USA, till 2021. She is currently Harpole-Pentair Assistant Professor with the Department of Electrical and Computer Engineering with Iowa State University, Ames, IA, USA. Her research interests include distributed decision-making, control and security of cyber-physical systems, game theory, sparse signal processing to improve system security, scalability, and efficiency. She was awarded the Best Research Thesis Award at IITB in 2019 and selected as a MIT-EECS Rising Star in 2019.