HETEROGENEOUS MULTI-AGENT CONTEXTUAL BANDITS UNDER BOUNDED MISSPECIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Misspecification is known to degrade the performance of contextual bandit algorithms, yet its impact has received little attention in the multi-agent setting. Motivated by this gap, we formulate the setting of arm-heterogeneous multi-agent linear contextual bandits under misspecification—a practically motivated model where each agent has access to only a subset of the global arm catalog. To minimize the joint regret of all agents without prior knowledge of the misspecification level, we propose HM-UCB, a collaborative algorithm that updates its parameter estimates via *size-weighted pooling* of agent observations. We first establish a high-probability regret upper bound of $\tilde{O}(R\sqrt{dNT\log(KN)} + \varepsilon\sqrt{d}\,NT)$ for HM-UCB. Next, by using geometric packing and information-theoretic techniques, we derive a generic regret lower bound in our model, which along with the upper bound of HM-UCB, establishing the near-optimality of HM-UCB in some practical regimes. To demonstrate the scalability of our framework, we instantiate HM-UCB under two practical protocols—Naïve and N-Selection. For each, we design customized algorithms and establish corresponding theoretical guarantees. Empirical evaluations corroborate our theoretical findings, underscoring the strong practical performance of HM-UCB and substantiating the tightness of our regret bounds.

1 INTRODUCTION

While modern regional platforms typically operate independently to serve local users, cross-platform performance can be significantly enhanced through collaborative learning. Consider online product recommendation across multiple platforms: each platform sequentially recommends items to its users from a local product catalog, observes user feedback as rewards, and updates its recommendation strategy. Due to geographic coverage, merchant partnerships, or inventory constraints, the available catalogs are heterogeneous across platforms. Despite these differing catalogs, all platforms share an invariant mapping from user-product features to expected rewards, allowing them to collaborate and accelerate learning. When the mapping can be modeled by a common linear function, this sequential recommendation process can be formulated as the learners' arm selection tasks in a linear contextual bandit (Li et al., 2010; Chapelle, 2014; Yang et al., 2020). Collaboration among platforms motivates the formulation of *multi-agent linear contextual bandits*, in which agents benefit from shared information (Dubey & Pentland, 2020a; Moradipari et al., 2022).

In reality, the expected rewards may not admit a fixed linear representation in the contextual features. Recall the example above, the platform reward may depend on joint effects of product price, promotional discounts and real-time user intent, which may not be fully captured by the available features. This *unavoidable* reward discrepancy is a vivid example of *model misspecification*: the expected reward differs from the best linear predictor by a bounded approximation error (Ghosh et al., 2017). Such scenario is commonly formalized as an *ε-bounded misspecified linear bandit*. While misspecification-robust algorithms have been established for a single agent (Foster et al., 2020; Takemura et al., 2021), existing multi-agent bandit frameworks predominantly rely on perfect linear reward assumptions (Moradipari et al., 2022; Do et al., 2023), highlighting a key methodological gap.

As each recommendation platform observes only a subset of the global product catalog, such arm heterogeneity naturally arises in real-world applications (Do et al., 2023; Yang et al., 2022). How-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ever, it remains largely unexplored in multi-agent contextual bandits under misspecification. To solve the platform's problem, we study arm-heterogeneous multi-agent linear contextual bandits under bounded misspecification and develop efficient algorithms for regret minimization in this setting.

1.1 CONTRIBUTION

Our main contributions are summarized as follows.

- **Model Specification.** We formulate misspecified multi-agent linear contextual bandits in which each agent selects arms from a local subset of a shared global catalog and arm rewards deviate from any linear parameter. We describe such reward discrepancy with global misspecification level ε, which is a weighted aggregation of agent-wise misspecification ε_n. The weight of ε_n is determined by the local catalog of agent n (Definition 1). This global misspecification ε characterizes heterogeneous arm availability and emphasizes the effect of local reward discrepancy on performance of the multi-agent system.
- **Algorithm Efficiency.** We design **HM-UCB** (Algorithm 1), which incorporates a size-weighted pooling mechanism into its parameter update, and establish its regret bound of

$$\tilde{O}\Big(R\sqrt{dNT\log(KN)} + \varepsilon\sqrt{d}\,NT + Bd\Big),$$

  where the first term stems from environment uncertainty as in standard linear bandit, and the second term reflects the cost of handling misspecification. We also establish a generic lower bound, which captures the problem hardness with both geometric packing and information-theoretic techniques. This lower bound nearly matches HM-UCB's upper bound, suggesting the near-optimality of HM-UCB in some practical regimes.
- **Scalability and Practical Efficacy.** We instantiate the framework of HM-UCB in two relevant protocols and establish theoretical regret guarantees for both: in the Naïve protocol, all agents share a common catalog; in the N-Selection protocol, the agents must pull N distinct arms in each round. Numerical experiment results corroborate our theoretical regret scalings and the robustness of HM-UCB against heterogeneous catalog sizes.

1.2 RELATED WORK

Bandit learning is commonly studied under two objectives: *regret minimization* (RM) and *best-arm identification* (BAI). Under RM, a learner sequentially selects arms over a horizon to minimize its cumulative regret, defined as the cumulative reward gap relative to an optimal oracle. Under BAI, the goal is to identify an optimal or near-optimal arm using as few samples as possible. This work aims to minimize the joint regret of all agents under model misspecification and arm heterogeneity in the multi-agent linear contextual setting. We now review some related work.

Linear Contextual Bandits (LCB). The LinUCB (Li et al., 2010) and OFUL algorithms (Abbasi-Yadkori et al., 2011) are UCB-type methods for standard linear contextual bandits. A stage-wise variant of LinUCB, SupLinUCB, attains $\tilde{O}(\sqrt{dT})$ regret with a finite catalog (Chu et al., 2011), when LinUCB and OFUL attain $\tilde{O}(d\sqrt{T})$ regret with either finite or infinite catalog (Abbasi-Yadkori et al., 2011). LinTS (Agrawal & Goyal, 2013; Abeille & Lazaric, 2017) is a Thompson-sampling alternative with regret of $\tilde{O}(d^{3/2}\sqrt{T})$. The generic regret lower bound is $\tilde{\Omega}(\sqrt{dT})$ or $\tilde{\Omega}(d\sqrt{T})$ when there is a finite or infinite catalog (Abbasi-Yadkori et al., 2011; Lattimore et al., 2020), which help us observe the optimality of these algorithms up to logarithmic factors. Linear BAI has also been studied: LT&S (Jedra & Proutiere, 2020) studies fixed-confidence identification setting, and OD-LinBAI (Yang & Tan, 2022) studies the fixed-budget setting.

Misspecified Linear Bandits. While LCB assumes expected rewards are strictly linear in the features, Ghosh et al. (2017) introduces the misspecified linear bandits, a more practical setting which allows a bounded discrepancy of rewards from any linear predictor. Such reward discrepancy is termed as *misspecification level*. On one hand, Ghosh et al. (2017) shows that standard optimism-based methods such as OFUL fail to minimize regret under misspecification. Besides, in this setting, Zhang et al. (2023) proposes the DS-OFUL algorithm which requires the prior knowledge of misspecification level ε, and Foster et al. (2020) designs an oracle-efficient algorithm which does not

know ε. Moreover, Takemura et al. (2021) adapts SupLinUCB to this misspecified setting; for brevity, we refer to this variant as M-SupLinUCB hereafter. On the other hand, recent works also study how model bias affects best arm identification. For example, Réda et al. (2021) designs MISLID for fixed-confidence linear Top-m identification and Alieva et al. (2021) investigates the fixed-budget setting.

Multi-Agent Bandits. In the multi-agent setting, agents share information to facilitate local decision-making (Zhu et al., 2024). Existing formulations differ along several axes, including interaction objective, communication pattern, and environment structure. Along the interaction objective, cooperative methods such as MA-LinUCB (Moradipari et al., 2022) enable agents to collaborate for minimization of a shared or aggregated regret, whereas competitive formulations model agents as opponents in a game (Li & Duan, 2025; Liu et al., 2020). Along the communication pattern, Landgren et al. (2016) studies distributed cooperative bandits over a network, while FedUCB (Dubey & Pentland, 2020a) is designed in a centralized federated system. Along the environment structure, Do et al. (2023) assumes that agents have dissimilar linear parameters and designs H-LINUCB for RM. Besides, Yang et al. (2022) and Dubey & Pentland (2020b) studies heterogeneous contexts or arm catalogs. In particular, the arm-heterogeneous setting assumes each agent pulls arms from a possibly distinct subset of a shared global catalog. All these works focus on linear reward model.

2 Problem Formulation

We study the **arm-heterogeneous multi-agent setting**. Hereafter, for brevity, we refer to this setting as **heterogeneous setting**. For any $n \in \mathbb{N}$, denote $[n] := \{1, \dots, n\}$. Let there be $N \in \mathbb{N}$ agents indexed by $[N]$, a horizon of $T \in \mathbb{N}$ rounds, and a global arm catalog $[K] := \{1, \dots, K\}$. Each agent $n \in [N]$ is equipped with a local catalog $[K_n] \subset [K]$ from which it pulls arms. At round $t \in [T]$, arm $i \in [K]$ is associated with a feature vector $x_t(i) \in \mathbb{R}^d$. If agent n pulls arm $i_{t,n} \in [K_n]$, it receives a random reward $r_{t,n}$ with mean $\mu_{t,n}(i_{t,n})$. The noise $\eta_{t,n} := r_{t,n} - \mu_{t,n}(i_{t,n})$ is conditionally R-sub-Gaussian given $\mathcal{F}_{t-1}$, that is, for all $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp(\lambda \eta_{t,n}) \mid \mathcal{F}_{t-1}] \le \exp\big(\lambda^2 R^2/2\big)$, where

$$\mathcal{F}_{t-1} := \sigma\big(\{x_s(i)\}_{s\in[t-1],\, i\in[K]},\ \{i_{s,n}, r_{s,n}\}_{s\in[t-1],\, n\in[N]}\big)$$

is the observation history. The expected rewards $\mu_{t,n}(i_{t,n})$ are assumed to be approximately linear in the observed features through a shared parameter θ, as formalized below.

Definition 1 (Misspecification in Heterogeneous Setting)**.** There exist $\varepsilon \ge 0$ and $\theta \in \mathbb{R}^d$ such that for all $n \in [N]$, $t \in [T]$ and $i_n \in [K_n]$,

$$\mu_{t,n}(i_{t,n}) = \theta^\top x_t(i_n) + \varepsilon_{t,n}(i_n),$$

where $|\varepsilon_{t,n}(i_n)| \le \varepsilon_n$ and $\sum_{n=1}^{N} \frac{|K_n|}{|K|}\varepsilon_n \le \varepsilon$. The *misspecification level* is the smallest such ε; the resulting setting is called ε*-bounded misspecified*.

We note that when $\varepsilon = 0$, this setting reduces to a standard linear contextual bandit. At round $t = 1, \dots, T$, each agent n first observes the feature vectors $\{x_t(i)\}_{i\in[K_n]}$. Given the history $\mathcal{F}_{t-1}$ and $\{x_t(i)\}_{i\in[K_n]}$, agent n selects an arm $i_{t,n} \in [K_n]$ and observes the corresponding reward $r_{t,n}$. The history for all agents then advances to

$$\mathcal{F}_t := \sigma\big(\{x_s(i)\}_{s\in[t],\, i\in[K]},\ \{i_{s,n}, r_{s,n}\}_{s\in[t],\, n\in[N]}\big)\,.$$

Throughout the horizon, the expected reward $\mu_{t,n}$, the shared parameter θ, the agent-wise misspecification levels ε_n, as well as the global misspecification level ε, are unknown to all agents.

An algorithm aims to minimize its *cumulative joint regret* of all agents

$$R(T) = \sum_{t\in[T]} \sum_{n\in[N]} \big(\mu_{t,n}(i^*_{t,n}) - \mu_{t,n}(i_{t,n})\big)\,,$$

where $i^*_{t,n} \in \arg\max_{i\in[K_n]} \mu_{t,n}(i)$ is an optimal arm in the local catalog of agent n at round t. Without loss of generality, assume there exist positive constants L, M and B such that (1) $\|x_t(i)\|_2 \le L$ for all $i \in [K]$, $n \in [N]$, $t \in [T]$; (2) $\|\theta\|_2 \le M$; (3) $|\mu_{t,n}(i) - \mu_{t,n}(j)| \le B$ for all $i, j \in [K]$, $n \in [N]$, $t \in [T]$.

3 ALGORITHM

In this section, we introduce **Heterogeneous Multi-agent SupLinUCB** (HM-UCB), which adapts the design of M-SupLinUCB (Chu et al., 2011; Takemura et al., 2021) to the heterogeneous setting. We first describe its procedure, which is also presented in Algorithm 1. Next, we elaborate the *size-weighted pooling* design in HM-UCB, a key design that sets our algorithm apart from existing methods and underpins its efficacy.

HM-UCB begins with initializing the matrices $V_{1,s}$ and $b_{1,s}$ as λI_d and $\mathbf{0}_d$, respectively. At round t, with $\{x_t(i)\}_{i=1}^{K_n}$ observed, agent n initializes $I_{t,n} = K_n$ and stage $s = 1$, and next compute the predictor $\hat{\theta}_{t,s}$ from $V_{t,s}$ and $b_{t,s}$. Subsequently, the agent computes confidence widths $w_{t,n,s}(i) = \alpha\|x_t(i)\|_{V_{t,s}^{-1}}$ and enters an inner loop. Concretely, each agent inspects its current confidence widths and proceeds along one of three branches: (i) Confidence branch: if all widths are sufficiently narrow, the agent commits to the UCB maximizer and the inner loop terminates; (ii) Stage branch: if all widths lie below the stage threshold, the agent eliminates low-UCB arms, increments s, and repeats the width test at the new stage; (iii) Exploration branch: otherwise the agent pulls an arm with a sufficiently large confidence width, records (t, n) in the global exploration set $\Psi_{t,s}$, and the inner loop terminates. Thus the inner loop iterates only on the Stage branch; besides, on the other two branches, the agent selects an arm and terminates the loop. After all agents have acted, $\Psi_{t,s}$ is used to update the shared parameters $(V_{t+1,s}, b_{t+1,s})$ through outer products and weighted rewards.

Algorithm 1 Heterogeneous Multi-agent SupLinUCB (HM-UCB)

Input: N, T, K, arm subsets $K_n \subseteq [K]$, dimension d, failure probability $\delta \in (0,1)$, $c > 1$, $\lambda > 0$
Set $\alpha \leftarrow R\sqrt{2\log(2KSNT/\delta)} + \sqrt{\lambda}\,M$ and $S \leftarrow \lceil \log_c(T/d)/2 \rceil$
Parameter Initialization: $V_{1,s} \leftarrow \lambda I_d$ and $b_{1,s} \leftarrow \mathbf{0}_d$
for round $t = 1$ to T **do**
  $\Psi_{t,s} \leftarrow \emptyset$
  for agent $n = 1$ to N **do**
    Observe arm features $\{x_t(i)\}_{i\in[K_n]}$
    $s \leftarrow 1$
    $I_{t,n} \leftarrow K_n$
    while true **do**
      Compute shared parameter $\hat{\theta}_{t,s} \leftarrow V_{t,s}^{-1} b_{t,s}$
      for each arm $i \in I_{t,n}$ **do**
        $\hat{r}_{t,n,s}(i) \leftarrow x_t(i)^\top \hat{\theta}_{t,s}$
        Global exploration set $w_{t,n,s}(i) \leftarrow \alpha\|x_t(i)\|_{V_{t,s}^{-1}}$
      end for
      if $\forall i \in I_{t,n}, w_{t,n,s}(i) \le \alpha\sqrt{d/NT}$ **then** ▷ Confidence branch
        Choose arm $i_{t,n} \leftarrow \arg\max_{i\in I_{t,n}} [\hat{r}_{t,n,s}(i) + w_{t,n,s}(i)]$ and receive reward $r_{t,n}$
        Global exploration set $\Psi_{t+1,s} \leftarrow \Psi_{t,s}$
        break
      else if $\forall i \in I_{t,n}, w_{t,n,s}(i) \le \alpha c^{-s}$ **then** ▷ Stage branch
        $U_{t,n} \leftarrow \max_{j\in I_{t,n}} [\hat{r}_{t,n,s}(j) + w_{t,n,s}(j)] - 2\alpha c^{1-s}$
        $I_{t,n} \leftarrow \{i \in I_{t,n} \mid \hat{r}_{t,n,s}(i) + w_{t,n,s}(i) \ge U_{t,n}\}$
        $s \leftarrow s + 1$
      else ▷ Exploration branch
        Choose arm $i_{t,n} \leftarrow i \in \{i \in I_{t,1} \mid w_{t,n,s}(i) > \alpha c^{-s}\}$ and receive reward $r_{t,n}$
        $\Psi_{t+1,s} \leftarrow \Psi_{t,s} \cup \{(t,n)\}$
        break
      end if
    end while
  end for
  $V_{t+1,s} \leftarrow V_{t,s} + \sum_{(t,n)\in\Psi_{t,s}} x_t(i_{t,n}) x_t(i_{t,n})^\top$
  Update weighted reward $b_{t+1,s} \leftarrow b_{t,s} + \sum_{(t,n)\in\Psi_{t,s}} \frac{|K_n|}{|K|} r_{t,n} x_t(i_{t,n})$
end for

We now detail the *size-weighted pooling* design in our algorithm. HM-UCB does not require the misspecification level ε as input. To achieve robustness without knowing ε, it maintains the *global exploration set* $\Psi_{t,s}$, *pooling* the observation of agents who enter the Exploration branch at stage s in

round t. All agents share a single linear predictor $\hat{\theta}_{t,s}$ built with statistics in this global exploration set $\Psi_{t,s}$ (line 11), which is the only cross-agent communication channel. The implementation of this channel is highly flexible and can be tailored to various operational constraints: when communication is costly, agents may upload stage histories to a central aggregator and download computed $\hat{\theta}_{t,s}$; when privacy is a concern, agents may compute (V, b) locally and exchange only the resulting parameters. The *weighted reward update* (line 32) scales the observation of each agent by the size of its catalog. Since an agent with a larger catalog typically provides more information that facilitate the learning of the entire system, its observation is assigned a larger weight in the parameter update. Relative to applying M-SupLinUCB independently, this *size-weighted pooling* is the distinctive ingredient of HM-UCB: it converts local, catalog-constrained observations into a single shared estimator without prior knowledge of ε, and thereby keeps the algorithm informative in ε-bounded misspecified heterogeneous setting.

4 Theoretical Findings

This section presents our key theoretical results. We first establish a high-probability upper bound on the regret of HM-UCB, quantifying both its regret efficiency and misspecification robustness. We also derive a generic lower bound, which along with the upper bound of HM-UCB indicates the near-optimality of HM-UCB in some practical regimes. The proof sketch and detailed proofs are deferred to Appendix C.

4.1 Upper bound

Theorem 1 (Regret upper bound of HM-UCB)**.** *Let $c - 1 = \Theta(1)$ and $\lambda = R^2 M^{-2}$. With probability at least $1 - \delta$, the cumulative regret of Algorithm 1 satisfies*

$$R(T) \in \tilde{O}\left(R\sqrt{dNT\log(KN)} + \varepsilon\sqrt{d}\,NT + Bd\right),$$

where $\tilde{O}(\cdot)$ hides polylogarithmic factors in (d, T, δ).

The first term in Theorem 1 reflects regret efficiency of HM-UCB and the second term reflects its misspecification robustness. To position this result more clearly, Table 1 compares representative bounds across four setting: single-agent contextual bandit, multi-agent contextual bandit, single-agent misspecified bandit, and our misspecified heterogeneous setting.

Table 1: Comparison of representative bounds across related settings.*DS-OFUL requires $\varepsilon \leq \tilde{O}(\Delta/\sqrt{d})$ with a known sub-optimality gap Δ.

Setting	Method/Lower bound	Bound
Single-agent linear contextual bandit	LinUCB (Chu et al., 2011)	$\tilde{O}(d\sqrt{T})$
	Lower bound (Chu et al., 2011)	$\tilde{\Omega}(d\sqrt{T})$
Multi-agent linear contextual bandit	H-LINUCB (Do et al., 2023)	$\tilde{O}(d\sqrt{NT})$
	Lower bound (Do et al., 2023)	$\tilde{\Omega}(d\sqrt{NT})$
Single-agent misspecified bandit	M-SupLinUCB (Takemura et al., 2021)	$\tilde{O}(R\sqrt{dT\log K} + \varepsilon\sqrt{d}\,T + Bd)$
	DS-OFUL* (Zhang et al., 2023)	$\tilde{O}(d^2/\Delta)$
	Lower bound (Lattimore et al., 2020)	$\varepsilon\min\{T, \frac{K-1}{2}\}\sqrt{\frac{d-1}{8\log K}}$
Multi-agent misspecified bandit (our work)	HM-UCB (Theorem 1)	$\tilde{O}(R\sqrt{dNT\log(KN)} + \varepsilon\sqrt{d}\,NT + Bd)$
	Lower bound (Theorem 2)	$\max\{\varepsilon\min\{NT, \frac{N(K-N)}{N+1}\}\sqrt{\frac{d-1}{8\log K}}, \varepsilon NT\big(1 - \frac{N}{K} - \varepsilon\sqrt{\frac{NT}{2K}}\big)\}$

As presented in Table 1, the upper bound of HM-UCB shares a similar structure with that of M-SupLinUCB. When $N = 1$, it recovers to the bound of M-SupLinUCB, which supports the misspecification robustness of HM-UCB. Unlike DS-OFUL, which requires knowledge of a sub-optimality gap Δ and a constraint on ε, HM-UCB operates without knowledge of ε. In comparison, our bound scales as $\tilde{O}(\sqrt{dNT})$, improving over the $\tilde{O}(d\sqrt{NT})$ regret of H-LINUCB in the multi-agent setting. This $\sqrt{d}$ factor improvement stems from a stage-wise elimination mechanism, generalizing the superiority of SupLinUCB over LinUCB to the multi-agent setting. This comparison of HM-UCB against H-LINUCB manifests HM-UCB's efficiency to minimize the regret.

4.2 Lower bound

To obtain a generic lower bound in the heterogeneous setting, we investigate the N-Selection protocol, which is different from the one defined in Section 2. In this protocol, N agents cooperatively pull N distinct arms in every round to minimize the aggregated regret of those N pulls. The N-Selection protocol and its similarity with heterogeneous setting are elaborated in Section 5.2. We begin our analysis with the geometric packing technique under the following condition.

Condition 1. Fix $K \geq 2$ and assume the feature dimension d satisfies $d \geq \left\lceil \frac{8 \log K}{\varepsilon^2} \right\rceil$.

Under Condition 1, there exists a family of contextual instances whose unit-norm arm features are nearly orthogonal, satisfying ε-bounded misspecification (Lattimore et al., 2020). We refer to this collection of instances as the *packing family*. In each instance, the global catalog comprises a set of *good* arms and a complementary set of *bad* arms, each set sharing a uniform expected reward. We say the instances sharing this characteristic are *good-versus-bad*. With good-versus-bad instances from the packing family, we establish the former term in Theorem 2.

Subsequently, we construct $K+1$ good-versus-bad instances to analyze the regret with information-theoretic technique (Zhong et al., 2021) under the following condition.

Condition 2. Fix N, K, T with $2 \leq N \leq K$ and suppose $\frac{N}{K} + \varepsilon\sqrt{\frac{NT}{2K}} < 1$.

Define the cyclic index $[i]_K := 1 + ((i-1) \bmod K)$. For each $k \in [K]$, define the set of optimal arms in instance k as $S^{*,k} = \{[k]_K, [k+1]_K, \ldots, [k+N-1]_K\}$. Instance 0 assigns expected reward $\mu_i^{(0)} = 1/2$, $x_i^{(0)} = (1, 0)$ to all arms $i \in [K]$, which indicates $S^{*,0} = [K]$. Under alternative instance k, we have

$$\mu_i^{(k)} = \begin{cases} 1/2 + \varepsilon, & i \in S^{*,k} \\ 1/2, & i \notin S^{*,k} \end{cases}, \qquad x_i^{(k)} = \begin{cases} (1, 1/k) & i \in S^{*,k} \\ (1, 0) & i \notin S^{*,k} \end{cases}.$$

The latent vector θ is set as $(1/2, 0)$ in all instances. This family satisfies ε-bounded misspecification, and the sets of optimal arms in K alternative instances differ by a cyclic shift. We refer to this family as the *cyclic family*. With this cyclic family, we establish the latter term in Theorem 2.

Theorem 2 (Generic regret lower bound)**.** *Under Condition 1 on the packing family and Condition 2 on the cyclic family, for any algorithm, there exists a heterogeneous instance such that the expected cumulative regret satisfies*

$$R(T) \geq \max\left\{ \varepsilon \min\left\{ NT, \frac{N(K-N)}{N+1} \right\} \sqrt{\frac{d-1}{8 \log K}},\ \varepsilon NT \left(1 - \frac{N}{K} - \varepsilon\sqrt{\frac{NT}{2K}} \right) \right\},$$

where the former term requires Condition 1, whereas the latter term requires Condition 2.

Theorem 2, alongside Table 1, shows that when $N = 1$, our lower bound recovers the classic single-agent misspecified bound (Lattimore et al., 2020), substantiating that it tightly characterizes the intrinsic problem hardness. Furthermore, the two lower bound terms are complementary: the former term dominates when T is sufficiently large or $N \approx K$, whereas the latter type term yields a tighter threshold otherwise. Jointly, they offer a two-fold perspective on the fundamental difficulty of regret minimization in the misspecified heterogeneous setting.

4.3 On the Near-optimality of HM-UCB

To illustrate the efficiency of HM-UCB, we compare the upper bound in Theorem 1 and the generic lower bound in Theorem 2.

Corollary 1 (Tightness). *Let $c-1=\Theta(1)$, $\lambda=R^2M^{-2}$, and $\varepsilon^2 \le 4\alpha^2 d/NT \le M^2L^2$. Under Condition 1, there exists an instance from the packing family such that with probability at least $1-\delta$, we have*

$$R(T) \in \tilde{\Omega}\left(\left(\frac{R\sqrt{d\log(KN)}}{\sqrt{NT}} + \frac{\varepsilon}{2}\sqrt{\frac{d-1}{8\log K}}\right)\cdot \min\big\{NT, N(K-N)/(N+1)\big\}\right) \cap \tilde{O}\Big(R\sqrt{dNT\log(KN)} + \varepsilon\sqrt{d}\,NT + Bd\Big). \tag{1}$$

The upper bound in (1) is achieved by HM-UCB.

Corollary 1 establishes the near-optimality of HM-UCB. Notably, as shown in Table 1, the gap between upper and lower regret bounds also persists in the single-agent misspecified setting. Closing the remaining gap constitutes a promising direction for future research in both single-agent and multi-agent misspecified bandits.

5 Case Study

This section instantiates the HM-UCB framework under two relevant protocols. In the Naïve protocol, N agents operate over a common catalog and thus share identical optimal arms, allowing HM-UCB to be directly applied. In the N-Selection protocol, agents are required to select N distinct arms per round; for this setting, we introduce two variants of HM-UCB, which are NS-UCB and NQ-UCB. We also establish the regret upper bound for these algorithms and the generic lower bound in both protocols.

Conceptually, these protocols anchor the spectrum of optimal-arm diversity: denoting $K_t' \in [1, N]$ as the number of distinct optimal arms across agents at round t, the Naïve protocol ($K_t'=1$) and N-Selection protocol ($K_t'=N$) form the boundary regimes of the general heterogeneous setting ($1 \le K_t' \le N$). This structural connection explains why N-Selection serves as a pivotal bridge in our general lower-bound analysis in Section 4.2.

5.1 Naïve Multi-Agent Bandits

We first investigate the homogeneous specialization of the setting introduced in Section 2, which is called the Naïve protocol. In this Naïve protocol, agents share the same catalog $K_n = [K]$. Misspecification is described by the maximum of all agent-wise misspecification levels: $|\varepsilon_{t,n}(i_n)| \le \varepsilon_n \le \varepsilon$, and the cumulative regret is defined as in the heterogeneous setting.

HM-UCB can naturally adapt to this setting for regret minimization. When HM-UCB is applied, all agents actually pull the identical arm in a given round. The upper bound of HM-UCB in this Naïve protocol is identical to that in Theorem 1 with similar analysis. Again by constructing the packing family, we can establish the lower bound $R(T) \in \Omega\Big(\varepsilon \min\{NT, \frac{K-1}{2}\}\sqrt{\frac{d-1}{8\log K}}\Big)$, which recovers the single-agent bound of Lattimore et al. (2020) with horizon NT. This is plausible due to the close affinity between this protocol and a single-agent setting with a total step budget of NT.

5.2 N-Selection Multi-Agent Bandits

In the N-Selection protocol, N agents collaboratively assemble a size-N group of distinct arms from a shared catalog $[K]$ in each round. The definition of misspecification follows Definition 1. Let $I_t^\star$ denotes the N arms with largest rewards at round t. The maximum per-round reward is $\sum_{i\in I_t^\star}\mu_t(i)$, and the cumulative regret is defined as

$$R(T) = \sum_{t=1}^{T}\left(\sum_{i\in I_t^\star}\mu_t(i) - \sum_{n=1}^{N}\mu_{t,n}(i_{t,n})\right).$$

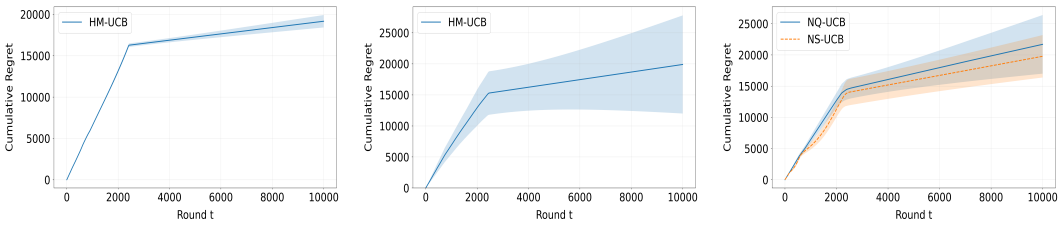
In this setting, we design two algorithms, NS-UCB and NQ-UCB, that retain the framework of HM-UCB, specially its *size-weighted pooling* mechanism. Without loss of generality, we assume agents act in the fixed order $1, 2, \ldots, N$ at every round. In this setting, the candidate catalog of agent n is the residual catalog after action selection of former agents, so $|K_n| = |K - n + 1|$. The two variants, NS-UCB and NQ-UCB, differ only in stage selection. NS-UCB forces every agent to pull arm from the same stage; NQ-UCB follows the same structure without this synchronization. Because all agents in NS-UCB share a common stage, the stage increments only when the Stage branch width test holds and the candidate catalog contains at least N arms after elimination. The pseudocode of NS-UCB is provided in Appendix E.

Under the N-Selection protocol, the upper bound for NS-UCB and the corresponding generic lower bound coincide with those established in Section 4. While we do not provide theoretical guarantees for NQ-UCB, empirical evaluations (see Section 6) show that it consistently underperforms NS-UCB. This performance discrepancy is likely attributed to the fact that stage synchronization aligns exploration sets across stages, thereby strengthening information coupling through the shared estimator.

6 Experiments

To corroborate the theoretical findings in Section 4 and Section 5, we conduct three synthetic studies of HM-UCB and its variants, observing their regret in the Naïve, N-Selection, and heterogeneous settings, as well as examining the effect of the misspecification level and demonstrating the algorithm robustness against catalog-size heterogeneity. All experiments are repeated across 30 random seeds, with mean values and standard deviations reported. Unless otherwise specified, we set $N = 10$, $K = 100$, $T = 10{,}000$, $d = 64$, and misspecification level $\varepsilon = 0.5$. In each round t, the misspecification terms $\varepsilon_{t,n}$ are drawn independently and uniformly from $[-\varepsilon_n, \varepsilon_n]$. Detailed experimental configurations are provided in Appendix F.

We evaluate algorithms in the heterogeneous setting and in the two protocols of Section 5: HM-UCB in the arm-heterogeneous and Naïve settings, as well as NQ-UCB and NS-UCB in N-Selection. For all settings, we construct good-versus-bad instances from the packing family, which has been introduced in Section 4.2.



(a) Heterogeneous Setting (b) Naïve Setting (c) N-Selection Setting

Figure 1: Cumulative regret in Heterogeneous, Naïve and N-Selection setting.

Figure 1 highlights three observations. First, all algorithms eventually enter a stable learning regime, with the cumulative regret exhibiting sublinear growth over the horizon. The curves do not become completely flat in the late-horizon regime due to the random bounded misspecification, which induces a persistent per-round regret even after the algorithms have a good estimation of the latent vector θ, corroborating the misspecification term in Theorem 1. Second, the three curves exhibit pronounced changes in slope over relatively short round windows. This behavior is consistent with the good-versus-bad structure of the instances: as many bad arms are eliminated, subsequent pulls concentrate on good arms with smaller instantaneous regret, and the regret growth speed decreases. Finally, in the N-Selection protocol, NS-UCB consistently achieves lower cumulative regret than NQ-UCB. This gap illustrates highlights the efficacy of the stage-synchronization design in NS-UCB, a benefit not realized in NQ-UCB.

To isolate the effect of misspecification on cumulative regret, we vary the global misspecification level $\varepsilon \in \{0.0, 0.1, \ldots, 0.5\}$ when keeping all other experimental parameters fixed. The results are

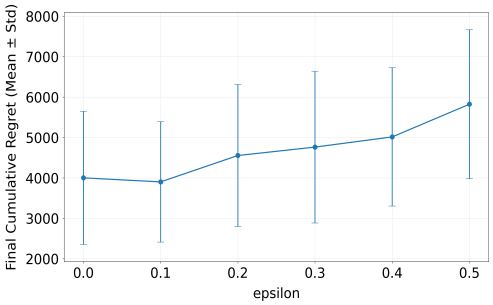


Figure 2: Final cumulative regret versus misspecification level ε.

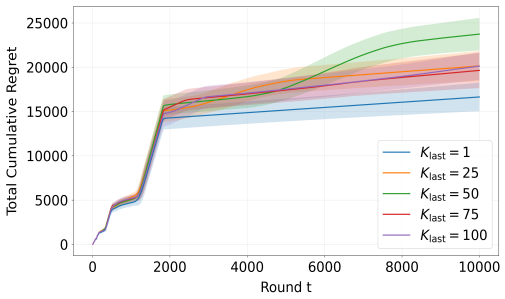


Figure 3: Arm-distribution ablation under heterogeneous K_{last} constructions.

presented in Figure 2, where we can observe an approximately linear increase in cumulative regret of HM-UCB as ε grows. This observation is consistent with our theoretical results in Section 4, in which the regret bound involves a misspecification term that scales linearly with the misspecification level ε.

Although one might expect the size of a local catalog to affect algorithm performance, Theorem 1 imposes no explicit penalty on the individual catalog sizes $|\mathcal{K}_n|$. To examine the impact of the catalog-size heterogeneity, we vary the catalog of the last agent while keeping the remaining settings fixed: agents $1, \ldots, N-1$ are assigned sliding-window catalogs of width $0.4K$, while the last agent is assigned a catalog of size $|\mathcal{K}_N| \in \{1, 25, 50, 75, 100\}$.

Figure 3 shows that the regret curves remain nearly overlapping across these catalog sizes of the Nth agent, with similar early-time slope changes and a common late-horizon growth rate. Thus an individual agent's catalog size is not a dominant factor of the joint regret, which corroborates the upper bound of HM-UCB in Theorem 1. This robustness stems directly from the size-weighted pooling mechanism, which dynamically balances information contributions according to local catalog capacities.

7 Conclusion and Future Work

This paper investigates arm-heterogeneous multi-agent linear contextual bandits under bounded misspecification for regret minimization. We formalize this problem by incorporating local misspecifications into a global level ε through catalog-weighted aggregation. Based on this formulation, we propose HM-UCB, which extends SupLinUCB to the multi-agent regime via *size-weighted pooling* of agent observations without requiring prior knowledge of ε. On the theoretical side, we establish a high-probability regret upper bound for HM-UCB that captures both regret efficiency and misspecification robustness, complemented by a generic lower bound characterizing two distinct sources of fundamental hardness. Together, these bounds demonstrate the near-optimality of HM-UCB in key practical regimes. Furthermore, we instantiate our framework under two representative protocols—Naïve and N-Selection—providing tailored algorithmic variants alongside theoretical guarantees. Finally, synthetic experiments empirically corroborate algorithm regret predicted by our theoretical analyses.

More broadly, these results mark an initial step toward understanding model misspecification in multi-agent sequential decision-making. Several promising directions remain open. Currently, our framework addresses misspecification primarily via agent-level quantities that are aggregated through weighted information sharing. A compelling line of future inquiry is to leverage inter-agent communication topologies to model and mitigate misspecification at a global scale, rather than relying on individual agent handling. Such communication-aware paradigms could significantly reduce systematic discrepancies, yielding more efficient algorithms and closing the gap between theoretical regret upper bound and lower bound. More generally, exploring how diverse communication protocols—such as centralized coordination or decentralized graph-structured networks—can robustify multi-agent bandit learning against misspecification offers a rich landscape for future research. We plan to investigate these questions in future research.

AI USE STATEMENT

We used generative AI tools to assist with writing polish and LaTeX formatting of this manuscript. We did not use generative AI tools to generate research ideas, theoretical results, proofs, or experimental findings. All technical claims were produced and verified by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work is a theoretical and empirical study of multi-agent contextual bandits. It does not involve human subjects, personally identifiable data, or the release of sensitive datasets. We do not foresee immediate ethical concerns beyond the standard dual-use considerations of sequential decision-making algorithms.

REPRODUCIBILITY STATEMENT

The problem setting, algorithms, and main theoretical statements are presented in Sections 2–4. Proof is sketched in Appendix C, and detailed in Appendix D. Experimental configurations are summarized in Section 6 and Appendix F. We submit the source code as a `.zip` archive in the supplementary materials.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320, 2011.

Marc Abeille and Alessandro Lazaric. Linear Thompson sampling revisited. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 176–184, 2017.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the International Conference on Machine Learning*, pp. 127–135, 2013.

Ayya Alieva, Ashok Cutkosky, and Abhimanyu Das. Robust pure exploration in linear bandits with limited budget. In *Proceedings of the International Conference on Machine Learning*, pp. 187–195, 2021.

Olivier Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1097–1105, 2014.

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

Anh Do, Thanh Nguyen-Tang, and Raman Arora. Multi-agent learning with heterogeneous linear contextual bandits. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78768–78790, 2023.

Abhimanyu Dubey and Alex Pentland. Differentially-private federated linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6003–6014, 2020a.

Abhimanyu Dubey and Alex Pentland. Kernel methods for cooperative multi-agent contextual bandits. In *Proceedings of the International Conference on Machine Learning*, pp. 2740–2750, 2020b.

Dylan J. Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11478–11489, 2020.

Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Misspecified linear bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 3761–3767, 2017.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10007–10017, 2020.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms. In *IEEE Conference on Decision and Control*, pp. 167–172, 2016.

Tor Lattimore, Csaba Szepesvári, and Gellért Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *Proceedings of the International Conference on Machine Learning*, pp. 5662–5670, 2020.

Hongbo Li and Lingjie Duan. Competitive multi-armed bandit games for resource sharing. *IEEE Transactions on Mobile Computing*, 24(9):8393–8404, 2025.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the International Conference on World Wide Web*, pp. 661–670, 2010.

Lydia T. Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1618–1628, 2020.

Ahmadreza Moradipari, Mohammad Ghavamzadeh, and Mahnoosh Alizadeh. Collaborative multi-agent stochastic linear bandits. In *American Control Conference*, pp. 2761–2766, 2022.

Clémence Réda, Andrea Tirinzoni, and Rémy Degenne. Dealing with misspecification in fixed-confidence linear top-m identification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25489–25501, 2021.

Kei Takemura, Shinji Ito, Daisuke Hatano, Hanna Sumita, Takuro Fukunaga, Naonori Kakimura, and Ken-ichi Kawarabayashi. A parameter-free algorithm for misspecified linear contextual bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 3367–3375, 2021.

Junwen Yang and Vincent Tan. Minimax optimal fixed-budget best arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12253–12266, 2022.

Lin Yang, Yu-Zhen Janice Chen, Mohammad H. Hajiesmaili, John C. S. Lui, and Don Towsley. Distributed bandits with heterogeneous agents. In *IEEE INFOCOM*, pp. 200–209, 2022.

Mengyue Yang, Qingyang Li, Zhiwei Qin, and Jieping Ye. Hierarchical adaptive contextual bandits for resource constraint based recommendation. In *Proceedings of the Web Conference*, pp. 292–302, 2020.

Weitong Zhang, Jiafan He, Zhiyuan Fan, and Quanquan Gu. On the interplay between misspecification and sub-optimality gap in linear contextual bandits. In *Proceedings of the International Conference on Machine Learning*, pp. 41111–41132, 2023.

Zixin Zhong, Wang Chi Cheung, and Vincent Y. F. Tan. Thompson sampling algorithms for cascading bandits. *Journal of Machine Learning Research*, 22(218):1–66, 2021.

Changxi Zhu, Mehdi Dastani, and Shihan Wang. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems*, 38(1):1–48, 2024.

A PAPER ORGANIZATION

The main text is organized as follows. Section 2 formalizes the setting; Section 3 presents HM-UCB; Section 4 states the theoretical findings; Section 5 instantiates two additional protocols; and Section 6 reports numerical results. The appendix is organized as follows. Appendix B records standard facts used in the proofs; Appendix C sketches the arguments of Section 4; Appendix D gives the complete proofs; Appendix E states the pseudocode of NS-UCB; and Appendix F lists experimental configurations.

B USEFUL FACTS

This appendix records standard facts used in Appendix D. Throughout, $V_{t,s} = \lambda I_d + \sum_{(\tau,n)\in\Psi_{t,s}} x_\tau(i_{\tau,n})x_\tau(i_{\tau,n})^\top$ denotes the ridge matrix of stage s, and $w_{t,n,s}(i) = \alpha\|x_t(i)\|_{V_{t,s}^{-1}}$ the associated confidence width.

Lemma 1 (Sub-Gaussian tail)**.** *Let $\{\eta_t\}_{t\geq 1}$ be adapted to a filtration $\{\mathcal{F}_t\}_{t\geq 1}$ and conditionally R-sub-Gaussian. Then for every $a > 0$ and $t \geq 1$,*

$$\mathbb{P}\left(\sum_{s=1}^{t}\eta_s > a\right) \leq \exp\left(-\frac{a^2}{2tR^2}\right).$$

Lemma 2 (Elliptical potential / stage-wise sample size)**.** *Suppose $\|x_t(i)\|_2 \leq L$ and every recorded pair in $\Psi_{t,s}$ satisfies $w_{t,n,s}(i_{t,n}) > \alpha c^{-s}$. Then*

$$|\Psi_{t,s}| \leq 2c^{2s}\, d\log\Big(1 + \frac{L^2|\Psi_{t,s}|}{d\lambda}\Big).$$

In particular, $|\Psi_{T,1}| = \tilde{O}(d)$.

Proof sketch. Each exploration sample increases $\log\det V_{t,s}$ by at least $\Omega(c^{-2s})$, while $\log\det V_{t,s} \leq d\log(1 + L^2|\Psi_{t,s}|/(d\lambda))$. Summing the increments yields the claimed bound; see Chu et al. (Chu et al., 2011) and Takemura et al. (Takemura et al., 2021). □

Lemma 3 (Ridge remainder)**.** *For any $x \in \mathbb{R}^d$, $\left|x^\top V_{t,s}^{-1}\theta\right| \leq M\|x\|_{V_{t,s}^{-1}}/\sqrt{\lambda}$. Consequently the ridge bias is absorbed into $\beta(\delta) = R\sqrt{2\log(2KSNT/\delta)} + \sqrt{\lambda}\,M$.*

Lemma 4 (Bretagnolle–Huber)**.** *For any pair of probability measures P, Q and any event A,*

$$P(A) + Q(A^c) \geq \tfrac{1}{2}\exp\big(-\mathrm{KL}(P\|Q)\big).$$

Lemma 5 (Bernoulli KL)**.** *For $0 < \varepsilon \leq 1/4$, $\mathrm{KL}\big(\mathrm{Bern}(1/2)\,\|\,\mathrm{Bern}(1/2+\varepsilon)\big) \leq 2\varepsilon^2$.*

Condition 1 is equivalent to the Johnson–Lindenstrauss threshold

$$d \geq \left\lceil \frac{8\log K}{\varepsilon^2} \right\rceil, \tag{2}$$

used for the near-orthogonal packing of Lemma 11 in the Yao-type term of Theorem 2. Condition 2 is the positivity requirement

$$\frac{N}{K} + \varepsilon\sqrt{\frac{NT}{2K}} < 1 \tag{3}$$

for the Divergence-type term of Theorem 2.

C PROOF SKETCHES

We sketch the proofs of the results in Section 4; complete arguments appear in Appendix D.

C.1 Proof sketch of Theorem 1

The analysis of HM-UCB follows the stage-wise regret decomposition of SupLinUCB (Takemura et al., 2021), with additional care for the weighted update and misspecification aggregation in the heterogeneous setting.

The cumulative regret decomposes as

$$R(T) = \underbrace{\sum_{t\in\Psi_{T+1,1}} (\mu_t(i_t^*) - \mu_t(i_t))}_{R^{\text{first}}(T)} + \underbrace{\sum_{s=2}^{S} \sum_{t\in\Psi_{T+1,s}} (\mu_t(i_t^*) - \mu_t(i_t))}_{R^{\text{main}}(T)} + \underbrace{\sum_{t\in\Psi_0} (\mu_t(i_t^*) - \mu_t(i_t))}_{R^{\text{confidence}}(T)}.$$

The Confidence-branch contribution is $R^{\text{confidence}}$; R^{first} and R^{main} are incurred by the Exploration branch, with R^{first} corresponding to stage $s = 1$ and R^{main} collecting the remaining stages. We bound each term in turn.

We bound $R^{\text{first}}(T)$ by applying Lemma 1 of Takemura et al. (2021):

$$|\Psi_{t,s}| \le 2c^{2s} d \log\Big(1 + \frac{L^2 |\Psi_{t,s}|}{d\lambda}\Big).$$

$$R^{\text{first}}(T) \le B|\Psi_{T,1}| \le B2c^2 d \log\Big(1 + \frac{L^2|NT|}{d\lambda}\Big) = O(Bd\log(NT/d)) = \tilde{O}(Bd). \tag{4}$$

To bound $R^{\text{main}}(T)$, we decompose the estimation error of $\hat{\theta}_{t,s}$ into a stochastic noise term and a misspecification bias term. We then invoke a concentration inequality for the former and a Cauchy–Schwarz–based bound for the latter. Substituting both into the stage-wise elimination inequality yields the stated regret. The argument differs from the single-agent template through the heterogeneous misspecification (Definition 1) and the weighted update. Recall the weighted ridge estimator update in Algorithm 1:

$$\hat{\theta}_{t,s} = V_{t,s}^{-1} \sum_{(\tau,n)\in\Psi_{t,s}} \frac{|K_n|}{|K|}\, r_{\tau,n}\, x_\tau(i_{\tau,n}),$$

where the observed reward admits the decomposition $r_{\tau,n} = \theta^\top x_\tau(i_{\tau,n}) + \varepsilon_{\tau,n}(i_{\tau,n}) + \eta_{\tau,n}$. Here $\eta_{\tau,n}$ is the noise, and $\varepsilon_{\tau,n}(i_{\tau,n})$ captures misspecification with $|\varepsilon_{\tau,n}(i_{\tau,n})| \le \varepsilon_n$ and $\sum_n (|K_n|/|K|)\varepsilon_n \le \varepsilon$ (Definition 1). Hence, for any arm i and any stage (t,s), the estimation error decomposes as

$$\begin{aligned} \big|(\hat{\theta}_{t,s} - \theta)^\top x_t(i)\big| \le & \underbrace{\Big| x_t(i)^\top V_{t-1,s}^{-1} \sum_{(\tau,n)\in\Psi_{t,s}} \frac{|K_n|}{|K|} \eta_{\tau,n} x_\tau(i_{\tau,n}) \Big|}_{\text{(I) stochastic noise}} \\ & + \underbrace{\Big| x_t(i)^\top V_{t-1,s}^{-1} \sum_{(\tau,n)\in\Psi_{t,s}} \frac{|K_n|}{|K|} \varepsilon_{\tau,n}(i_{\tau,n}) x_\tau(i_{\tau,n}) \Big|}_{\text{(II) residual bias}}. \end{aligned} \tag{5}$$

Bounding (I). The concentration lemma gives

$$\Big| x_t(i)^\top V_{t-1,s}^{-1} \sum_{(\tau,n)\in\Psi_{t,s}} \frac{|K_n|}{|K|} \eta_{\tau,n} x_\tau(i_{\tau,n}) \Big| \le \beta(\delta)\, \|x_t(i)\|_{V_{t,s}^{-1}}. \tag{6}$$

Bounding (II). Using $|\varepsilon_{\tau,n}(i_{\tau,n})| \le \varepsilon_n$ and Cauchy–Schwarz,

$$\begin{aligned} \text{(II)} &\le \Big(\sum_n \tfrac{|K_n|}{|K|}\varepsilon_n\Big) \cdot \sum_{(\tau,n)\in\Psi_{t,s}} \big| x_t(i)^\top V_{t-1,s}^{-1} x_\tau(i_{\tau,n}) \big| \\ &\le \sqrt{2}\varepsilon \cdot \sqrt{d \log\big(1 + \tfrac{L^2 N t}{d\lambda}\big)}. \end{aligned} \tag{7}$$

After these steps, the weighted scheme collapses to the global ε-budget. By the definition of $\Psi_{t,s}$ and the termination stage s_t, together with equation 6, equation 7 and the confidence-width threshold in the Stage branch (line 21), we obtain, for any stage s_t, the bound:

$$\mu_t(i_t^*) - \mu_t(i) \le 5\beta(\delta)c^{1-s_t} + 2\varepsilon\left(1 + \sqrt{2d\log\left(1 + \frac{L^2 N t}{d\lambda}\right)}\right) s_t. \tag{8}$$

By aggregating stage-wide results, we finally obtain that, with probability $1 - \frac{\delta(S-1)}{S}$,

$$R^{\text{main}}(T) \le C_1\sqrt{dNT\log(KN)} + C_2\,\varepsilon\sqrt{d}\,NT = \tilde{O}(R\sqrt{dNT\log(KN)} + \varepsilon\sqrt{d}NT), \tag{9}$$

where the explicit constants are $C_1 = 10Rc\sqrt{\log(ST)/\delta}\sqrt{\log(1 + L^2NT/(d\lambda))}$ and $C_2 = 2(1 + 2\sqrt{\log(1 + L^2NT/(d\lambda))})$.

Given the confidence widths (line 16 in Algorithm 1), we can bound $R^{\text{confidence}}(T)$ by a similar argument to that used for $R^{\text{main}}(T)$:

$$R^{\text{confidence}}(T) = \tilde{O}(R\sqrt{dNT} + \varepsilon\sqrt{d}NT) \tag{10}$$

with probability $1 - \frac{\delta}{S}$.

Summing the three parts proves Theorem 1.

C.2 Proof sketch of Theorem 2

The two terms of Theorem 2 rely on a finite-dimensional geometric packing argument (Yao's principle) and an information-theoretic change-of-measure argument, following Lattimore et al. (2020) and Zhong et al. (2021) respectively.

Before proving the bounds, we introduce the number of distinct optimal arms in the heterogeneous setting. Let $i_{t,n}^*$ denote an optimal arm for agent n at time t, and let K_t' count the number of distinct arms in $\{i_{t,n}^*\}_{n\in[N]}$:

$$K_t' = \left|\bigcup_{n\in[N]} \{i_n^\star :\ i_n^\star \in \arg\max_{i\in K_n} \mu_n(i)\}\right|.$$

In the heterogeneous setting, the collective objective in every round is to minimize regret while selecting K_t' distinct arms. The value of K_t' connects the heterogeneous setting to the two protocols in Section 5: (1) $K_t' = 1$ (Naïve); (2) $1 < K_t' < N$ (heterogeneous); (3) $K_t' = N$ (N-Selection). N-Selection is thus a special case of the heterogeneous setting and appears naturally in the heterogeneous lower-bound proofs.

Yao-type term of Theorem 2. Under Condition 1, we can construct instances with K unit vectors $\{a_1, \dots, a_K\} \subset \mathbb{R}^d$ satisfying

$$|\langle a_i, a_i\rangle| = 1, \qquad |\langle a_i, a_j\rangle| \le \sqrt{\frac{8\log K}{d-1}} \quad (i \ne j).$$

Then a hidden anchor $a^\star$ is chosen as the context of a good arm, and we construct the packing family so that exactly N good arms share this elevated reward, while the remaining (bad) arms are near-orthogonal alternatives. In detail, we set

$$\theta = \delta a^\star, \qquad \delta = \varepsilon\sqrt{\frac{d-1}{8\log K}},$$

and set linear rewards so that good arms have gap δ in expected reward over bad arms, while still satisfying Definition 1. Then each pull of a bad arm reveals only weak information about which N arms are good. Let τ be the first round by which all hidden good arms are identified. It follows that

$$\mathbb{E}[R(T)] \ge \delta\,\mathbb{E}[\tau].$$

To obtain the covering-time bound, consider an algorithm that pulls arms from the catalog sequentially and stops upon collecting N arms with reward δ. Let Y be the index of the last high-reward

arm encountered in a uniformly randomized embedding. Then $\tau = Y - N$, and the packing lower bound follows from computing $\mathbb{E}[Y]$:

$$P(Y = e) = \frac{\binom{e-1}{N-1}}{\binom{K}{N}}, \qquad e = N, \dots, K,$$

$$\mathbb{E}[Y] = \sum_{e=N}^{K} e\, P(Y = e) = \frac{N(K+1)}{N+1}, \mathbb{E}[\tau] = \mathbb{E}[Y] - N = \frac{N(K-N)}{N+1},$$

hence

$$\mathbb{E}[\tau] \geq \min\left\{ NT, \frac{N(K-N)}{N+1} \right\}.$$

Combining it with $\mathbb{E}[R(T)] \geq \delta\, \mathbb{E}[\tau]$ and the value of δ yields the Yao-type term of Theorem 2.

Divergence-type term of Theorem 2. We use the cyclic family introduced in Section 4 and compare each alternative instance k with the null instance 0. Under instance k, the good set is $S^{*,k}$, so the reward of round t for instance k is

$$r(S_t \mid r^{(k)}) = \sum_{i \in S_t} r^{(k)}(i).$$

The per-round regret is

$$r(S_k^* \mid r^{(k)}) - r(S_t \mid r^{(k)}) = \varepsilon |S_k^* - S_t|.$$

By Pinsker's inequality, the difference between the expected number of times that an arm is selected under instance k and under instance 0 can be bounded in terms of the KL divergence between the corresponding observation distributions. In particular, for each alternative instance $k \in [K]$, we have

$$\mathbb{E}^{(k)}\left[\sum_{t=1}^{T} \left(r(S^{*,k} \mid r^{(k)}) - r(S_t^\pi \mid r^{(k)}) \right) \right] = \varepsilon \sum_{j \in [K] \setminus S^{*,k}} \mathbb{E}^{(k)}\left[\sum_{t=1}^{T} \mathbb{I}(j \in S_t^\pi) \right]. \tag{11}$$

We next relate the selection probability under instance k to that under instance 0. By Pinsker's inequality, we obtain

$$\left| \mathbb{P}^{(k)}(j \in S_t^\pi) - \mathbb{P}^{(0)}(j \in S_t^\pi) \right| \leq \sqrt{\frac{1}{2} \mathrm{KL}\left(P_{t,j}^{(0)} \,\middle\|\, P_{t,j}^{(k)} \right)}.$$

Summing this inequality over t and j and then averaging over the K alternative instances yield the following

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}^{(k)}[R(T)] \geq \varepsilon NT \left(1 - \frac{N}{K} - \sqrt{\frac{T}{2} \frac{N^2}{K} \overline{\mathrm{KL}}(P^0, P^k)} \right), \tag{12}$$

where $\overline{\mathrm{KL}}(P^0, P^k) := \frac{1}{K} \sum_{k=1}^{K} \mathrm{KL}(P^0 \,\|\, P^k)$ denotes the average KL divergence between instance P^0 and the alternative instances $\{P^k\}_{k=1}^K$.

It remains to bound the KL divergence appearing in equation 12. Since the observations are Bernoulli distributed, we have

$$\mathrm{KL}\left(\mathrm{Bern}\left(\frac{1}{2}\right) \middle\| \mathrm{Bern}\left(\frac{1}{2} + \varepsilon\right) \right) = \frac{1}{2} \log \frac{1/2}{1/2 + \varepsilon} + \frac{1}{2} \log \frac{1/2}{1/2 - \varepsilon} = O(\varepsilon^2),$$

for sufficiently small ε. Consequently, the averaged KL divergence satisfies

$$\overline{\mathrm{KL}}(P^0, P^k) = O(\varepsilon^2),$$

and substituting this bound into equation 12 contributes to

$$R(T) \geq \varepsilon NT \left(1 - \frac{N}{K} - \varepsilon \sqrt{\frac{NT}{2K}} \right). \tag{13}$$

Condition 2 ensures the right-hand side is positive, yielding the Divergence-type term of Theorem 2.

C.3 Proof sketch of Corollary 1

Under the constraint $\varepsilon^2 \le \frac{4\alpha^2 d}{NT} \le M^2 L^2$, and Condition 1, Corollary 1 considers the following instance, which is obtained from the hard instance used in the proof of the Yao-type term of Theorem 2 by redefining the reward function. Specifically, we set

$$\delta = \frac{\varepsilon}{2}\sqrt{\frac{d-1}{8\log K}}, \qquad \gamma = \frac{\alpha\sqrt{d}}{\sqrt{NT}},$$

and define

$$\theta = (\delta + \gamma)a^\star.$$

The above constraint ensures that the resulting expected rewards remain within the prescribed bounded range. Following the same argument as in the proof of the Yao-type term of Theorem 2, on this instance we have the regret satisfying $\mathbb{E}[R(T)] \ge (\delta+\gamma)\,\mathbb{E}[\tau]$, yielding

$$\mathbb{E}[R(T)] = \tilde{\Omega}\left(\left[\frac{R\sqrt{d\log(KN)}}{\sqrt{NT}} + \frac{\varepsilon}{2}\sqrt{\frac{d-1}{8\log K}}\right] m\right),$$

where

$$m = \min\left\{NT,\ \frac{N(K-N)}{N+1}\right\}.$$

On the other hand, the upper bound directly follows from Theorem 1, which provides a matching-order upper control for HM-UCB. Hence, the upper and lower bounds match up to logarithmic factors, establishing Corollary 1.

D Detailed proofs

D.1 Proof of Theorem 1: Upper Bound

This section gives the complete proof of Theorem 1, including the explicit constants $C_1, C_2, \ldots$.

We prove the high-probability bound $R(T) = \tilde{O}\Big(R\sqrt{dNT\log(KN)} + \varepsilon\sqrt{d}\,NT + Bd\Big)$ under the parameter choices $\alpha = \beta(\delta)$, $c - 1 = \Theta(1)$, and $\lambda = R^2M^{-2}$ stated in Theorem 1. Following SupLinUCB (Takemura et al., 2021), we decompose regret using exploration records $\Psi_{t,s}$ (Algorithm 1) and per-agent termination stages.

$$R(T) = R^{\text{first}}(T) + R^{\text{main}}(T) + R^{\text{confident}}(T),$$

where

$$R^{\text{first}}(T) = \sum_{n\in[N]} \sum_{t\in\Psi_{T,1,n}} (\mu_{t,n}(i^*_{t,n}) - \mu_{t,n}(i_{t,n})),$$

$$R^{\text{main}}(T) = \sum_{n\in[N]} \sum_{s=2}^{S} \sum_{t\in\Psi_{T,s,n}} (\mu_{t,n}(i^*_{t,n}) - \mu_{t,n}(i_{t,n})), \quad \text{and}$$

$$R^{\text{confident}}(T) = \sum_{n\in[N]} \sum_{t\in\Psi_{0,n}} (\mu_{t,n}(i^*_{t,n}) - \mu_{t,n}(i_{t,n})).$$

To emphasize, in the algorithm we do not separate exploration sets $\Psi_{t,s}$ of different agents in the same stage and round, that is

$$\Psi_{t,s} = \bigcup_{n\in[N]} \Psi_{t,s,n}$$

where we can conclude $|\Psi_{t,s}| = |\bigcup_{n\in[N]} \Psi_{t,s,n}| = O(NT)$. In the following proof, we use $|\Psi_{t,s}|$ for clarity, and use $\{t, n\}$ to denote the record in the exploration set from agent n in round t. This notation is consistent with the algorithm description.

The following lemma controls stage-1 termination; it corresponds to equation equation 4 in Section C and is the $s = 1$ case of Lemma 2.

Lemma 6. *For all $t \in [T]$ and $s \geq 1$,*

$$|\Psi_{t,s}| \leq 2c^{2s} d \log\big(1 + L^2|\Psi_{t,s}|/(d\lambda)\big)\,.$$

In particular, $|\Psi_{T,1}| = \tilde{O}(d)$.

Proof. The claim is the general-stage form of Lemma 2: every recorded pair in $\Psi_{t,s}$ is an Exploration-branch pull and therefore satisfies $w_{t,n,s}(i_{t,n}) > \alpha c^{-s}$. □

With Lemma 6, we bound the first part of regret. In Section C, $c_1 = 2c^2 \log\big(1 + L^2 NT/(d\lambda)\big)$.

$$\begin{aligned} R^{\text{first}}(T) &\leq B|\Psi_{T,1}| \\ &\leq 2c^2 Bd \log\big(1 + L^2|\Psi_{T,1}|/(d\lambda)\big) \\ &= O\big(Bd \log\big(1 + L^2 NT/(d\lambda)\big)\big) \\ &= \tilde{O}\,(Bd) \end{aligned}$$

We first bound parameter estimation error (stochastic noise, Lemma 7; residual bias, Lemma 8), then apply the per-round bound (Lemma 9) as in equation equation 9 of Section C.

Lemma 7 (Noise concentration)**.** *Let $w_n := |K_n|/|K| \in (0, 1]$ and let $V_{t,s}$ be as in Algorithm 1, i.e. $V_{t,s} = \lambda I_d + \sum_{(\tau,n)\in\Psi_{t,s}} x_\tau(i_{\tau,n}) x_\tau(i_{\tau,n})^\top$. For every $t \in [T]$, $s \in [S]$ and arm i, with probability at least $1 - \delta/(2KSNT)$,*

$$\left| x_t(i)^\top V_{t,s}^{-1} \sum_{(\tau,n)\in\Psi_{t,s}} w_n \eta_{\tau,n} x_\tau(i_{\tau,n}) \right| + \left|\lambda x_t(i)^\top V_{t,s}^{-1}\theta\right| \leq \beta(\delta)\, \|x_t(i)\|_{V_{t,s}^{-1}},$$

where $\beta(\delta) = R\sqrt{2\log(4KSNT/\delta)} + \sqrt{\lambda}\, M$.

Proof. Write $z := x_t(i)$ and $\Psi := \Psi_{t,s}$. The ridge identity $V_{t,s} = \lambda I_d + \sum_{(\tau,n)\in\Psi} x_\tau(i_{\tau,n}) x_\tau(i_{\tau,n})^\top$ gives

$$\sum_{(\tau,n)\in\Psi} x_\tau(i_{\tau,n}) x_\tau(i_{\tau,n})^\top - V_{t,s} = -\lambda I_d,$$

hence $\left|\lambda z^\top V_{t,s}^{-1}\theta\right| \leq \sqrt{\lambda}\, M \|z\|_{V_{t,s}^{-1}}$ by Lemma 3.

For the martingale term, $w_n\eta_{\tau,n}$ is conditionally R-sub-Gaussian because $w_n \leq 1$. Lemma 1 therefore yields, for any $a > 0$,

$$\mathbb{P}\left(\Big| \sum_{(\tau,n)\in\Psi} w_n \big(z^\top V_{t,s}^{-1} x_\tau(i_{\tau,n})\big)\eta_{\tau,n} \Big| > a \right) \leq 2\exp\left(-\frac{a^2}{2R^2 \sum_{(\tau,n)\in\Psi} w_n^2 \big(z^\top V_{t,s}^{-1} x_\tau(i_{\tau,n})\big)^2} \right).$$

Since $w_n^2 \leq w_n \leq 1$,

$$\begin{aligned} \sum_{(\tau,n)\in\Psi} w_n^2 \big(z^\top V_{t,s}^{-1} x_\tau(i_{\tau,n})\big)^2 &= z^\top V_{t,s}^{-1} \Big(\sum_{(\tau,n)\in\Psi} w_n^2\, x_\tau(i_{\tau,n}) x_\tau(i_{\tau,n})^\top \Big) V_{t,s}^{-1} z \\ &\leq z^\top V_{t,s}^{-1} \Big(\sum_{(\tau,n)\in\Psi} x_\tau(i_{\tau,n}) x_\tau(i_{\tau,n})^\top \Big) V_{t,s}^{-1} z \\ &= z^\top V_{t,s}^{-1} (V_{t,s} - \lambda I_d) V_{t,s}^{-1} z \leq \|z\|^2_{V_{t,s}^{-1}}. \end{aligned}$$

Taking $a = \beta_{\mathrm{n}}\|z\|_{V_{t,s}^{-1}}$ with $\beta_{\mathrm{n}} = R\sqrt{2\log(4KSNT/\delta)}$ makes the right-hand side at most $\delta/(2KSNT)$. Adding the ridge remainder completes the proof. □

Lemma 8 (Parameter error with misspecification)**.** *On the event of Lemma 7, for all $t \in [T]$, all stages s with $\|x_t(i)\|_{V_{t,s}^{-1}} \le c^{-s}$, and all surviving arms $i \in I_{t,s}$,*

$$\left|(\hat{\theta}_{t,s} - \theta)^\top x_t(i)\right| \le \beta(\delta)\,\|x_t(i)\|_{V_{t,s}^{-1}} + \varepsilon\sqrt{2d\log\bigl(1 + L^2NT/(d\lambda)\bigr)}.$$

In particular, a union bound over at most $KSNT$ triples (t, s, i) yields that the bound holds simultaneously with probability at least $1 - \delta/2$.

Proof. Fix t, s and $i \in I_{t,s}$. As in equation 5,

$$\left|(\hat{\theta}_{t,s} - \theta)^\top x_t(i)\right| \le \underbrace{\left|x_t(i)^\top V_{t,s}^{-1} \sum_{(\tau,n)\in\Psi_{t,s}} w_n \eta_{\tau,n} x_\tau(i_{\tau,n})\right| + \left|\lambda x_t(i)^\top V_{t,s}^{-1}\theta\right|}_{\text{(I)}}$$
$$+ \underbrace{\left|x_t(i)^\top V_{t,s}^{-1} \sum_{(\tau,n)\in\Psi_{t,s}} w_n \varepsilon_{\tau,n}(i_{\tau,n}) x_\tau(i_{\tau,n})\right|}_{\text{(II)}}.$$

Term (I) is Lemma 7. For (II), Definition 1 and $w_n\varepsilon_n$ summing to at most ε give

$$\begin{aligned}
\text{(II)} &\le \varepsilon \sum_{(\tau,n)\in\Psi_{t,s}} \left|x_\tau(i_{\tau,n})^\top V_{t,s}^{-1} x_t(i)\right| \\
&\le \varepsilon\sqrt{|\Psi_{t,s}|}\sqrt{\sum_{(\tau,n)\in\Psi_{t,s}} \left(x_\tau(i_{\tau,n})^\top V_{t,s}^{-1} x_t(i)\right)^2} \\
&= \varepsilon\sqrt{|\Psi_{t,s}|}\,\left\|x_t(i)\right\|_{V_{t,s}^{-1}(V_{t,s}-\lambda I_d)V_{t,s}^{-1}} \\
&\le \varepsilon\sqrt{|\Psi_{t,s}|}\,\|x_t(i)\|_{V_{t,s}^{-1}}.
\end{aligned}$$

The stage constraint $\|x_t(i)\|_{V_{t,s}^{-1}} \le c^{-s}$ and the general-s sample-size bound of Lemma 2 (equivalently Lemma 6), $|\Psi_{t,s}| \le 2c^{2s}d\log(1 + L^2|\Psi_{t,s}|/(d\lambda)) \le 2c^{2s}d\log(1 + L^2NT/(d\lambda))$, imply

$$\sqrt{|\Psi_{t,s}|}\,c^{-s} \le \sqrt{2d\log\bigl(1 + L^2NT/(d\lambda)\bigr)}.$$

This is the s-dependent form of the stage-1 count: the extra c^s in $|\Psi_{t,s}|$ cancels the extra c^{-s} in the width, so the residual-bias contribution remains $\varepsilon\sqrt{2d\log(\cdots)}$ at every stage. □

Lemma 9 (Per-round regret in main stages)**.** *For all $t \in [T]$ with $s_t \ge 2$, on the event of Lemma 8,*

$$\mu_t(i_t^*) - \mu_t(i) \le 5\beta(\delta)\,c^{1-s_t} + 2\varepsilon\bigl(1 + \sqrt{2d\log(1 + L^2Nt/(d\lambda))}\bigr)s_t$$

for every arm i still in the candidate set at stage s_t.

Proof. Write $\Delta_{\mathrm{mis}}(t) := \varepsilon\bigl(1 + \sqrt{2d\log(1 + L^2Nt/(d\lambda))}\bigr)$. Lemma 8 and $|\varepsilon_{t,n}(\cdot)| \le \varepsilon$ yield, for every surviving arm j at stage s,

$$\left|\mu_t(j) - \hat{r}_{t,s}(j)\right| \le w_{t,s}(j) + \Delta_{\mathrm{mis}}(t),$$

where $w_{t,s}(j) = \beta(\delta)\|x_t(j)\|_{V_{t,s}^{-1}}$ and $\hat{r}_{t,s}(j) = x_t(j)^\top\hat{\theta}_{t,s}$. Thus

$$\mu_t(i_t^*) - \mu_t(i) \le \bigl(\hat{r}_{t,s}(i_t^*) + w_{t,s}(i_t^*)\bigr) - \bigl(\hat{r}_{t,s}(i) - w_{t,s}(i)\bigr) + 2\Delta_{\mathrm{mis}}(t).$$

The optimal arm remains in the candidate set throughout stages $1, \dots, s_t - 1$: at a stage-s elimination, any arm with UCB below $\max \mathrm{UCB} - 2\alpha c^{1-s}$ is discarded, and the displayed inequality together with $w_{t,s}(i_t^*) \le \alpha c^{-s}$ (the stage-branch trigger) prevents i_t^* from being eliminated.

After $s_t - 1$ eliminations the candidate set satisfies $w_{t,s}(j) \le \alpha c^{1-s_t}$ for all remaining j, and the last elimination constraint gives

$$\hat{r}_{t,s}(i_t^*) + w_{t,s}(i_t^*) \le \hat{r}_{t,s}(i) + w_{t,s}(i) + 2\alpha c^{1-s_t}.$$

Substituting and using $w_{t,s}(\cdot) \le \alpha c^{1-s_t}$ produces a gap of at most $5\alpha c^{1-s_t} + 2\Delta_{\mathrm{mis}}(t)$. Each of the s_t stages can contribute an additional misspecification slack $\Delta_{\mathrm{mis}}(t)$ through the UCB comparisons, which accounts for the factor s_t on the ε term. Since $\alpha = \beta(\delta)$, the claim follows. □

We can now bound $R^{\text{main}}(T)$. With Lemma 9, we have

$$\begin{aligned}
&\sum_{s=2}^{S} \sum_{t \in \Psi_{T+1,s}} \left(\mu_t\left(i_t^*\right) - \mu_t\left(i_t\right)\right) \\
&\quad \le \sum_{s=2}^{S} 5\beta(\delta) c^{1-s} \left|\Psi_{T+1,s}\right| \\
&\qquad + \sum_{s=2}^{S} 2\varepsilon \left(1 + 2\sqrt{d \log\left(1 + L^2 NT/(d\lambda)\right)}\right) s \left|\Psi_{T+1,s}\right|.
\end{aligned}$$

We bound the first term by Lemma 6, that is

$$\begin{aligned}
&\sum_{s=2}^{S} 5\beta(\delta) c^{1-s} \left|\Psi_{T+1,s}\right| \\
&\quad \le 5\beta(\delta) c \sqrt{2d \log\left(1 + L^2 NT/(d\lambda)\right)} \sum_{s=2}^{S} \sqrt{\left|\Psi_{T+1,s}\right|}.
\end{aligned}$$

Recall that $|\Psi_{t,s}| = O(NT)$. We then have

$$\sum_{s=2}^{S} \sqrt{\left|\Psi_{T+1,s}\right|} \le \sqrt{S \sum_{s=2}^{S} \left|\Psi_{T+1,s}\right|} \le \sqrt{SNT}.$$

Recall that $S = \lceil \log_c(T/d)/2 \rceil$ and $\beta(\delta) = \tilde{O}(R\sqrt{\log(KN)})$. From the assumption $c - 1 = \Theta(1)$, there exists a universal constant $C > 0$ such that $\frac{1}{2}\log_c(T/d) \le C \log(T/d)$. This implies that $S = O(\log(T))$. Thus, we have

$$\sum_{s=2}^{S} 5\beta(\delta) c^{1-s} \left|\Psi_{T+1,s}\right| = \tilde{O}(R\sqrt{dNT \log(KN)}).$$

Furthermore, from the fact that $\sum_{s \in [S]} |\Psi_{T+1,s}| \le NT$, we have that the second term is $\tilde{O}(\varepsilon\sqrt{d}NT)$. Hence, we obtain

$$R^{\text{main}}(T) = \tilde{O}\left(R\sqrt{dNT \log(KN)} + \varepsilon\sqrt{d}NT\right)$$

on the event of Lemma 8 (probability at least $1 - \delta/2$).

The confident branch uses the threshold $\alpha\sqrt{d/(NT)}$ (Algorithm 1, line 15), replacing the single-agent threshold $\alpha\sqrt{d/T}$; see also equation equation 10 in Section C.

Lemma 10 (Confident-branch instantaneous regret)**.** *We have*

$$\mu_t\left(i_{t,s_t}^*\right) - \mu_t\left(i_t\right) \le 2\beta(\delta)\sqrt{d/NT} + 2\varepsilon(1 + \sqrt{d})$$

for all $t \in \Psi_0$, on the event of Lemma 8.

We now bound $R^{\text{confident}}(T)$. Write

$$R^{\text{confident}}(T) = \sum_{t\in\Psi_0} \left(\mu_t\left(i_t^*\right) - \mu_t\left(i_{t,s_t}^*\right)\right) + \sum_{t\in\Psi_0} \left(\mu_t\left(i_{t,s_t}^*\right) - \mu_t\left(i_t\right)\right).$$

On the confident branch the optimal arm remains in the candidate set, so $i_t^* = i_{t,s_t}^*$ and the first sum vanishes. The second sum is controlled by Lemma 10. Since $|\Psi_0| \le NT$, we conclude

$$R^{\text{confident}}(T) = \tilde{O}\left(R\sqrt{dNT} + \varepsilon\sqrt{d}NT\right)$$

on the same event.

Proof of Lemma 10. Fix a round $t \in \Psi_0$ and let $i_{t,n}$ be the arm selected by agent n (or by the global selection rule in the homogeneous special cases). On the confident branch, for the active stage s_t and the active candidate set $I_{t,n}$ we have $w_{t,n,s_t}(i) \le \alpha\sqrt{d/(NT)}$ for all $i \in I_{t,n}$, hence in particular for $i = i_{t,n}$ and $i = i_{t,n}^*$. By the UCB selection rule,

$$\hat{r}_{t,n,s_t}(i_{t,n}^*) + w_{t,n,s_t}(i_{t,n}^*) \le \hat{r}_{t,n,s_t}(i_{t,n}) + w_{t,n,s_t}(i_{t,n}).$$

Rearranging and using $w_{t,n,s_t}(\cdot) \le \alpha\sqrt{d/(NT)}$ gives

$$\hat{r}_{t,n,s_t}(i_{t,n}^*) - \hat{r}_{t,n,s_t}(i_{t,n}) \le 2\alpha\sqrt{\frac{d}{NT}}.$$

Now relate prediction error to true mean differences using Lemma 8 and Definition 1. For any arm i,

$$\mu_{t,n}(i) = \theta^\top x_t(i) + \varepsilon_{t,n}(i), \qquad |\varepsilon_{t,n}(i)| \le \varepsilon_n, \qquad \sum_n \frac{|K_n|}{|K|}\varepsilon_n \le \varepsilon.$$

Thus,

$$\mu_{t,n}(i_{t,n}^*) - \mu_{t,n}(i_{t,n}) = (\theta - \hat{\theta}_{t,s_t})^\top x_t(i_{t,n}^*) - (\theta - \hat{\theta}_{t,s_t})^\top x_t(i_{t,n}) + \hat{r}_{t,n,s_t}(i_{t,n}^*) - \hat{r}_{t,n,s_t}(i_{t,n}) \;+\; \varepsilon_{t,n}(i_{t,n}^*) - \varepsilon_{t,n}(i_{t,n}).$$

Applying Lemma 8 and $|\varepsilon_{t,n}(\cdot)| \le \varepsilon_n \le \varepsilon$ (Definition 1) yields the stated bound. □

A union bound over the at most $KSNT$ applications of Lemma 7 costs $\delta/2$. The residual-bias contribution in Lemma 8 is deterministic given Definition 1, and Lemmas 9–10 are consequences of that same event. Hence, with probability at least $1-\delta$,

$$R(T) = R^{\text{first}}(T) + R^{\text{main}}(T) + R^{\text{confident}}(T) = \tilde{O}\Big(R\sqrt{dNT\log(KN)} + \varepsilon\sqrt{d}\,NT + Bd\Big),$$

matching Theorem 1.

D.2 Proof of the Yao-type term of Theorem 2

This section proves the Yao-type term of Theorem 2, following the near-orthogonal packing construction of Lattimore et al. (2020, Lemma 3.1 and Corollary 3.3) and the N-Selection counting argument in Section C. The hard features match the packing family of Section 4 and the lower-bound experiments (Section 6): pairwise inner products at most $\sqrt{8\log K/(d-1)}$ and gap $\delta = \varepsilon\sqrt{(d-1)/(8\log K)}$.

The dimension condition equation 2 is the JL threshold in Lattimore et al. (2020, Lemma 3.1).

Lemma 11 (Near-orthogonal unit features)**.** *If $d \ge \lceil 8\log K/\varepsilon^2\rceil$, then there exist unit vectors $a_1,\dots,a_K \in \mathbb{R}^d$ such that $\|a_i\|_2 = 1$ for all i and*

$$|\langle a_i, a_j\rangle| \le \rho := \sqrt{\frac{8\log K}{d-1}} \qquad (i \ne j).$$

Proof. Sample $Z_1, \dots, Z_K$ i.i.d. uniform on the unit sphere in $\mathbb{R}^d$. For fixed $i \neq j$, a standard concentration bound (Lattimore et al., 2020, proof of Lemma 3.1) gives $\mathbb{P}(|\langle Z_i, Z_j\rangle| > \rho) \leq 2\exp(-(d-1)\rho^2/8) = 2/K^2$. Union bound over $\binom{K}{2}$ pairs yields failure probability at most $1/K$, so such $\{a_i\}$ exist. □

Fix $\{a_i\}$ and set $\delta := \varepsilon\sqrt{(d-1)/(8\log K)} = \varepsilon\rho^{-1}$, so $\delta\rho = \varepsilon$.

We use the N-Selection protocol (Section 5.2) with $K'_t = N$: each round selects $i_{t,n} \in K_n$ and $R(T) = \sum_{t,n}(\mu_{t,n}(i^\star_{t,n}) - \mu_{t,n}(i_{t,n}))$. For the lower bound, take $K_n = [K]$ and static features $x_t(i) = a_i$.

Draw $O \subset [K]$ uniformly among $\binom{K}{N}$ subsets of size N. Pick $a^\star \in \{a_i : i \in O\}$ and set $\theta = \delta a^\star$,

$$\mu_{t,n}(i) = \theta^\top a_i + \Delta_{t,n}(i),$$

with $|\Delta_{t,n}(i)| \leq \varepsilon$ (Definition 1) and

$$\mu_{t,n}(i) = \delta + \Delta_{t,n}(i) \text{ for } i \in O, \qquad \mu_{t,n}(i) \leq \delta\rho + \varepsilon \text{ for } i \notin O.$$

For $i \in O$, choose $\Delta_{t,n}(i) = \delta(1 - \langle a^\star, a_i\rangle)$; for $i \notin O$, $|\theta^\top a_i| \leq \delta\rho$ and pick $\Delta_{t,n}(i)$ within $[-\varepsilon, \varepsilon]$. Suboptimal pulls incur regret at least $c_0\delta$ for a universal $c_0 > 0$ (e.g. $c_0 = 1/2$ when $\rho \leq 1/2$).

Let $\mathcal{O}_t = \{i \in O : \exists s \leq t,\ n,\ i_{s,n} = i\}$. Until $\mathcal{O}_t = O$, some agent pulls a suboptimal arm, hence

$$R(T) \ \geq\ c_0\delta \sum_{t=1}^{T} \mathbf{1}\{\mathcal{O}_{t-1} \neq O\} \ \geq\ c_0\delta\,\tau, \tag{14}$$

where $\tau := \min\{t \leq T :\ \mathcal{O}_t = O\}$ (or $\tau = T+1$ if O is never covered). Thus $\mathbb{E}[R(T)] \geq c_0\delta\,\mathbb{E}[\min\{\tau, T\}]$.

By Yao's principle (Lattimore et al., 2020, Appendix A), lower bound $\mathbb{E}[\min\{\tau, T\}]$ for a deterministic algorithm when O is uniform. Let $Y := \max O$.

Lemma 12 (Last-discovery index)**.** *If O is uniform among $\binom{K}{N}$ subsets of $[K]$, then for $e \in \{N, \dots, K\}$,*

$$\mathbb{P}(Y = e) = \frac{\binom{e-1}{N-1}}{\binom{K}{N}}, \qquad \mathbb{E}[Y] = \frac{N(K+1)}{N+1}, \qquad \mathbb{E}[Y - N] = \frac{N(K-N)}{N+1}.$$

Proof. $\{Y = e\}$ holds iff $e \in O$ and $O \subseteq \{1, \dots, e\}$ with $|O| = N$, giving $\binom{e-1}{N-1}$ choices. Summing $e\,\mathbb{P}(Y = e)$ and using $\sum_{e=N}^{K} e\binom{e-1}{N-1} = \binom{K+1}{N+1}$ yields $\mathbb{E}[Y] = N(K+1)/(N+1)$. □

Any policy that has discovered all of O by time T must have selected each arm in O at least once. Identifying τ with the covering time of the N hidden optima in a uniformly randomized embedding gives $\tau = Y - N$ in the one-by-one probing argument of Section C, hence $\mathbb{E}[\min\{\tau, T\}] \geq \min\{NT,\ \mathbb{E}[Y - N]\}$. Combining with equation 14,

$$\mathbb{E}[R(T)] \ \geq\ c\,\varepsilon \min\left\{NT, \frac{N(K-N)}{N+1}\right\}\sqrt{\frac{d-1}{8\log K}}$$

for universal $c > 0$, proving the Yao-type term of Theorem 2.

Remark (heterogeneous K_n). For general K'_t (Section C), the same JL family yields gap δ and a search over K'_t hidden optima; the Yao-type term of Theorem 2 is the specialization $K'_t = N$ with $\mathbb{E}[Y - N] = \frac{N(K-N)}{N+1}$.

D.3 Proof of the Divergence-type term of Theorem 2

This section proves the Divergence-type term of Theorem 2 by adapting the information-theoretic template of Zhong et al. (combinatorial bandits, Theorem 6.1 and Lemmas 6.2–6.5) to our N-Selection multi-agent protocol (Section 5.2).

Fix N, K, T with $2 \le N \le K$ and $\varepsilon \in (0, 1/4)$, and take the cyclic family of Section 4. Write $S_k^\star$ for the good set $S^{*,k} = \{[k]_K, \dots, [k+N-1]_K\}$.

- **Instance** 0**.** $\mathbb{P}(r = 1 \mid i) = 1/2$ for all $i \in [K]$.
- **Instance** k **(**$1 \le k \le K$**).** $\mathbb{P}(r = 1 \mid i) = 1/2 + \varepsilon$ if $i \in S_k^\star$, else $1/2$.

These Bernoulli means are realized in the linear contextual model as in Section 4: a shared intercept $x_t(i) = 1$ and parameter $\theta = 1/2$, with an ε-bounded residual on $S^{*,k}$ (Definition 1).

At round t the policy selects $i_{t,1}, \dots, i_{t,N}$; let $S_t := \{i_{t,1}, \dots, i_{t,N}\}$. Then (Zhong et al., Lemma 6.2, additive form)

$$r(S_k^\star \mid k) - r(S_t \mid k) = \varepsilon \, |S_k^\star \setminus S_t|, \tag{15}$$

and

$$\mathbb{E}^{(k)}[R(T)] = \varepsilon \sum_{j \in [K] \setminus S_k^\star} \mathbb{E}^{(k)}[N_T(j)], \qquad N_T(j) := \sum_{t=1}^{T} \mathbf{1}\{j \in S_t\}. \tag{16}$$

Let $\mathbb{P}^{(k)}$ be the law of $\mathcal{H}_T = \{(i_{t,n}, r_{t,n})_{n \in [N]}\}_{t \in [T]}$ under instance k.

Lemma 13 (Coverage lower bound)**.** *For every policy,*

$$\frac{1}{K} \sum_{k=1}^{K} \sum_{j \in [K] \setminus S_k^\star} \mathbb{E}^{(k)}[N_T(j)] \;\ge\; NT \left(1 - \frac{N}{K} - \sqrt{\frac{1}{2K} \sum_{k=1}^{K} \mathrm{KL}\big(\mathbb{P}^{(0)}, \mathbb{P}^{(k)}\big)}\right). \tag{17}$$

Proof. For fixed k, let $A_k := \{\sum_{j \notin S_k^\star} N_T(j) \le NT(1 - N/K - \Delta)\}$. Lemma 4 (Bretagnolle–Huber) gives $\mathbb{P}^{(0)}(A_k) + \mathbb{P}^{(k)}(A_k^c) \ge \frac{1}{2} \exp(-\mathrm{KL}(\mathbb{P}^{(0)}, \mathbb{P}^{(k)}))$. Averaging over k and choosing $\Delta = \sqrt{\frac{1}{2K} \sum_k \mathrm{KL}(\mathbb{P}^{(0)}, \mathbb{P}^{(k)})}$ yields equation 17 (Zhong et al., Lemma 6.3). □

By the chain rule (Cover and Thomas, 2012; Zhong et al., Lemma 6.4),

$$\mathrm{KL}(\mathbb{P}^{(0)}, \mathbb{P}^{(k)}) = \sum_{t=1}^{T} \sum_{n=1}^{N} \mathbb{E}^{(0)}\left[\mathrm{KL}\big(\mathrm{Bern}(\tfrac{1}{2}) \,\big\|\, \mathrm{Bern}(\tfrac{1}{2} + \varepsilon \, \mathbf{1}\{i_{t,n} \in S_k^\star\})\big)\right]. \tag{18}$$

Each $j \in [K]$ lies in exactly N sets $S_1^\star, \dots, S_K^\star$ (uniform cover, Zhong et al., Lemma 6.5). By Lemma 5, $\mathrm{KL}(\mathrm{Bern}(1/2) \,\|\, \mathrm{Bern}(1/2 + \varepsilon)) \le 2\varepsilon^2$ for $0 < \varepsilon \le 1/4$.

Summing equation 18 over k and applying Lemma 5,

$$\frac{1}{K} \sum_{k=1}^{K} \mathrm{KL}(\mathbb{P}^{(0)}, \mathbb{P}^{(k)}) \le \frac{2\varepsilon^2}{K} \sum_{k=1}^{K} \sum_{j \in S_k^\star} \mathbb{E}^{(0)}[N_T(j)] \le 2\varepsilon^2 \cdot \frac{N}{K} \cdot NT.$$

Combine equation 16, Lemma 13, and $\sqrt{\frac{1}{2K} \sum_k \mathrm{KL}} \le \varepsilon N \sqrt{T/(2K)}$:

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}^{(k)}[R(T)] \ge \varepsilon NT \left(1 - \frac{N}{K} - \varepsilon \sqrt{\frac{NT}{2K}}\right).$$

Some instance $k \in [K]$ therefore satisfies the same bound in expectation, proving the Divergence-type term of Theorem 2 when equation 3 holds.

D.4 Proof of Corollary 1

This section proves Corollary 1, combining the Yao-type term of Theorem 2 (Section D.2) with the SupLinUCB-style upper bound (Section D.1, Theorem 1) on a *single* synchronized hard instance. The idea parallels Lattimore et al. (2020), where a misspecification gap ε and an estimation tolerance of order $\varepsilon\sqrt{d}$ are traded off on the same feature family; here we additionally align the confident-branch width $\alpha\sqrt{d}/\sqrt{NT}$ with the packing gap.

Throughout, let $m := \min\{NT, N(K-N)/(N+1)\}$ as in Corollary 1, assume condition equation 2, and take $\alpha = \beta(\delta)$ from Theorem 1.

Fix the near-orthogonal features $\{a_i\}_{i\in[K]}$ from Lemma 11 and a hidden optimal set $O \subset [K]$ with $|O| = N$ (N-Selection specialization, Section 5.2). Choose $a^\star \in \{a_i : i \in O\}$ and define the *synchronized* parameters from Section C:

$$\delta = \frac{\varepsilon}{2}\sqrt{\frac{d-1}{8\log K}}, \qquad \gamma = \frac{\alpha\sqrt{d}}{\sqrt{NT}}, \qquad \theta = (\delta+\gamma)\, a^\star. \tag{19}$$

Set static contexts $x_t(i) = a_i$ and expected rewards

$$\mu_{t,n}(i) = \theta^\top a_i + \Delta_{t,n}(i), \tag{20}$$

with residuals $|\Delta_{t,n}(i)| \le \varepsilon$ chosen as in Section D.2: for $i \in O$, $\Delta_{t,n}(i) = (\delta+\gamma)(1-\langle a^\star, a_i\rangle)$ so $\mu_{t,n}(i) = \delta+\gamma$; for $i \notin O$, $|\theta^\top a_i| \le (\delta+\gamma)\rho$ with $\rho = \sqrt{8\log K/(d-1)}$, and $\Delta_{t,n}(i)$ is picked in $[-\varepsilon, \varepsilon]$ to realize the required suboptimal rewards.

Lemma 14 (Feasibility under Corollary 1 conditions)**.** *If $\varepsilon^2 \le 4\alpha^2 d/(NT) \le 1$ and $d \ge \lceil 8\log K/\varepsilon^2\rceil$, then equation 20 satisfies Definition 1 and $|\mu_{t,n}(i)| \le 1$ for all t, n, i.*

Proof. On O, $|\mu_{t,n}(i)| = \delta+\gamma \le \varepsilon/2 + \alpha\sqrt{d/(NT)} \le 1/2 + 1/2 \le 1$ using $\delta \le \varepsilon/2$ and $\gamma \le 1$ from $\varepsilon^2 \le 4\alpha^2 d/(NT) \le 1$ with $\alpha \ge 1$ after scaling (if $\alpha < 1$, the bound is even smaller). Off O, $|\mu_{t,n}(i)| \le (\delta+\gamma)\rho + \varepsilon = \varepsilon/2 + \varepsilon + \gamma\rho \le 2\varepsilon + \gamma\rho \le 1$ for $\varepsilon \le 1/2$ and $\gamma \le 1$. Residuals satisfy $|\Delta_{t,n}(i)| \le \varepsilon$ by construction. □

Lemma 15 (Instantaneous gap on the synchronized instance)**.** *There exists $c_0 > 0$ (e.g. $c_0 = 1/4$ for $\rho \le 1/2$ and $\varepsilon \le 1/4$) such that every suboptimal pull (t, n) with $i_{t,n} \notin O$ satisfies*

$$\mu_{t,n}(i^\star_{t,n}) - \mu_{t,n}(i_{t,n}) \;\ge\; c_0(\delta+\gamma).$$

Proof. For $i^\star \in O$ and $i \notin O$, $\mu_{t,n}(i^\star) = \delta+\gamma$ while $\mu_{t,n}(i) \le (\delta+\gamma)\rho + \varepsilon$. Since $\delta\rho = \varepsilon/2$ and $\gamma\rho \le \gamma$,

$$\mu_{t,n}(i^\star) - \mu_{t,n}(i) \ge (\delta+\gamma)(1-\rho) - \varepsilon \ge \tfrac{1}{2}(\delta+\gamma) - \varepsilon \ge c_0(\delta+\gamma)$$

for universal c_0 when $\rho \le 1/2$. □

Existence of a hard embedding. The value m is achieved (up to constants) by the uniform-over-$\binom{K}{N}$ embedding in Lemma 12: there exists a choice of O such that any policy must incur search cost of order m before identifying all optima. Corollary 1 only requires *some* instance in the packing family; we fix such an O (e.g. $O = \{K-N+1, \dots, K\}$ when $T \ge K$) for the remainder.

Reuse the coverage time τ and sets $\mathcal{O}_t$ from Section D.2, now with gap $\delta+\gamma$ in place of δ. Until $\mathcal{O}_T = O$, some round has a suboptimal pull, so by Lemma 15,

$$R(T) \;\ge\; c_0(\delta+\gamma)\sum_{t=1}^{T} \mathbf{1}\{\mathcal{O}_{t-1} \ne O\} \;\ge\; c_0(\delta+\gamma)\,\min\{\tau, T\}. \tag{21}$$

Taking expectations and using $\mathbb{E}[\min\{\tau, T\}] \ge m$ (Section C, Lemma 12),

$$\mathbb{E}[R(T)] \;\ge\; c_0(\delta+\gamma)\, m \;=\; c_0\, m\left(\frac{\varepsilon}{2}\sqrt{\frac{d-1}{8\log K}} + \frac{\alpha\sqrt{d}}{\sqrt{NT}}\right). \tag{22}$$

Thus $R(T) \in \tilde{\Omega}\big((\varepsilon\sqrt{d/(8\log K)})\, m + (\alpha\sqrt{d/NT})\, m\big)$ in expectation; the statement in Corollary 1 records the same scaling (constants and polylogarithmic factors absorbed into $\tilde{\Omega}$).

Apply Theorem 1 to Algorithm 1 on the instance equation 20–equation 19, with $\alpha = \beta(\delta)$ and the same $(N, K, T, d, R, \varepsilon, c, \lambda)$. With probability at least $1 - \delta$,

$$R(T) = \tilde{O}\Big(R\sqrt{dNT\log(KN)} + \varepsilon\sqrt{d}\, NT + Bd\Big)\,. \tag{23}$$

Matching the dominant terms. Compare equation 22 and equation 23:

- **Misspecification.** The lower bound contributes $\tilde{\Omega}(\varepsilon\sqrt{d/(8\log K)}\, m)$ while the upper bound has $\tilde{O}(\varepsilon\sqrt{d}\, NT)$. On the synchronized instance, $m \le NT$, so $\varepsilon\sqrt{d}\, m \le \varepsilon\sqrt{d}\, NT$: the upper misspecification term dominates the lower misspecification term up to $\sqrt{\log K}$ and polylog factors—the same $\varepsilon\sqrt{d}$ scaling.
- **Statistical / confident branch.** The lower bound adds $\tilde{\Omega}((\alpha\sqrt{d/NT})\, m)$, which is exactly the threshold $\alpha\sqrt{d/(NT)}$ from the confident branch (Lemma 10) multiplied by the search horizon m. The upper bound contributes $\tilde{O}(R\sqrt{dNT\log(KN)})$ from exploration and $\tilde{O}(R\sqrt{dNT} + \varepsilon\sqrt{d}NT)$ from the confident branch (Lemma 10); when $\varepsilon^2 \le 4\alpha^2 d/(NT)$, $\gamma = \alpha\sqrt{d/NT}$ is the same order as the per-round UCB width that governs exploitation on this instance.
- **Burn-in.** The Bd term in equation 23 is instance-independent and negligible relative to NT for large T.

Hence, on one synchronized packing instance, the algorithm's high-probability regret lies in the $\tilde{\Omega}(\cdot) \cap \tilde{O}(\cdot)$ sandwich stated in Corollary 1.

Combining equation 22 (existence of the hard instance and $\mathbb{E}[R(T)]$ scale) with equation 23 (Algorithm 1 on that same instance) proves Corollary 1. The $\tilde{\Omega}$ part is inherited from the Yao-type term of Theorem 2 with the tuned split (δ, γ); the $\tilde{O}$ part is Theorem 1. Unlike Corollary 1, the Divergence-type term of Theorem 2 uses the cyclic family and does not require near-orthogonal features; the matching statement is specific to the synchronized packing construction.

E PSEUDO CODE OF VARIANT ALGORITHMS

Algorithm 2 Top-N sequential Multi-Agent SupLinUCB (NS-UCB)

Input: Number of agents N, number of rounds T, number of arms K, dimension d, parameters α, $c > 1$, λ, $S = \lceil \log_c(T/d)/2 \rceil$
Initialize for each stage $s \in [S]$:
  $V_s \leftarrow \lambda I_d$ ▷ Global covariance matrix for stage s
  $b_s \leftarrow \mathbf{0}_d$ ▷ Global vector for stage s
for round $t = 1$ to T **do**
  $I_t \leftarrow [K]$ ▷ Current candidate catalog, shared across agents
  Observe features $\{x_t(i) : i \in I_t\}$
  $s \leftarrow 1$
  while true **do**
    for each arm $i \in I_t$ **do**
      $\hat{r}_{t,s}(i) \leftarrow x_t(i)^\top V_s^{-1} b_s$
      $w_{t,s}(i) \leftarrow \alpha \|x_t(i)\|_{V_s^{-1}}$
    end for
    if $\forall i \in I_t,\ w_{t,s}(i) \le \alpha\sqrt{d/NT}$ **then** ▷ Confidence branch
      for agent $n = 1$ to N **do**
        Choose arm $i_{t,n} \leftarrow \arg\max_{i \in I_t} [\hat{r}_{t,s}(i) + w_{t,s}(i)]$
        Receive reward $r_{t,n}$
        $I_t \leftarrow I_t \setminus \{i_{t,n}\}$
      end for
      break
    else if $\forall i \in I_t,\ w_{t,s}(i) \le \alpha c^{-s}$ **then** ▷ Stage branch
      $U_t \leftarrow \max_{j \in I_t} [\hat{r}_{t,s}(j) + w_{t,s}(j)] - 2\alpha c^{1-s}$
      $I'_t \leftarrow \{i \in I_t \mid \hat{r}_{t,s}(i) + w_{t,s}(i) \ge U_t\}$
      if $|I'_t| \ge N$ **then**
        $I_t \leftarrow I'_t$
        $s \leftarrow s + 1$ ▷ Remaining catalog still large enough for all agents
      else
        for agent $n = 1$ to N **do**
          Choose arm $i_{t,n} \leftarrow \arg\max_{i \in I_t} [\hat{r}_{t,s}(i) + w_{t,s}(i)]$
          Receive reward $r_{t,n}$
          $I_t \leftarrow I_t \setminus \{i_{t,n}\}$
        end for
        break
      end if
    else ▷ Exploration branch
      for agent $n = 1$ to N **do**
        Choose arm $i_{t,n} \leftarrow i \in \{i \in I_t \mid w_{t,s}(i) > \alpha c^{-s}\}$
        Receive reward $r_{t,n}$
        Record exploration: agent n at stage s in round t
        $I_t \leftarrow I_t \setminus \{i_{t,n}\}$
      end for
      break
    end if
  end while
  for each stage $s \in [S]$ **do**
    for each agent n exploring an arm at stage s in round t **do**
      $V_s \leftarrow V_s + x_t(i_{t,n}) x_t(i_{t,n})^\top$ ▷ Update global covariance matrix
      $b_s \leftarrow b_s + r_{t,n} x_t(i_{t,n})$ ▷ Update global vector
    end for
  end for
end for

F EXPERIMENTAL CONFIGURATION

Table 2 summarizes hyperparameters shared across modules (Section 6):

Table 2: Default experimental hyperparameters (Section 6).

Parameter	Value
N, K, T, d	10, 100, 10,000, 64
$R, \varepsilon, c, \lambda$	0.5, 0.5, 2.0, 0.25
Misspecification	per-round $\mathrm{Unif}([-\varepsilon_n, \varepsilon_n])$ for every agent n
Seeds	`BASE_SEED`+2i, $i = 0, \ldots, n_{\text{seeds}} - 1$

- Figure 1: near-orthogonal features. In the heterogeneous setting (Figure 1c), agents $n \leq 2$ (indices $0, 1, 2$) use arms $[0, 50)$; agents $n \geq 7$ (indices $7, 8, 9$) use $[50, 100)$; the remaining agents use the full catalog $[K]$.
- Figure 2: $\varepsilon \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ with a structured arm distribution.
- Figure 3: agents $1, \ldots, N-1$ use sliding windows of width $\approx 0.4K$; agent N uses a random catalog of size $|K_n| \in \{1, 25, 50, 75, 100\}$. Good/bad clusters are generated with $\varepsilon_n \propto |K_n|/K$.