# Reward Teaching for Federated Multiarmed Bandits

Chengshuai Shi, Wei Xiong, Cong Shen, *Senior Member, IEEE*, and Jing Yang, *Senior Member, IEEE*

***Abstract*—Most of the existing federated multi-armed bandits (FMAB) designs are based on the presumption that clients will implement the specified design to collaborate with the server. In reality, however, it may not be possible to modify the clients' existing protocols. To address this challenge, this work focuses on clients who always maximize their individual cumulative rewards, and introduces a novel idea of "reward teaching", where the server guides the clients towards global optimality through implicit local reward adjustments. Under this framework, the server faces two *tightly coupled* tasks of bandit learning and target teaching, whose combination is non-trivial and challenging. A phased approach, called *Teaching-After-Learning (TAL)*, is first designed to encourage and discourage clients' explorations separately. General performance analyses of TAL are established when the clients' strategies satisfy certain mild requirements. With novel technical approaches developed to analyze the *warm-start* behaviors of bandit algorithms, particularized guarantees of TAL with clients running UCB or $\varepsilon$-greedy strategies are then obtained. These results demonstrate that TAL achieves logarithmic regrets while only incurring logarithmic adjustment costs, which is order-optimal w.r.t. a natural lower bound. As a further extension, the *Teaching-While-Learning (TWL)* algorithm is developed with the idea of successive arm elimination to break the non-adaptive phase separation in TAL. Rigorous analyses demonstrate that when facing clients with UCB1, TWL outperforms TAL in terms of the dependencies on sub-optimality gaps thanks to its adaptive design. Experimental results demonstrate the effectiveness and generality of the proposed algorithms.**

***Index Terms*—Federated learning, multi-armed bandits, reward teaching, upper confidence bound.**

## I. Introduction

FEDERATED multi-armed bandits (FMAB) [2], [3], [4], [5], [6], [7] is a recently proposed framework that introduces the core principles of federated learning (FL) [8], [9] into multi-armed bandits (MAB) [10], [11], [12]. In particular, FMAB often considers a system of one global server and multiple *heterogeneous* local clients with the goal of having the clients converge to the *global* optimality. Since proposed by [2], [3], FMAB has found applications in cognitive radio, recommender systems, and beyond.

One practical difficulty of realizing FMAB is that the existing designs have to implement new protocols for both the server and clients [3], [4], [13]. Specifically, the server and clients must strictly follow the design collaboratively. In real-world applications, it is relatively easy to update the server's protocols for FMAB. However, given the typically large number of clients, it is often not realistic to assume that all of their protocols can be updated due to infrastructure cost and complicated agent behaviors.

We first use the example of *cognitive radio systems*, a common motivating application for FMAB [3], [14], [15], for a more concrete illustration. Specifically, the base station (i.e., the central server) wants to find a good channel to broadcast information to mobile devices in its coverage area. However, different mobile devices, which are modeled as clients in FMAB, typically have different local channel availabilities due to their different geographic locations. As aforementioned, previous designs (e.g., [2], [3]) typically require mobile devices (i.e., clients) to follow the new FMAB protocols to collaborate with the base station. However, in reality, mobile devices are often configured to optimize their individual communication qualities following their built-in protocols. It is typically hard and expensive to update all mobile devices to follow the new FMAB designs, especially since such changes are often needed for both software and hardware.

Moreover, in the recommender system, another well-accepted application of FMAB [3], [4], [16], [17], [18], the online sellers (i.e., clients) often need to select items (i.e., actions) for promotions on the shopping platform (i.e., the server). However, these sellers typically follow their own strategies to optimize profits and often ignore other social influences, such as environmental effects and health concerns (e.g., for cigarettes). It is thus unrealistic to assume that the selfish sellers would strictly perform the previously proposed FMAB designs.

This work removes this limitation for FMAB by *designing mechanisms only at the server side*. Especially, the clients can still follow the original routines to optimize their individual performances (as in the aforementioned examples of cognitive radio and recommender systems) and no change of their protocols is required. Towards this end, a novel **"reward**

Manuscript received 3 April 2023; revised 3 September 2023 and 31 October 2023; accepted 3 November 2023. Date of publication 22 November 2023; date of current version 29 November 2023. The work of Chengshuai Shi and Cong Shen was supported in part by the U.S. National Science Foundation (NSF) under Awards 2029978, 2143559, and 2002902, Virginia Commonwealth Cyber Initiative, and the Bloomberg Data Science Ph.D. Fellowship. The work of Jing Yang was supported in part by the U.S. NSF under Awards 2030026, 2114542, and 1956276. An earlier version of this paper was presented at the 2023 IEEE International Symposium on Information Theory (ISIT) [DOI: 10.1109/ISIT54713.2023.10206444]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiao Fu. *(Corresponding author: Cong Shen.)*

Chengshuai Shi and Cong Shen are with Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: cs7ync@virginia.edu; cong@virginia.edu).

Wei Xiong is with the Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL 61801 USA (e-mail: wx13@illinois.edu).

Jing Yang is with the Department of Electrical Engineering, The Pennsylvania State University, University Park, PA 16802 USA (e-mail: yangjing@psu.edu).

Digital Object Identifier 10.1109/TSP.2023.3333658

**teaching"** approach is proposed: the server implicitly adjusts the local rewards perceived by the clients to influence their decision-making indirectly. We note that this idea is practical for the aforementioned applications. For cognitive radio, it is widely adopted in standard communication protocols for the base station to measure rewards (e.g., throughput) and send designed signals to mobile devices. In recommender systems, the bonuses received by the sellers are commonly designed and distributed by the shopping platform.

From a different perspective, this work can also be viewed as breaking the barrier of *naive clients* in the previous FMAB designs, where the clients unconditionally follow the server's instructions. Such naive behaviors are often unrealistic, while a more reasonable scenario (as in this work) is that the clients take actions to optimize their local performances, which may not always align with the server's global objective.

Note that the seemingly simple idea of reward adjustment brings considerable challenges for the server strategy. In particular, the server needs to determine how to adjust rewards to handle the following two tasks *simultaneously*: **bandit learning** and **target teaching**. On one hand, the server has to learn the *unknown* global model through the clients' actions, which are based on local observations and may not align with the server's global objective. Thus, reward adjustments should be carefully placed to have the clients explore with respect to (w.r.t.) the global information (instead of their local ones). On the other hand, even if the global model is learned successfully, the corresponding learning history has a cumulative effect on guiding the clients towards the learned target, as all historical (adjusted) rewards are considered by the client in her future decision-making. As a result, while having been studied individually (e.g., learning in MAB and teaching in data-poisoning MAB), the combination of these two tasks is novel and challenging as they are *tightly coupled*, which is detailed in Sec. IV.

The contributions of this work are summarized as follows.

- **A reward-teaching framework.** A novel idea of reward teaching is proposed to let the server design reward signals to guide clients with their own local strategies. This idea is practically appealing for FMAB systems as existing client protocols do not have to be modified – only the reward signals they receive are adjusted. From another perspective, it also provides a method to handle non-naive FMAB clients.
- **Client strategy-agnostic algorithm designs.** A phased approach, coined *"Teaching-After-Learning" (TAL)*, is proposed. It addresses the challenge of teaching in an unknown environment by separately encouraging and discouraging explorations in two phases. A more adaptive *"Teaching-While-Learning" (TWL)* algorithm is then developed to break the strict two-phased structure via the idea of *successive arm elimination*. It is worth noting that both TAL and TWL are agnostic to the clients' local strategies.
- **Client strategy-dependent analysis.** When the clients' local strategies satisfy some general properties, theoretical regret and cost guarantees of TAL are established. Particularizing these properties to UCB1 and $\varepsilon$-greedy [19] strategies at clients reveals that TAL can achieve a logarithmic regret while only incurring a logarithmic adjustment cost, which is order-optimal w.r.t. a natural lower bound. Regarding TWL, its advantage is rigorously established with clients running UCB1, where TWL achieves an improved performance dependency on the sub-optimality gaps than TAL due to its adaptive design. Moreover, one key ingredient to obtain these results is the novel technical approaches developed to analyze the *warm-start* behaviors of bandit algorithms, which may be of independent merit.
- **Experimental results.** The performance of the proposed designs is verified empirically. Especially, their effectiveness and generality are corroborated with different client strategies (i.e., UCB1, $\varepsilon$-greedy, Thompson sampling [20], and their mixtures), where the advantage of TWL is also evidenced.

## II. Related Works

**FMAB.** FMAB can be viewed as a variant of the general problem of multi-agent bandits [11], [12], [21], [22], [23], where global rewards instead of local ones measure the performance. Recent studies have investigated its robustness [24], personalization [16] and privacy protection [13], and extended the studies to contextual bandits [5], [6], [7]. However, almost all of the previous studies assume the clients follow updated local protocols, which either require clients to directly follow the server's instructions or have them work collaboratively. Instead, the designs in this work are purely on the server's side and no change is needed on the client side, which broadens the applicability of FMAB.

**Reward adjustments in MAB and RL.** One line of research on reward adjustments focuses on the malicious poisoning attacks [25], [26], [27]. The most relevant works are under the "strong attack" model [28], [29], [30], [31], where the attacker perturbs the rewards *after* observing the player's actions and tricks her into converging to a pre-selected sub-optimal arm (see Sec. IV). Other forms of attacks are also studied [32], [33], [34], [35], [36], including the "weak attack" model [37], [38], [39] where attacks are performed *before* observing actions. Note that the attackers in all these works have no desire to explore the environment, while the reward-teaching server has to actively learn the global model.

Another line is more *conceptually* related to this work: performing adjustments for positive purposes, such as reward shaping [40], [41], [42]. Especially, in reward shaping, the goal is to accelerate learning using a newly designed set of rewards; thus the optimal policy is kept the same. However, for reward teaching, the goal is to use modified rewards to guide clients to a different optimal policy (i.e., the optimal global arm). The recent work by [43] shares a similar idea of "teaching" the player via certain adjustments in reinforcement learning (RL); however, the target is still pre-selected. While differences exist between these previous attempts and this work, they all demonstrate the potential of "teaching" in MAB and RL.

In addition, the reward-teaching idea shares similarities with the design of implicit rewards in hierarchical RL [44], [45].

TABLE I
A SUMMARY OF KEY NOTATIONS USED IN THIS WORK

| Notations | Explanations |
|---|---|
| $M$ | The number of clients and local models |
| $K$ | The number of available arms |
| $X_{k,m}(t)$ | The reward of arm $k$ on local model $m$ at time $t$ |
| $Y_k(t)$ | The reward of arm $k$ on the global model at time $t$ |
| $\mu_{k,m}$ | The expected reward of arm $k$ on local model $m$ |
| $\nu_k$ | Th expected reward of arm $k$ on the global model |
| $k_{*,m}$ | The optimal arm for local model $m$ |
| $k_\dagger$ | The optimal arm for the global model |
| $\mu_{*,m}$ | The expected reward of the locally optimal arm $k_{*,m}$ on local model $m$ |
| $\mu_{\dagger,m}$ | The expected reward of the globally optimal arm $k_\dagger$ on local model $m$ |
| $\nu_\dagger$ | The expected reward of the globally optimal arm $k_\dagger$ on the global model |
| $X'_{\pi_m(t),m}(t)$ | The modified observation for client $m$'s action $\pi_m(t)$ at time $t$ |
| $\sigma_m(t)$ | The adjustment amount performed on client $m$'s observation at time $t$ |
| $R_m(T)$ | The cumulative global regret caused by client $m$ |
| $R_F(T)$ | The cumulative global regret caused by all clients |
| $C_m(T)$ | The cumulative cost for adjusting client $m$'s observations |
| $C_F(T)$ | The cumulative cost for adjusting all clients' observations |
| $\Delta_k$ | The suboptimality gap of arm $k$, i.e., $\nu_\dagger - \nu_k, \forall k \neq k_\dagger$ |
| $\Delta_{\min}$ | The minimal suboptimality gap, i.e., $\min_{k\neq k_\dagger} \Delta_k$ |
| $\Delta_{\max}$ | The maximal suboptimality gap, i.e., $\max_{k\neq k_\dagger} \Delta_k$ |

Thus, the designs in this work may contribute to improving the theoretical understanding of hierarchical RL, which is currently lacking. In particular, our work may be useful in demonstrating that client behaviors can be guided via a small number of modifications on their original rewards.

**Incentivized explorations in MAB and RL.** Another related research domain is the incentivized explorations in MAB and RL. Especially, a principal leverages either strategically designed signals [46], [47], [48] or additional compensations [49], [50], [51] to motivate the agent to perform certain actions. In particular, [52] leverages additional bonuses to motivate non-naive FMAB clients to perform certain explorations. However, comparing incentivized explorations with this work, we note that major differences exist: the incentivizing principal's signals or compensations are *explicit* to the agent, who then takes corresponding actions; however, the reward adjustment used by the server in this work is *implicit* to the clients, who autonomously perform their own local strategies.

## III. PROBLEM FORMULATION

### A. Federated Multiarmed Bandits

**Local and global models.** Following [2], [3], [4], a standard FMAB system of $M$ local models and one global model is considered. With the same set of $K$ arms shared by all the models, at each time step $t \in [T]$, each arm $k \in [K]$ is associated with a local reward $X_{k,m}(t) \in [0,1]$ for each local model $m \in [M]$ and a global reward $Y_k(t) \in [0,1]$ for the global model. These rewards of each arm $k$ are all independently sampled with unknown expectations denoted as $\mu_{k,m} := \mathbb{E}[X_{k,m}(t)], \forall m \in [M]$ and $\nu_k := \mathbb{E}[Y_k(t)]$. In general, the local arm utilities are model-dependent, i.e., $\mu_{k,m} \neq \mu_{k,n}$ for all $n \neq m$. The optimal local arm for each local model $m$ is denoted as $k_{*,m} := \arg\max_{k\in[K]} \mu_{k,m}$ with $\mu_{*,m} := \mu_{k_{*,m},m}$, and the optimal global arm as $k_\dagger := \arg\max_{k\in[K]} \nu_k$ with $\nu_\dagger := \nu_{k_\dagger}$.

As in [2], [3], [4], we consider the setting where each arm $k$'s mean reward on the global model is the average of its mean rewards on the local models[1], i.e.,

$$\nu_k := \mathbb{E}\left[Y_k(t)\right] = \frac{1}{M}\sum\nolimits_{m\in[M]} \mu_{k,m}.$$

As a result, a global-local misalignment may occur as the global optimality may not align with each local optimality, i.e., $k_\dagger \neq k_{*,m}$ for all or part of $m \in [M]$.

**Clients and server.** In FMAB, there exist $M$ clients and one server. At time $t$, each client $m \in [M]$ selects an arm $\pi_m(t)$ (referred to as "local actions") and then observes its local reward $X_{\pi_m(t),m}(t)$ on local model $m$. Additionally, each client $m$'s action $\pi_m(t)$ would also generate a reward $Y_{\pi_m(t)}(t)$ from the global model. It would be helpful to interpret the local and global rewards as the individual-level and system-level impact of the clients' actions.

The server in FMAB does not perform any arm-pulling action herself. Instead, she focuses on guiding the local actions to optimize their incurred *global rewards*. However, the global rewards are not directly observable by the server and the clients, which is often a result of practical measurement limitations [3]. Instead, the server is assumed to be able to observe the local actions and the corresponding local rewards, i.e., $\{\pi_m(t), X_{\pi_m(t),m}(t) : m \in [M]\}$.

To better optimize global performance, previous FMAB studies require that all clients work collaboratively following the updated local protocols. On the contrary, this work considers that clients are fully committed to interacting with their *own* local models (i.e., client $m$ with local model $m$). Then, the clients would naturally adopt their own MAB policies to maximize their local rewards. This setting is practically appealing as in many applications (e.g., the examples of cognitive radio and recommender systems in Sec. I), the local clients are inherently configured to perform local policies to optimize their local performance (e.g., IoT devices maximizing their own data rate and selfish sellers optimizing their profits). Specifically, at time $t$, each client $m$ *individually* makes an arm-pulling decision $\pi_m(t)$ based on her own history observed on local model $m$, i.e., $H_m(t-1) := \{\pi_m(\tau), X_{\pi_m(\tau),m}(\tau) : 1 \leq \tau \leq t-1\}$.

### B. Reward Teaching

As mentioned, each client $m$ would select suitable actions w.r.t. her own local model, which however may not necessarily meet the server's preference due to the global-local model misalignment. To address this challenge, the following reward-teaching mechanism is introduced for the server to indirectly influence the clients' action selections.

Specifically, after observing $\{X_{\pi_m(t),m}(t) : m \in [M]\}$, the server can adjust each client $m$'s local reward $X_{\pi_m(t),m}(t)$ to $X'_{\pi_m(t),m}(t)$ by an amount of $\sigma_m(t)$, i.e.,

$$X'_{\pi_m(t),m}(t) := X_{\pi_m(t),m}(t) + \sigma_m(t),$$

[1]Other global-local model relationships can also be considered, e.g., the weighted sum in [16]. To better convey the key idea of reward teaching, the exact average, which is simple while representative, is adopted in this work.

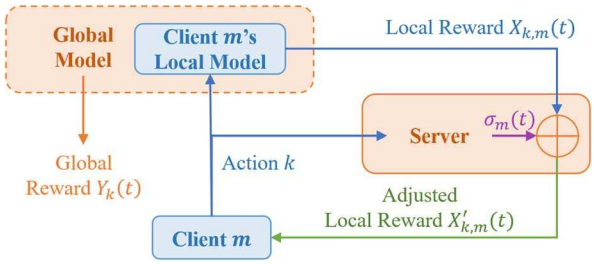



Fig. 1. The reward-teaching process with client $m$ (among the overall $M$ clients) and action $\pi_m(t) = k$.

which is then revealed to the client (instead of $X_{\pi_m(t),m}(t)$). Note that one implicit constraint is that the adjusted rewards must still be in $[0, 1]$, which is the system limitation.[2] If this constraint is satisfied, the clients are assumed to be unable to detect the reward adjustments by any means. The adjusted rewards lead to an adjusted history of $H'_m(t) := \{\pi_m(\tau), X'_{\pi_m(\tau),m}(\tau) : 1 \le \tau \le t\}$ for client $m$, which ideally can shape her future actions in favor of the server.

It is worth emphasizing that such reward adjustments are practical for FMAB applications. In the cognitive radio example, it is common for the base station to first measure the communication quality (via pilot signals) and then send *designed* feedback to the devices; this is the case in both cellular and WiFi. Adjusting rewards can be achieved via either sending modified feedback signals or modifying the allocated resources (e.g., retransmission bandwidth [53]) to boost or reduce client performance, which is standard in modern communication protocols. The devices, on the other hand, are oblivious to such adjustments thanks to their built-in protocols. In the application of recommender systems, the shopping platform can implicitly leverage extra or decreased bonuses to guide the decisions of the selfish sellers, e.g., to promote more environmentally friendly and healthier items.

The reward-teaching process is summarized as the following steps, which is also illustrated in Fig. 1:

- Each client $m$ chooses $\pi_m(t)$ using history $H'_m(t-1)$;
- The server observes $\{\pi_m(t), X_{\pi_m(t),m}(t) : m \in [M]\}$;
- The server adjusts $X_{\pi_m(t),m}(t)$ into $X'_{\pi_m(t),m}(t)$ by the amount of $\sigma_m(t)$ for each client $m \in [M]$;
- Each client $m$ observes the adjusted $X'_{\pi_m(t),m}(t)$.

### C. Learning Objectives

Following previous FMAB studies, the global view by the server is the focus of our design, which leads to a two-fold objective. First, the server's main goal is to maximize the cumulative **global** rewards and can be characterized by minimizing the *global regret*, defined as

$$R_F(T) := \sum\nolimits_{m\in[M]} R_m(T),$$

[2]In fact, if there is no restriction on the adjustment range, the server is more powerful and the algorithm design is thus easier.

where $R_m(T)$ is the regret incurred by client $m$'s actions w.r.t. the global model (instead of her local model) defined as

$$R_m(T) := T\nu_\dagger - \mathbb{E}\left[\sum\nolimits_{t\in[T]} Y_{\pi_m(t)}(t)\right].$$

The expectation is w.r.t. both the reward generations and the client-system interactions. Second, the server's adjustments on local rewards are often costly. For example, in the aforementioned application of cognitive radio, the base station naturally needs to make additional efforts when modifying the originally allocated resources, e.g., infrastructure costs for deviating from the default transmission bandwidth. This work, thus, further introduces the objective of *cumulative cost*, defined as

$$C_F(T) := \sum\nolimits_{m\in[M]} C_m(T),$$

where $C_m(T)$ denotes the overall cost spent on client $m$ and is further defined as

$$C_m(T) := \mathbb{E}\left[\sum\nolimits_{t\in[T]} |\sigma_m(t)|\right].$$

The subscripts $F$ in $R_F(T)$ and $C_F(T)$ refer to the global model (i.e., the federation).

Intuitively, there exists a trade-off between these two objectives: with more adjustments on rewards, i.e., larger $C_F(T)$, the server can have a bigger impact on the clients' decisions, which ideally would decrease the regret $R_F(T)$. It is thus important to strike a balance between these two objectives, which is the focus of the remainder of this paper.

### D. Client Strategies

To facilitate discussion, we denote client $m$'s local bandit policy as $\Pi_m$. Note that while performing their own policies, the clients are assumed not to be strategically against the server, which is reasonable for most of the real-world applications of FMAB, e.g., autonomous but not fully flexible mobile devices in cognitive radio [3]. In addition, we denote $N_{k,m}(t)$ as the number of pulls by client $m$ on arm $k$ by time $t$, and $N^{-1}_{k,m}(\tau)$ refers to the time step $t$ such that $N_{k,m}(t) = \tau$.

The proposed designs are general and agnostic to clients' strategies, which will be evident in Sections V and VII. For the theoretical analysis, general performance bounds are first provided without specifying the clients' strategies. This is accomplished by identifying the properties of client strategies that lead to the desired theoretical results. More specifically, client-strategy-dependent bounds are then derived (i.e., clients with UCB1 or $\varepsilon$-greedy). Finally, experiments with varying (and even mixing) strategies for clients are reported.

## IV. Two Coupled Tasks and Design Objectives

In this section, two tightly coupled tasks faced by the reward-teaching server, bandit learning, and target teaching, are elaborated. A system design objective is also proposed.

**Bandit learning.** One major distinction between learning in FMAB and in classical MAB [10], [54] is the server can only gather information through clients' local actions. Previous

FMAB studies tackled this challenge by implementing new protocols for clients to naively follow [2], [3], [4], [13]. In contrast, in this work, such information collection can only be indirectly guided via carefully designed rewards.

**Target teaching.** To understand teaching, a special case is first considered where the optimal arm $k_\dagger$ is known by the server. Then, the goal is to assign adjustments to have the clients pull the *pre-specified* arm $k_\dagger$ as much as possible, which is mathematically the same as the *data-poisoning MAB* problem [26], [28], [29], [55], where adjustments are phrased as "attacks". In such scenarios, the server can achieve $R_m(T) = O(\log(T))$ and $C_m(T) = O(\log(T))$ for each $m \in [M]$ by adjusting rewards from all arms except arm $k_\dagger$ to 0's [30]. The underlying philosophy is to "discourage explorations" with the adjusted reward 0's.

**Combination leads to a tight coupling.** While both tasks have been separately investigated (to some extent), the reward-teaching server faces a combination of them. On one hand, even if the server can perfectly learn the global model, she still needs to teach it to the clients. On the other hand, to teach correctly, sufficient information must be learned by the server. The resulted *tight coupling* is the main challenge of the design. Specifically, the learning attempt has a cumulative effect on teaching, which in return relies on the learned target. Technically, the main resultant difficulty is the analysis of the "warm-start" behaviors of bandit algorithms, which is elaborated in Sec. VI.

**Design objective.** For the cost, with a known target arm, [30], [31] prove lower bounds that *with UCB1 and $\varepsilon$-greedy clients (defined in Sec. VI), it is necessary to spend a cost $C_m(T) = \Omega(\log(T))$ to obtain a regret $R_m(T) = O(\log(T))$*. Thus, with $M$ *independent* FMAB clients, a cost of $C_F(T) = \Omega(M\log(T))$ is required to obtain a regret of $R_F(T) = O(M\log(T))$ while knowing arm $k_\dagger$, which naturally holds for the more stringent case of not knowing the target $k_\dagger$. For the regret, UCB1 and $\varepsilon$-greedy clients can be shown to be conservative [30] as each client $m$ would pull each arm at least $\Omega(\log(T))$ times regardless of the rewards; thus $R_m(T) = \Omega(\log(T))$ and $R_F(T) = \Omega(M\log(T))$.

With these results, the following system design goal is established, which is order-wise tight w.r.t. both criteria:

> ***Goal:*** *Design algorithms to achieve both* $R_F(T) = O(M\log(T))$ *and* $C_F(T) = O(M\log(T))$.

To verify that this goal is non-trivial, two intuitive baseline policies, NG and NA, are discussed as follows, whose limitations are further illustrated experimentally in Sec. VIII.

- **"Naively-Guess" (NG).** The server may randomly initialize one arm $k'$ as the target to adopt the aforementioned approach from [30]. However, the regret would be $R_F^{\text{NG}}(T) = \Omega(MT)$ if $k' \neq k_\dagger$, although achieving $C_F^{\text{NG}}(T) = O(M\log(T))$.
- **"Naively-Align" (NA).** Another natural idea is to have the server align $X'_{\pi_m(t),m}(t)$ with $Y_{\pi_m(t)}(t)$ via $\sigma_m(t) = Y_{\pi_m(t)}(t) - X_{\pi_m(t),m}(t)$.[3] While achieving $R_F^{\text{NA}}(T) = O(M\log(T))$, adjustments would be needed nearly all the time steps, i.e., $C_F^{\text{NA}}(T) = \Omega(MT)$.

*Remark 1:* A refined lower bound beyond $\Omega(M\log(T))$ can be instructive, especially for determining the optimal dependencies on parameters other than $M$ and $T$. However, such lower bounds are also challenging, even with a known target [30], [31]; thus it is left as an open question for future works.

## V. TAL: Algorithm Design

To address the coupled tasks of bandit learning and target teaching, one idea is to first learn the server's target and then teach the clients to converge to it, which leads to the proposed "Teaching-After-Learning" (TAL) algorithm (presented in Alg. 1). Specifically, TAL starts with the learning phase where the goal is to identify the optimal global arm. Then, in the teaching phase, the server guides the clients toward the learned global optimality. Note that although there is a separation of phases, the teaching phase must handle clients that accumulate observations from the learning phase (i.e., "warm-start" clients), whose effect will be more evident in the analysis.

In the learning phase, TAL uniformly adjusts each client $m$'s observed rewards to $\gamma_1$, i.e., $\sigma_m(t) \leftarrow \gamma_1 - X_{\pi_m(t),m}(t)$, where $\gamma_1 \in [0,1]$ is a to-be-specified input parameter. Intuitively, this uniform reward adjustment encourages sufficient (or ideally, uniform) explorations among all arms, since their rewards are all at the same value $\gamma_1$. If clients are indeed sufficiently exploring, the server can collect enough information on each arm to identify her optimal arm $k_\dagger$.

This identification is designed to proceed in epochs indexed by counter $\psi$ to ensure statistical independence. If at time $t$, each client $m$ has pulled each arm $k$ at least $F(\psi) := \sum_{\tau\in[\psi]} f(\tau)$ times, where $f(\psi) := \frac{1}{M}\cdot 2^{2\psi+3}\log(2KT^2)$, the server updates upper and lower confidence bounds (UCB and LCB) for each arm $k \in [K]$ using its rewards collected between its $F(\psi-1)+1$ and $F(\psi)$ pulls (i.e., overall $f(\psi)$ pulls) by each client as follows:

$$\text{UCB}_k(\psi), \text{LCB}_k(\psi) := \frac{1}{M}\sum\nolimits_{m\in[M]} \hat{\mu}_{k,m}(\psi) \pm \text{CB}(\psi), \tag{1}$$

where

$$\hat{\mu}_{k,m}(\psi) := \sum\nolimits_{\tau=F(\psi-1)+1}^{F(\psi)} X_{k,m}(N_{k,m}^{-1}(\tau))/f(\psi),$$
$$\text{CB}(\psi) := \sqrt{\log(2KT^2)/(2Mf(\psi))} = 2^{-\psi-2}.$$

Note that with the estimation of $\mu_{k,m}$ from local samples, the first term in Eqn. (1) is essentially an estimation $\hat{\nu}_k(\psi)$ of $\nu_k$. The confidence bound $\text{CB}(\psi)$ is specifically designed such that $\text{LCB}(\psi) \le \nu_k \le \text{UCB}(\psi)$ holds for each arm $k$ and each epoch $\psi$ in the learning phase with high probability.

The learning phase ends in epoch $\psi$ if the confidence interval of one arm $k_\ddagger$ dominates that of all other arms, i.e., $\text{LCB}_{k_\ddagger}(\psi) \ge \text{UCB}_k(\psi), \forall k \neq k_\ddagger$, which is recognized as the

[3] $Y_{\pi_m(t)}(t)$ is assumed to be observable here for the baseline, which is not the case in our design.

**Algorithm 1** TAL

**Input:** Parameter $\gamma_1, \gamma_2 \in [0,1]$; Time Horizon $T$

Initialize: $F \leftarrow 1$ (i.e., the learning phase); $\psi \leftarrow 1$; $k_\ddagger \leftarrow 0$
**for** $t \le T$ **do**
  Observe $\{\pi_m(t), X_{\pi_m(t),m}(t) : m \in [M]\}$
  **if** $F = 1$ & $N_{k,m}(t) \ge F(\psi), \forall m \in [M], k \in [K]$ **then**
    Update $\{\mathrm{UCB}_k(\psi), \mathrm{LCB}_k(\psi) : k \in [K]\}$ as Eqn. (1)
    **if** $\exists j \in [K], \mathrm{LCB}_j(\psi) \ge \mathrm{UCB}_k(\psi), \forall k \ne j$ **then**
      Set $k_\ddagger \leftarrow j$; $F \leftarrow 2$ (i.e., the teaching phase)
    **else** Set $\psi \leftarrow \psi + 1$
    **end if**
  **end if**
  **if** $F = 1$ **then** $\sigma_m(t) \leftarrow \gamma_1 - X_{\pi_m(t),m}(t), \forall m \in [M]$
  **else if** $F = 2$ **then** Set $\sigma_m(t)$ as Eqn. (2), $\forall m \in [M]$
  **end if**
  Set $X'_{\pi_m(t),m}(t) \leftarrow X_{\pi_m(t),m}(t) + \sigma_m(t), \forall m \in [M]$
  Reveal $X'_{\pi_m(t),m}(t)$ to each client $m \in [M]$
**end for**

optimal arm. Otherwise, a new epoch $\psi + 1$ begins. With the designed confidence bound, this identification is guaranteed to be correct with high probability.

With the identified arm $k_\ddagger$, the server utilizes the following adjustments to guide the clients in the teaching phase:

$$\sigma_m(t) \leftarrow \begin{cases} \gamma_2 - X_{\pi_m(t),m}(t) & \text{if} \pi_m(t) \ne k_\ddagger \\ 0 & \text{if} \pi_m(t) = k_\ddagger \end{cases}, \tag{2}$$

where $\gamma_2$ is another to-be-specified input parameter and typically should be small. In other words, if the client does not pull arm $k_\ddagger$, her reward is adjusted to a small value $\gamma_2$ to discourage explorations; otherwise, the original reward of arm $k_\ddagger$ is kept unchanged to save adjustments.

From Alg. 1, it can be observed that TAL is a pure server protocol and agnostic to the clients' local strategies – the only interaction with the clients is the adjusted rewards.

## VI. TAL: Theoretical Analysis

In this section, we first provide a general analysis of TAL (Theorem 5) under some abstract characterizations of clients' strategies (i.e., sufficient-exploring and warm-starting in Definitions 1 and 3, respectively). Then, we consider clients with UCB1 or $\varepsilon$-greedy in the following two subsections, respectively. In particular, the adopted abstract characterizations are particularized (Lemmas 6, 7, 9 and 10), and then specific performance guarantees are obtained (Theorems 8 and 11), which show that TAL achieves the design goals in Section IV with these clients. Detailed proofs are deferred to Appendix A.

Some useful notations are introduced as follows: $\Delta_k := \nu_\dagger - \nu_k, \forall k \ne k_\dagger$, $\Delta_{\min} = \Delta_{k_\dagger} := \min_{k \ne k_\dagger} \Delta_k$, $\Delta_{\max} := \max_{k \in [K]} \Delta_k$, and $\mu_{\dagger,m} := \mu_{k_\dagger,m}$. Moreover, $\delta_{k,m}(\gamma) := \mathbb{E}[|\gamma - X_{k,m}(t)|]$ and $\psi_{\max} := \lceil \log_2(1/\Delta_{\min}) \rceil$. Also, without loss of generality, it is assumed that $K, M \ll T$.

We first define sufficiently exploring algorithms for the learning phase in TAL, which states that a bandit algorithm would sufficiently explore when facing uniform rewards.

*Definition 1:* (Sufficiently Exploring Algorithms). Consider a $K$-armed bandit environment where rewards from arms in a set $\mathcal{I} \subseteq [K]$ are always a fixed constant $\gamma \in [0,1]$. In this environment, a bandit algorithm $\Pi$ is said to be $(\mathcal{I}, \gamma, \underline{\eta}, \overline{\eta})$-sufficiently exploring if it would pull each arm in the set $\mathcal{I}$ at least $\underline{\eta}(\tau; \gamma, \mathcal{I})$ and at most $\overline{\eta}(\tau; \gamma, \mathcal{I})$ times when total $\tau$ pulls have been performed on set $\mathcal{I}$.

If local strategies are sufficiently exploring as in Definition 1, enough information can be collected in the learning phase to identify the global optimal arm, as stated in the following lemma, where $\underline{\eta}^{-1}(N; \gamma, [K])$ denotes the value $\tau$ such that $\underline{\eta}(\tau; \gamma, [K]) = N$.

*Lemma 2:* (Learning Phase in TAL). If $\Pi_m$ is $([K], \gamma_1, \underline{\eta}_m, \overline{\eta}_m)$-sufficiently exploring for all $m \in [M]$, with probability (w.p.) at least $1 - 1/T$, the learning phase ends with $k_\ddagger = k_\dagger$ by time step $T_1$, and the regret and cost in the learning phase of TAL are bounded, respectively, as

$$R^{\mathrm{TAL}}_{F,1}(T) \le \sum\nolimits_{m \in [M]} \sum\nolimits_{k \ne k_\dagger} \Delta_k \cdot \overline{\eta}_m(T_1; \gamma_1, [K]);$$
$$C^{\mathrm{TAL}}_{F,1}(T) \le \sum\nolimits_{m \in [M]} \sum\nolimits_{k \in [K]} \delta_{k,m}(\gamma_1) \cdot \overline{\eta}_m(T_1; \gamma_1, [K]),$$

where $T_1 \le \max_{m \in [M]} \{\underline{\eta}_m^{-1}(F(\psi_{\max}); \gamma_1, [K])\}$.

Note that the time step $T_1$ bounded via the sufficiently exploring lower bound (i.e., $\underline{\eta}$) ensures sufficient information collection, while the corresponding upper bound (i.e., $\overline{\eta}$) guarantees performance, i.e., regret and adjustment cost.

Then, for the teaching phase, since the cumulative observations from the learning phase are inherited to the client strategies, we can view the clients as "warm-started". The following notion of warm-start pulls is introduced, which measures the warm-start behavior of an algorithm.

*Definition 3:* (Warm-start Pulls). In a $K$-armed bandit environment $\mathcal{B}$, if a reward sequence $H = \{H_k : k \in [K]\}$ is input to a bandit algorithm $\Pi$, where $H_k$ is a reward sequence for arm $k$, warm-start pulls on arm $k$ is defined as $\iota_k(T; H, \mathcal{B}, \Pi) := \mathbb{E}_\Pi[\sum_{t \in [T]} \mathbb{1}\{\pi(t) = k\} | H, \mathcal{B}]$, which represents the expected pulls performed by $\Pi$ on each arm $k$ during $T$ steps in environment $\mathcal{B}$ with prior input $H$.

Using this notion of warm-start pulls, the following guarantee on the teaching phase is established.

*Lemma 4:* (Teaching Phase in TAL). If the event in Lemma 2 occurs, the regret and cost in the teaching phase of TAL are bounded, respectively, as

$$R^{\mathrm{TAL}}_{F,2}(T) \le \sum_{m \in [M]} \max_{H_m \in \mathcal{H}_m} \sum_{k \ne k_\dagger} \Delta_k \cdot \iota_k(T; H_m, \mathcal{B}_m, \Pi_m);$$
$$C^{\mathrm{TAL}}_{F,2}(T) \le \sum_{m \in [M]} \max_{H_m \in \mathcal{H}_m} \sum_{k \ne k_\dagger} \delta_{k,m}(\gamma_2) \cdot \iota_k(T; H_m, \mathcal{B}_m, \Pi_m),$$

where $\mathcal{B}_m$ denotes an environment with constant rewards as $\gamma_2$ for arm $k \ne k_\dagger$ and stochastic rewards with expectation $\mu_{\dagger,m}$ for arm $k_\dagger$. The set $\mathcal{H}_m$ is defined with each element of it as a reward sequence $H_m = \{H_{k,m} : k \in [K]\}$ where $H_{k,m} \in \{\{\gamma_1\}^\tau : \tau \in [\underline{\eta}_m(T_1; \gamma_1, [K]), \overline{\eta}_m(T_1; \gamma_1, [K])]\}$.

Note that $\mathcal{B}_m$ characterizes the environment of client $m$ in the teaching phase while $\mathcal{H}_m$ represents the cumulative observation inherited from the learning phase.

Finally, the overall performance guarantee can be obtained by combining the regrets from two phases.

*Theorem 5:* (Overall Performance of TAL). Under the assumption in Lemma 2, with $R_{F,1}^{\mathrm{TAL}}(T), C_{F,1}^{\mathrm{TAL}}(T)$ defined in Lemma 2 and $R_{F,2}^{\mathrm{TAL}}(T), C_{F,2}^{\mathrm{TAL}}(T)$ in Lemma 4, the regret and cost of TAL are bounded, respectively, as

$$R_F^{\mathrm{TAL}}(T) \le R_{F,1}^{\mathrm{TAL}}(T) + R_{F,2}^{\mathrm{TAL}}(T) + O(M);$$
$$C_F^{\mathrm{TAL}}(T) \le C_{F,1}^{\mathrm{TAL}}(T) + C_{F,2}^{\mathrm{TAL}}(T) + O(M).$$

The key difficulty behind this analysis resides in leveraging the quantities in Definitions 1 and 3. In particular, how to specify $\underline{\eta}$, $\overline{\eta}$ and $\iota$ is non-trivial, which is one of the main technical challenges in proving Thm. 5. Furthermore, Thm. 5 implies that the desired logarithmic regret and cost can be achieved by TAL when $R_{F,1}^{\mathrm{TAL}}(T), R_{F,2}^{\mathrm{TAL}}(T), C_{F,1}^{\mathrm{TAL}}(T)$ and $C_{F,2}^{\mathrm{TAL}}(T)$ are all bounded in logarithmic orders. The analyses of these terms are further determined by the sufficiently exploring property and the warm-start pulls of the specific clients' strategies as stated in Lemmas 2 and 4.

In the following, to particularize the general guarantee in Thm. 5, we analyze several well-known bandit algorithms as clients' strategies (i.e., UCB and $\varepsilon$-greedy).

### A. UCB Clients

The popular UCB-type algorithms are first considered. In particular, we analyze the celebrated UCB1 algorithm [19] while noting that the analysis generalizes to other UCB variants [56], [57]. Especially, at time $t$, the UCB1 algorithm for client $m$ chooses arm as follows:

$$\pi_m(t) = \arg\max_{k\in[K]} \left\{ \hat{\mu}'_{k,m}(t-1) + \sqrt{2\log(t)/N_{k,m}(t-1)} \right\},$$

which considers both the perceived sample mean

$$\hat{\mu}'_{k,m}(t) := \sum\nolimits_{\tau\in[N_{k,m}(t)]} X'_{k,m}(N_{k,m}^{-1}(\tau))/N_{k,m}(t)$$

and the associated confidence bound.

First, the sufficiently exploring assumption in Lemma 2 is verified for UCB1 in Lemma 6. This is intuitive as with constant rewards, the sample means are the same while additional pulls decrease the confidence bound in UCB1.

*Lemma 6:* For any $\gamma \in [0,1]$ and set $\mathcal{I} \subseteq [K]$, UCB1 is $(\mathcal{I}, \gamma, \underline{\eta}, \overline{\eta})$-sufficiently exploring with $\underline{\eta}(\tau;\gamma,\mathcal{I}) = \lfloor \tau/|\mathcal{I}| \rfloor$ and $\overline{\eta}(\tau;\gamma,\mathcal{I}) = \lceil \tau/|\mathcal{I}| \rceil$.

Then, the performance of TAL in the learning phase (in Lemma 2) can be bounded by recognizing $T_1 = O(K\log(T)/(M\Delta_{\min}^2))$, which further specifies the reward sequence set $\mathcal{H}_m$ in Lemma 4 and leads to the following lemma on the warm-start pulls of UCB1.

*Lemma 7:* If $\gamma_1 \ge \mu_{\dagger,m} > \gamma_2$ and $\Pi_m$ is UCB1, for all $k \ne k_\dagger$, it holds that $\max_{H_m\in\mathcal{H}_m}\{\iota_k(T; H_m, \mathcal{B}_m, \Pi_m)\} = O\left(\frac{(\gamma_1-\gamma_2)T_1}{K(\mu_{\dagger,m}-\gamma_2)} + \frac{\log(T)}{(\mu_{\dagger,m}-\gamma_2)^2}\right)$.

Proving this lemma is non-trivial and may be of independent interest in understanding the warm-start behavior of UCB1. Essentially, the result can be interpreted as first offsetting the "warm-start" history (the first term) and then converging to arm $k_\dagger$ (the second term) in an environment $\mathcal{B}_m$, whose rewards for arm $k \ne k_\dagger$ are constant $\gamma_2$'s and rewards for arm $k_\dagger$ have an expectation $\mu_{\dagger,m}$ (see Lemma 4).

It is noted that Lemma 7 requires $\gamma_1 \ge \mu_{\dagger,m}$, which maintains the optimism for the estimation of arm $k_\dagger$ on each local model $m$. The other requirement $\mu_{\dagger,m} > \gamma_2$ is intuitive as otherwise, the local client $m$ would not converge to arm $k_\dagger$. Since there is no prior information about $\mu_{\dagger,m}$, a feasible and sufficient solution is to set $\gamma_1 = 1$ while $\gamma_2 = 0$, which leads to the following theorem.

*Theorem 8:* (TAL with UCB1 clients). For TAL with $\gamma_1 = 1$ and $\gamma_2 = 0$, if all clients run UCB1 locally and $\mu_{\dagger,m} \ne 0$ for all $m \in [M]$, it holds that

$$R_F^{\mathrm{TAL}}(T) = O\left( \sum_{m\in[M]} \sum_{k\ne k_\dagger} \left[ \frac{\Delta_k \log(T)}{\mu_{\dagger,m} M\Delta_{\min}^2} + \frac{\Delta_k \log(T)}{\mu_{\dagger,m}^2} \right] \right);$$

$$C_F^{\mathrm{TAL}}(T) = O\Bigg( \sum_{m\in[M]} \sum_{k\in[K]} \frac{(1-\mu_{k,m})\log(T)}{M\Delta_{\min}^2} + \sum_{m\in[M]} \sum_{k\ne k_\dagger} \left[ \frac{\mu_{k,m}\log(T)}{\mu_{\dagger,m} M\Delta_{\min}^2} + \frac{\mu_{k,m}\log(T)}{\mu_{\dagger,m}^2} \right] \Bigg).$$

We note that with a focus on the dependencies on $M$ and $T$, the regret and cost are both of order $O(M\log(T))$; thus TAL is order-optimal w.r.t. both criteria stated in Sec. IV, i.e., the general design goal is achieved. Moreover, the regret bound shows two dominating terms, which are from Lemma 7, i.e., the teaching phase. In fact, there is another non-dominating (thus hidden) term from Lemma 2 for the learning phase; see more details in section B of Appendix A. A similar three-part form is shared by the cost. In particular, the first term is from the learning phase (thus the sum is over all arms $k \in [K]$ and each term scales with $1-\mu_{k,m}$), and the last two terms are from the teaching phase (thus the sum is over sub-optimal global arms $k \ne k_\dagger$ and scales with $\mu_{k,m}$).

### B. $\varepsilon$-Greedy Clients

The analysis is further extended to the clients running the $\varepsilon$-greedy algorithm [58], another well-known bandit strategy. Especially, the $\varepsilon$-greedy algorithm for client $m$ is as follows:

$$\pi_m(t) \leftarrow \begin{cases} \arg\max_{k\in[K]} \hat{\mu}'_{k,m}(t-1) & \text{w.p.} 1-\varepsilon_m(t) \\ \text{a random arm in } [K] & \text{w.p.} \varepsilon_m(t) \end{cases},$$

where the exploration probability $\varepsilon_m(t) \in [0,1]$ is taken as $\varepsilon_m(t) = O(K/t)$, following [19].

First, the following lemma states that $\varepsilon$-greedy is sufficiently exploring, which is intuitive as the constant rewards lead to the same sample mean for different arms.

*Lemma 9:* For any $\gamma \in [0,1]$, if ties among arms are broken uniformly at random, with probability at least $1-1/T$, $\varepsilon$-greedy is $([K], \gamma, \underline{\eta}, \overline{\eta})$-uniformly exploring with $\underline{\eta}(\tau;\gamma,[K])$ and $\overline{\eta}(\tau;\gamma,[K]) = O(\tau/K \pm \log(KT))$.

Due to the randomness in $\varepsilon$-greedy, it is complicated to analyze its warm-start pulls in general. Instead, the following lemma focuses on $\gamma_1 = \gamma_2 = 0$. Under this setting, the sample means are all kept as zero in the learning phase. Thus, once a

non-zero reward is collected in the teaching phase, that arm will immediately have the highest sample mean.

*Lemma 10:* If $\Pi_m$ is $\varepsilon$-greedy and $\mu_{\dagger,m} > \gamma_1 = \gamma_2 = 0$, with probability at least $1 - 1/T$, it holds that $\max_{H_m \in \mathcal{H}_m} \{\sum_{k \neq k_\dagger} \iota_{k,m}(T; H_m, \mathcal{B}_m, \Pi_m)\} = O(K \log(KT)/\mu_{\dagger,m}^2)$.

Combining these results with Thm. 5, the following performance guarantees can be obtained.

*Theorem 11:* (TAL with $\varepsilon$-greedy clients). For TAL with $\gamma_1 = \gamma_2 = 0$, if clients run $\varepsilon$-greedy and break ties uniformly at random, and $\mu_{\dagger,m} \neq 0, \forall m \in [M]$, it holds that

$$R_F^{\text{TAL}}(T) = O\left(\frac{K\Delta_{\max}\log(T)}{\Delta_{\min}^2} + \sum_{m\in[M]} \frac{K\Delta_{\max}\log(T)}{\mu_{\dagger,m}^2}\right),$$

$$C_F^{\text{TAL}}(T) = O\left(\sum_{m\in[M]} \left[\frac{K\mu_{*,m}\log(T)}{M\Delta_{\min}^2} + \frac{K\mu_{*,m}\log(T)}{\mu_{\dagger,m}^2}\right]\right).$$

The two parts in regret and cost are from the learning and teaching phases, respectively. It can be observed that TAL with $\varepsilon$-greedy clients also achieves the goal illustrated in Section IV. Moreover, compared with Theorem 8, dependencies on $\Delta_{\max}$ and $\mu_{*,m}$ (instead of $\Delta_k$ and $\mu_{k,m}$) can be observed, which is a worst-case consideration to capture the random actions generated from the $\varepsilon$-greedy policy.

### C. Discussions: Thompson Sampling and Beyond

Another popular bandit strategy is Thompson sampling (TS) [20]. Experiment results in Sec. VIII verify the performance of TAL with TS clients; however, the theoretical analysis remains open. In particular, unlike the sufficiently exploring UCB and $\varepsilon$-greedy, [59] indicates that when facing two arms with constant reward 1's, the pulls by TS can be arbitrarily imbalanced. Instead, balanced pulls can be achieved with reward 0's for these two arms. This phenomenon motivates using $\gamma_1 = 0$ to encourage TS explorations in the learning phase, whose effectiveness is verified empirically but not analytically. On the other hand, the complicated warm-start behavior of TS also requires further investigation.

Furthermore, in Secs. VI-A and VI-B, the hyper-parameter $\gamma_1$ is set to different values (i.e., 1 for UCB clients and 0 for $\varepsilon$-greedy clients). These choices are made to facilitate the corresponding "warm-start" analyses required in Definition 3 (i.e., to maintain the optimism of estimations in UCB and to avoid complicated analyses due to the randomness in $\varepsilon$-greedy). However, the capabilities of TAL extend beyond these theoretically sound options. Especially, experiments in Sec. VIII show that various other choices (e.g., $\gamma_1 = 0$ for UCB clients and $\gamma_1 = 1$ for $\varepsilon$-greedy clients) can also lead to reasonable performances. Thus, it would be an interesting future direction to investigate whether a unified hyper-parameter $\gamma_1$ in TAL is sufficient for certain classes of client strategies (e.g., UCB and $\varepsilon$-greedy). The main difficulty along this direction is still to analyze the "warm-start" behaviors, which are largely determined by the specific strategy.

Moreover, Thm. 5 has established conditions on clients' strategies to obtain performance guarantees of TAL, i.e., sufficiently exploring and low sub-optimal warm-start pulls. An interesting direction is to verify the client-strategy-agnostic nature of TAL in an even broad sense, e.g., with any no-regret client strategy. Experimental results are provided later to enlighten future works on this open problem.

**Algorithm 2** TWL

**Input:** Parameter $\gamma_1, \gamma_2 \in [0, 1]$; Time Horizon $T$

Initialize: active arm set $\Upsilon \leftarrow [K]$; iteration counter $\psi \leftarrow 1$
**for** $t \leq T$ **do**
  Observe $\{\pi_m(t), X_{\pi_m(t),m}(t) : m \in [M]\}$
  **if** $|\Upsilon| > 1$ and $N_{k,m}(t) \geq F(\psi), \forall k \in \Upsilon, m \in [M]$ **then**
    Update $\{\text{UCB}_k(\psi), \text{LCB}_k(\psi) : k \in \Upsilon\}$ as in Eqn. (1)
    Update $\Upsilon \leftarrow \{j \in \Upsilon : \text{UCB}_j(\psi) \geq \text{LCB}_k(\psi), \forall k \in \Upsilon\}$
    Set $\psi \leftarrow \psi + 1$
  **end if**
  $\forall m \in [M]$, set

$$\sigma_m(t) \leftarrow \begin{cases} \gamma_2 - X_{\pi_m(t),m}(t) & \text{if } \pi_m(t) \notin \Upsilon \\ \gamma_1 - X_{\pi_m(t),m}(t) & \text{if } \pi_m(t) \in \Upsilon \text{ and } |\Upsilon| > 1 \\ 0 & \text{if } \pi_m(t) \in \Upsilon \text{ and } |\Upsilon| = 1 \end{cases},$$

  Set $X'_{\pi_m(t),m}(t) \leftarrow X_{\pi_m(t),m}(t) + \sigma_m(t)$
  Reveal $X'_{\pi_m(t),m}(t)$ to each client $m \in [M]$
**end for**

## VII. TWL: A More Adaptive Extension

### A. Algorithm Design

To further optimize the performance, a more adaptive "Teaching-While-Learning" (TWL) algorithm (presented in Alg. 2) is proposed, which breaks the non-adaptive phased structure of TAL by leveraging a different idea of *successive arm elimination* [60], [61]. In TWL, the server maintains a set $\Upsilon$ of active arms (on the global model), which is initialized as $[K]$. If $|\Upsilon| > 1$, the following update is performed after each active arm $k \in \Upsilon$ has been pulled at least $F(\psi)$ times by each client:

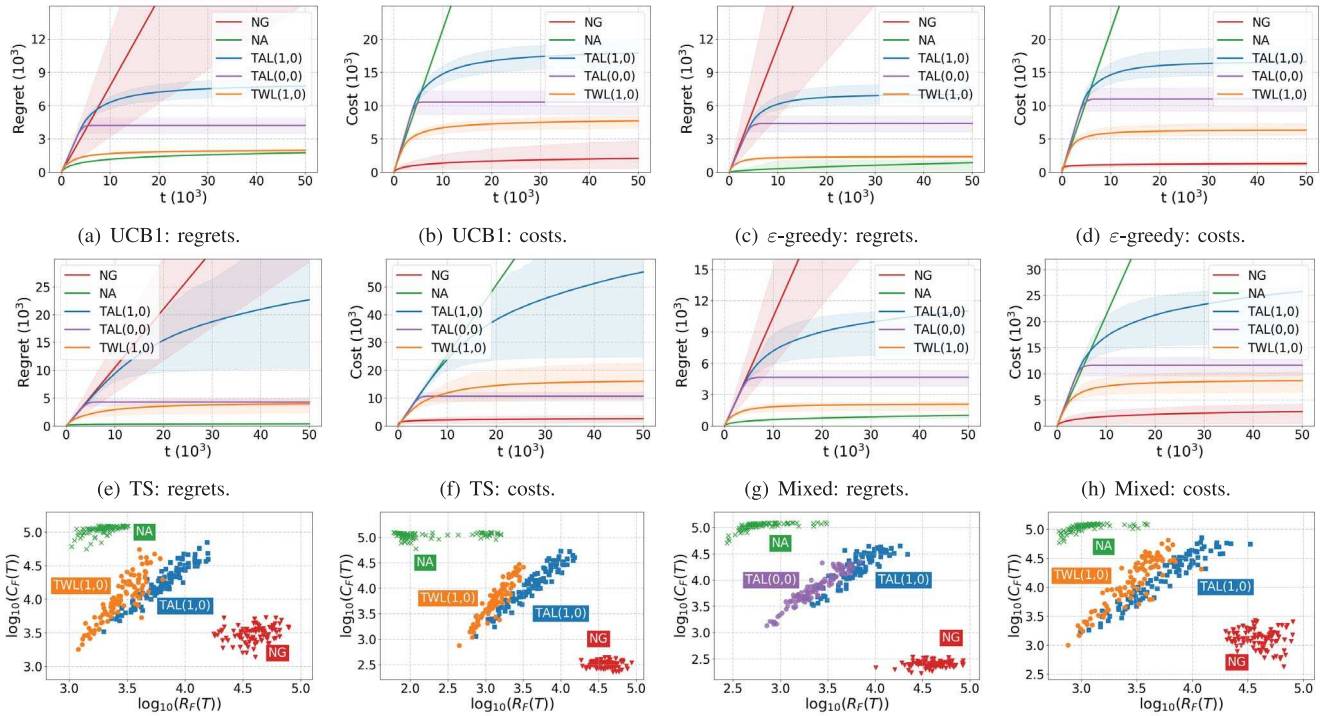
$$\Upsilon \leftarrow \{j \in \Upsilon : \text{UCB}_j(\psi) \geq \text{LCB}_k(\psi), \forall k \in \Upsilon\},$$

where $\text{UCB}_k(\psi)$ and $\text{LCB}_k(\psi)$ are defined in Eqn. (1) and $\psi$ is the epoch counter as in TAL. In this process, the arms that do not satisfy the requirement are eliminated (i.e., marked as inactive). Then, based on the set $\Upsilon$, the following adjustment is performed for client $m$:

$$\sigma_m(t) \leftarrow \begin{cases} \gamma_2 - X_{\pi_m(t),m}(t) & \text{if } \pi_m(t) \notin \Upsilon \\ \gamma_1 - X_{\pi_m(t),m}(t) & \text{if } \pi_m(t) \in \Upsilon \text{ and } |\Upsilon| > 1 \\ 0 & \text{if } \pi_m(t) \in \Upsilon \text{ and } |\Upsilon| = 1 \end{cases},$$

where $\gamma_1, \gamma_2 \in [0, 1]$ are to-be-specified input parameters.

In other words, the local rewards of all inactive arms are adjusted to $\gamma_2$ (typically small) to discourage explorations. For an active arm, when there are other active arms (i.e., $|\Upsilon| > 1$), the server uniformly adjusts its rewards to $\gamma_1$ to encourage explorations. When an arm is the only active one (which is arm $k_\dagger$ with high probability), its original rewards are kept to save server adjustments, which is sufficient as all other arms are inactive with a small perceived reward $\gamma_2$.




(a) UCB1: regrets. (b) UCB1: costs. (c) $\varepsilon$-greedy: regrets. (d) $\varepsilon$-greedy: costs.

(e) TS: regrets. (f) TS: costs. (g) Mixed: regrets. (h) Mixed: costs.

(i) UCB1: random insts. (j) $\varepsilon$-greedy: random insts. (k) TS: random insts. (l) Mixed: random insts.

Fig. 2. Experimental results on synthetic datasets with clients running UCB1, $\varepsilon$-greedy, TS, and mixed strategies. Evaluations of (a–h) are under a fixed 5-clients-5-arms instance, where the curves represent the empirically averaged values and the shadowed areas represent the upper and lower 80% confidence intervals. Evaluations of (i–l) are with 100 randomly generated 5-clients-5-arms instances, where each dot reports the performance (in a log-log scale) under one instance and plots of a few algorithms are omitted for a better presentation here. The mixed strategies are two UCB1, two $\varepsilon$-greedy, and one TS. All time horizons are $T = 50000$.

TWL is more refined than TAL as it only encourages explorations on the active arms (instead of all arms), which is important in two aspects. First, only necessary arm-dependent explorations are encouraged. Second, fewer cumulative rewards on the sub-optimal arms also alleviate the server's burden of teaching clients to converge to the optimal arm.

### B. Theoretical Analysis

The general performance of TWL can be similarly analyzed as that of TAL in Sec. VI. The following result establishes the performance guarantee for UCB1 clients.

*Theorem 12:* (TWL with UCB1 clients). For TWL with $\gamma_1 = 1$ and $\gamma_2 = 0$, if all clients run UCB1 locally and $\mu_{\dagger,m} \neq 0$ for all $m \neq [M]$, it holds that

$$R_F^{\mathrm{TWL}}(T) = O\left( \sum_{m\in[M]} \sum_{k\neq k_\dagger} \left[ \frac{\log(T)}{\mu_{\dagger,m} M \Delta_k} + \frac{\Delta_k \log(T)}{\mu_{\dagger,m}^2} \right] \right),$$

$$C_F^{\mathrm{TWL}}(T) = O\left( \sum_{m\in[M]} \sum_{k\in[K]} \frac{(1-\mu_{k,m})\log(T)}{M\Delta_k^2} + \sum_{m\in[M]} \sum_{k\neq k_\dagger} \left[ \frac{\mu_{k,m}\log(T)}{\mu_{\dagger,m} M \Delta_k^2} + \frac{\mu_{k,m}\log(T)}{\mu_{\dagger,m}^2} \right] \right).$$

The proof can be found in Appendix B. The above guarantees can be interpreted in similar ways as those of TAL in Thm. 8, i.e., one part from learning the global optimal arm and the other part from guiding agents towards it. More importantly, it is noted that with UCB1 clients, TWL strictly outperforms TAL w.r.t. both criteria since the dependency on the minimum gap $\Delta_{\min}$ is replaced by arm-dependent gaps $\Delta_k \geq \Delta_{\min}$, which comes precisely from its adaptive design.

*Remark 2:* For $\varepsilon$-greedy clients, with $\gamma_1 = \gamma_2 = 0$, the same performance guarantee as Thm. 11 can be established for TWL because the active and non-active arms are not distinctly treated under this specification, which degrades TWL to TAL. However, experimental results show that better empirical performance is achieved with $\gamma_1 = 1$ and $\gamma_2 = 0$, whose theoretical analyses are left open for future works.

*Remark 3:* While TWL improves the regret of TAL regarding the dependency on $\Delta_k$, it is unclear whether its dependencies on parameters other than $M$ and $T$ are tight. One the one hand, as mentioned in Remark 1, a refined lower bound would be instructive in evaluating such tightness. On the other hand, it is equally worth exploring whether a refined upper bound can be obtained, which is left for further investigations.

## VIII. Experimental Results

In this section, the proposed algorithms are empirically evaluated against two baselines, NG and NA from Sec. IV, to demonstrate their superiority and generality.

### A. Synthetic Dataset

First, experimental results with synthetic datasets are reported in Fig. 2. In particular, two sets of experiments are

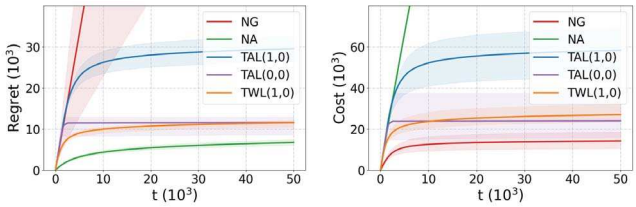
performed: (1) the first environment is a fixed instance with $M = 5$ clients and $K = 5$ arms, where each client's local model is specified (left to right: arm 1 to arm 5) with the following mean rewards: Client 1–$[0.2, 0.9, 0.1, 0.8, 0.6]$, Client 2–[0.4, 0.1, 0.9, 0.4, 0.8], Client 3–[0.2, 0.2, 0.5, 0.5, 0.9], Client 4–[0.4, 0.3, 0.8, 0.9, 0.4], Client 5–[0.3, 0.5, 0.2, 0.4, 0.8]. The corresponding global game then has the following mean rewards with a gap of $\Delta_{\min} = 0.1$ (left to right: arm 1 to arm 5): [0.3, 0.4, 0.5, 0.6, 0.7]; (2) the second setting is 100 randomly generated instances with $M = 5$ clients and $K = 5$ arms. Especially, the mean reward of each local arm for each client is sampled from a uniform distribution in $[0, 1]$. The obtained results from these two sets of environments are reported with different client strategies in Fig. 2(a)–2(h) and Fig. 2(i)–2(l), respectively, and discussed in the following. To facilitate presentations, we denote TAL$(\gamma_1, \gamma_2)$ (resp. TWL$(\gamma_1, \gamma_2)$) as TAL (resp. TWL) with specific parameters $\gamma_1$ and $\gamma_2$. We note that with the randomly generated instances in the second environment, the reported observations are sufficiently general.

**UCB1 clients.** First, with UCB1 clients, from Fig. 2(a) and 2(b), it can be observed that the proposed algorithms are capable of converging while the superiority of TWL over TAL is verified. However, as claimed in Sec. IV, the baselines are at two extremes: NG (resp. NA) is almost linear in regret (resp. cost) although performing well w.r.t. cost (resp. regret). Fig. 2(i) further demonstrates that TAL and TWL strike a balance between regret and cost, while the advantage of TWL is evident again. In particular, their performance scatter plots from 100 randomly generated instances are concentrated in the diagonal between the two axes. However, the plots of the two baselines are near one axis but far from the other.

***ε*-greedy clients.** Fig. 2(c) and 2(d) report that TAL and TWL can successfully teach $\varepsilon$-greedy clients with a reasonably low regret and cost at the same time. Somewhat unexpectedly, TWL$(1, 0)$ has a better performance even over the theoretically sound TAL$(0, 0)$ (equivalently, TWL$(0, 0)$), which warrants further investigations with $\varepsilon$-greedy clients. Fig. 2(j) verifies that the above observations hold in general.

**TS clients.** Although not theoretically studied, Fig. 2(e), 2(f) and 2(k) report the performances of the proposed algorithms with TS clients. While converging, the performance of TAL$(1, 0)$ and TWL$(1, 0)$ are highly unstable, which verifies the imbalanced exploration of TS discussed in Sec. VI-C. On the other hand, TAL$(0, 0)$ has stable and competitive behaviors.

**Mixed clients.** Beyond one single local strategy, TAL and TWL are also tested with mixed strategies for clients. Especially, with two UCB1 clients, two $\varepsilon$-greedy clients, and one TS client, the results are reported in Fig. 2(g), 2(h) and 2(l). It can be observed that the proposed designs are capable of effectively guiding the clients to the global optimal arm in the face of mixed client strategies while achieving a good balance between regret and cost. These results further demonstrate the broad applicability of the designs and their appealing property of being client-strategy-agnostic.




(a) Mixed: regrets. (b) Mixed: costs.

Fig. 3. Experimental results on the real-world MovieLens dataset with clients running mixed strategies. Evaluations (a) and (b) are under a fixed 15-clients-15-arms instance, which is extracted by grouping users and movies in the MovieLens dataset. The curves represent the empirically averaged values and the shadowed areas represent the upper and lower 80% confidence intervals.

### B. Real-World Dataset

To further complement the observations obtained from synthetic datasets, the empirical performances of the proposed designs are further evaluated on the MovieLens dataset [62]. The available users and movies in the dataset are both randomly divided into 15 groups to form an FMAB environment with 15 clients and 15 arms. The average movie ratings from each group of users are used to construct their local rewards. Also, the clients are considered to adopt mixed bandit strategies: 5 clients using each choice of UCB1, $\varepsilon$, and TS, respectively. From the results reported in Fig. 3, the aforementioned key observation is further verified that the proposed designs, i.e., TAL and TWL, are capable of effectively guiding the clients towards the global optimal arm with a reasonable amount of adjustment cost, i.e., balancing between regret and cost. These results further demonstrate the practicability of the proposed designs.

## IX. Conclusion

A novel idea of reward teaching was proposed to have the server guide autonomous clients in an unknown FMAB environment via reward adjustments, which avoids any changes to the clients' protocols and removes the previous requirement of naive clients in FMAB. Two client-strategy-agnostic algorithms, TAL and TWL, were proposed. The TAL algorithm was designed with two phases to separately encourage and discourage explorations. The TWL algorithm further optimized the performance by breaking the non-adaptive phased structure into a flexible interleaving scheme. General performance analysis was established for TAL when the clients' strategies satisfy certain requirements. Especially, for the representative UCB1 and $\varepsilon$-greedy clients, rigorous analyses showed that TAL strikes a balance between regret and adjustment cost (logarithmic in both metrics), which is order-optimal w.r.t. the natural lower bound. Moreover, the analyses also demonstrated that TWL achieves an improved dependency on the sub-optimality gap than TAL due to its adaptive design. Experimental results further demonstrated the effectiveness and efficiency of the proposed algorithms. Under the reward teaching framework, many interesting questions were left open for further investigations, e.g., theoretical analysis on TAL and TWL with Thompson sampling clients.

# APPENDIX A
# TAL: PERFORMANCE ANALYSIS

## A. General Analysis: Theorem 5

First, the following good event is established to demonstrate the effectiveness of the proposed confidence bounds.

*Lemma 13:* Denoting event $\mathcal{E}_F$ as

$$\mathcal{E}_F := \left\{\forall \psi \le T, \forall k \in [K], |\hat{\nu}_k(\psi) - \nu_k| \le 2^{-\psi-2}\right\}$$

where $\hat{\nu}_k(\psi) := \frac{1}{M}\sum_{m=1}^{M} \hat{\mu}_{k,m}(\psi)$, it holds that $\mathbb{P}(\mathcal{E}_F) \ge 1 - 1/T$.

*Proof:* With Hoeffding's inequality and the design that

$$\hat{\nu}_k(\psi) = \sum_{m=1}^{M} \sum_{\tau=F(\psi-1)+1}^{F(\psi)} X_{k,m}(N_{k,m}^{-1}(\tau))/(Mf(\psi)),$$

at epoch $\psi$, for arm $k$, we have

$$\begin{aligned}\mathbb{P}\left(|\hat{\nu}_k(\psi) - \nu_k| > 2^{-\psi-2}\right) &\le 2\exp(-2Mf(\psi)2^{-2\psi-4}) \\ &= 1/(KT^2).\end{aligned}$$

With a union bound over $\psi \le T$ and $k \in [K]$, the lemma can be proved. □

*Lemma 14:* (Learning Phase in TAL; Restatement of Lemma 2). If $\Pi_m$ is $([K], \gamma_1, \underline{\eta}_m, \overline{\eta}_m)$-sufficiently exploring for all $m \in [M]$, with probability (w.p.) at least $1 - 1/T$, the learning phase ends with $k_\ddagger = k_\dagger$ by time step $T_1$, and the regret and cost in the learning phase of TAL are bounded, respectively, as

$$\begin{aligned}R_{F,1}(T) &\le \sum_{m\in[M]} \sum_{k\ne k_\dagger} \Delta_k \cdot \overline{\eta}_m(T_1; \gamma_1, [K]); \\ C_{F,1}(T) &\le \sum_{m\in[M]} \sum_{k\in[K]} \delta_{k,m}(\gamma_1) \cdot \overline{\eta}_m(T_1; \gamma_1, [K]),\end{aligned}$$

where $T_1 \le \max_{m\in[M]} \left\{\underline{\eta}_m^{-1}(F(\psi_{\max}); \gamma_1, [K])\right\}$.

*Proof of Lemma 14:* With event $\mathcal{E}_F$ in Lemma 13 happening, we assume the learning phases end at time step $T_1$ such that

$$T_1 \ge \max_{m\in[M]} \left\{\underline{\eta}_m^{-1}(F(\psi_{\max}); \gamma_1, [K])\right\}.$$

Since each local algorithm $\Pi_m$ is $([K], \gamma_1, \underline{\eta}_m, \overline{\eta}_m)$-sufficiently exploring and the rewards on all arms are constant $\gamma_1$'s, it holds that $N_{k,m}(T_1) \ge F(\psi_{\max}), \forall k \in [K], \forall m \in [M]$, which means that epoch $\psi_{\max}$ is reached. Thus, the confidence bound can be bounded as $\mathrm{CB}(\psi_{\max}) \le \frac{1}{4} \cdot 2^{-\psi_{\max}} \le \frac{1}{4}\Delta_{\min}$, which results in

$$\begin{aligned}\mathrm{LCB}_\dagger(\psi_{\max}) &= \hat{\nu}_\dagger(\psi_{\max}) - \mathrm{CB}(\psi_{\max}) \\ &\ge \nu_\dagger - 2\mathrm{CB}(\psi_{\max}) \ge \nu_\dagger - \frac{\Delta_{\min}}{2} \\ &\ge \nu_k + \frac{\Delta_{\min}}{2} \ge \nu_k + 2\mathrm{CB}(\psi_{\max}) \\ &\ge \hat{\nu}_k(\psi_{\max}) + \mathrm{CB}(\psi_{\max}) \\ &= \mathrm{UCB}_k(\psi_{\max}), \qquad \forall k \ne k_\dagger.\end{aligned}$$

Thus, the learning phase should already end. Similarly, it can be obtained that arm $k_\dagger$ would not be dominated by any other arm; thus $k_\ddagger = k_\dagger$. Then, with the observation that

$$N_{k,m}(T_1) \le \overline{\eta}_m(T_1; \gamma_1, [K]), \qquad \forall k \in [K], m \in [M],$$

the lemma can be proved. □

*Lemma 15:* (Teaching Phase in TAL; Restatement of Lemma 4). If the event in Lemma 14 occurs, the regret and cost in the teaching phase of TAL are bounded, respectively, as

$$\begin{aligned}R_{F,2}(T) &\le \sum_{m\in[M]} \max_{H_m\in\mathcal{H}_m} \sum_{k\ne k_\dagger} \Delta_k \cdot \iota_k(T; H_m, \mathcal{B}_m, \Pi_m); \\ C_{F,2}(T) &\le \sum_{m\in[M]} \max_{H_m\in\mathcal{H}_m} \sum_{k\ne k_\dagger} \delta_{k,m}(\gamma_2) \cdot \iota_k(T; H_m, \mathcal{B}_m, \Pi_m),\end{aligned}$$

where $\mathcal{B}_m$ denotes an environment with constant rewards as $\gamma_2$ for arm $k \ne k_\dagger$ and stochastic rewards with expectation $\mu_{\dagger,m}$ for arm $k_\dagger$. The set $\mathcal{H}_m$ is defined with each element of it as a reward sequence $H_m = \{H_{k,m} : k \in [K]\}$ where $H_{k,m} \in \{\{\gamma_1\}^\tau : \tau \in [\underline{\eta}_m(T_1; \gamma_1, [K]), \overline{\eta}_m(T_1; \gamma_1, [K])]\}$.

*Proof of Lemma 15:* This lemma can be obtained by realizing that if the event in Lemma 2 happens, at the beginning of the teaching phase, i.e., time step $T_1$, client $m$ has observed constant reward $\gamma_1$ on each arm $k \in [K]$ for at least $\underline{\eta}_m(T_1; \gamma_1, [K])$ and at most $\overline{\eta}_m(T_1; \gamma_1, [K])$ times, which leads to the definition of $\mathcal{H}_m$.

Starting at time step $T_1$, the local bandit algorithm $\Pi_m$ can be viewed as interacting with environment $\mathcal{B}_m$ with prior input $H_m \in \mathcal{H}_m$. By recognizing that with the reward sequence $H_m \in \mathcal{H}_m$, $\mathbb{E}[N_{k,m}(T - T_1)|H_m, \mathcal{B}_m] \le \mathbb{E}[N_{k,m}(T)|H_m, \mathcal{B}_m] = \iota_{k,m}(T; H_m, \mathcal{B}_m, \Pi_m)$, the lemma can be proved. □

*Theorem 16:* (Overall Performance of TAL; Restatement of Theorem 5). Under the assumption in Lemma 14, with $R_{F,1}(T), C_{F,1}(T)$ defined in Lemma 14 and $R_{F,2}(T), C_{F,2}(T)$ in Lemma 15, the regret and cost of TAL are bounded, respectively, as $R_F(T) \le R_{F,1}(T) + R_{F,2}(T) + O(M)$ and $C_F(T) \le C_{F,1}(T) + C_{F,2}(T) + O(M)$.

*Proof of Theorem 5:* When event $\mathcal{E}_F$ happens, the regret and cost can be obtained as the combination of Lemmas 14 and 15. Otherwise, the regret and cost can be bounded linearly by $MT$. The lemma can then be proved with the guarantee that $\mathbb{P}(\mathcal{E}_F) \ge 1 - 1/T$ as shown in Lemma 14. □

## B. UCB1 Clients: Theorem 8

*Lemma 17:* (Restatement of Lemma 6). For any $\gamma \in [0, 1]$ and set $\mathcal{I} \subseteq [K]$, UCB1 is $(\mathcal{I}, \gamma, \underline{\eta}, \overline{\eta})$-sufficiently exploring with $\underline{\eta}(\tau; \gamma, \mathcal{I}) = \lfloor \tau/|\mathcal{I}| \rfloor$ and $\overline{\eta}(\tau; \gamma, \mathcal{I}) = \lceil \tau/|\mathcal{I}| \rceil$.

*Proof:* The UCB1 algorithm is defined in Sec. VI-A and the subscript $m$ is ignored in the following to denote a general UCB1 algorithm. To prove the lemma, it is essential to obtain that if at time step $t$, $\sum_{k\in\mathcal{I}} N_k(t) = \tau$, then $\max_{k\in\mathcal{I}} N_k(t) - \min_{k\in\mathcal{I}} N_k(t) \le 1$. If this claim does not hold, there exist arms $k, k'$ such that $N_k(t) \ge N_{k'}(t) + 2$. Then, at the last time step that the arm $k$ is pulled, denoted as $t'$, it holds that

$$\begin{aligned}\mu_k'(t') + \sqrt{\frac{2\log(t')}{N_k(t')}} &\overset{(a)}{\le} \gamma + \sqrt{\frac{2\log(t')}{N_{k'}(t)+1}} \le \gamma + \sqrt{\frac{2\log(t')}{N_{k'}(t)}} \\ &\le \gamma + \sqrt{\frac{2\log(t')}{N_{k'}(t')}} \overset{(b)}{=} \mu_{k'}'(t') + \sqrt{\frac{2\log(t')}{N_{k'}(t')}},\end{aligned}$$

where steps (a) and (b) leverages the fact that both arm $k$ and $k'$ receive reward $\gamma$'s. A contradiction is thus raised as arm $k$ would not be pulled, and the lemma is proved. □

Then, with Lemma 17, we can observe that

$$T_1 \leq KF(\psi_{\max}) = O\left(\frac{K\log(KT)}{M\Delta_{\min}^2}\right). \tag{3}$$

*Lemma 18:* (Restatement of Lemma 7). If $\gamma_1 \geq \mu_{\dagger,m} > \gamma_2$ and $\Pi_m$ is UCB1, for all $k \neq k_\dagger$, it holds that $\max_{H_m \in \mathcal{H}_m}\{\iota_k(T; H_m, \mathcal{B}_m, \Pi_m)\} = O\left(\frac{(\gamma_1-\gamma_2)T_1}{K(\mu_{\dagger,m}-\gamma_2)} + \frac{\log(KT)}{(\mu_{\dagger,m}-\gamma_2)^2}\right)$.

*Proof of Lemma 18:* For $H_m$ in the set $\mathcal{H}_m$, it contains $\tau_{k,m}$ times reward $\gamma_1$ on each arm $k \in [K]$, where $\tau_{k,m} \in [\lfloor T_1/K \rfloor, \lceil T_1/K \rceil]$. We denote $t_H = \sum_{k\in[K]} \tau_{k,m} \leq T_1 + K$ as the length of reward sequence in $H_m$ and

$$\begin{aligned} L_{k,m} &:= \frac{(\gamma_1-\gamma_2)\tau_{k,m}}{4(\mu-\gamma_2)^2} + \frac{2\log(T+t_H)}{4(\mu-\gamma_2)^2} \\ &\leq \frac{(\gamma_1-\gamma_2)(T_1/K+1)}{4(\mu-\gamma_2)^2} + \frac{2\log(T+T_1+K)}{4(\mu-\gamma_2)^2}. \end{aligned}$$

It holds that

$$\begin{aligned} &\iota_{k,m}(T; H_m, \mathcal{B}_m, \Pi_m) \leq L_{k,m}(T) \\ &\quad + \mathbb{E}\left[\sum_{t\in[T]} \mathbb{1}\left\{\pi_m(t) = k, N_{k,m}(t-1) > L_{k,m}(T)\right\}\right] \\ &= L_{k,m}(T) + \sum_{t\in[T]} \mathbb{P}(\pi_m(t) \\ &= k, N_{k,m}(t-1) > L_{k,m}(T)). \end{aligned}$$

With UCB1 as $\Pi_m$, it further holds that

$$\begin{aligned} &\mathbb{P}\left(\pi_m(t) = k, N_{k,m}(t-1) > L_{k,m}(T)\right) \\ &\leq \mathbb{P}\left(\hat{\mu}'_{k,m}(t-1) + \sqrt{\frac{2\log(t+t_H)}{\tau_{k,m}+N_{k,m}(t-1)}} \geq \hat{\mu}'_{\dagger,m}(t-1)\right. \\ &\quad \left. + \sqrt{\frac{2\log(t+t_H)}{\tau_{\dagger,m}+N_{\dagger,m}(t-1)}}, N_{k,m}(t-1) > L_{k,m}(T)\right) \end{aligned}$$

where the last inequality is due to a union bound.

Let us separately consider $N_{k,m}(t-1) = n_{k,m} \in [L_{k,m}(T), t]$ and $N_{\dagger,m}(t-1) = n_{k,m} \in [t]$. It holds that

$$\begin{aligned} &\mathbb{P}\left(\hat{\mu}'_{\dagger,m}(t-1) + \sqrt{\frac{2\log(t+t_H)}{\tau_{\dagger,m}+n_{\dagger,m}}} \leq \mu_{\dagger,m}\right) \\ &= \mathbb{P}\left(\frac{\tau_{\dagger,m}\gamma_1 + \sum_{\tau=1}^{n_{\dagger,m}} X^\tau_{\dagger,m}}{\tau_{\dagger,m}+n_{\dagger,m}} + \sqrt{\frac{2\log(t+t_H)}{\tau_{\dagger,m}+n_{\dagger,m}}} \leq \mu_{\dagger,m}\right) \\ &\overset{(a)}{\leq} \mathbb{P}\left(\frac{\tau_{\dagger,m}\mu_{\dagger,m} + \sum_{\tau=1}^{n_{\dagger,m}} X^\tau_{\dagger,m}}{\tau_{\dagger,m}+n_{\dagger,m}} + \sqrt{\frac{2\log(t+t_H)}{\tau_{\dagger,m}+n_{\dagger,m}}} \leq \mu_{\dagger,m}\right) \\ &\overset{(b)}{\leq} \frac{1}{(t+t_H)^4}, \end{aligned}$$

where inequality (a) is an essential step of "optimism" due to $\gamma_1 \geq \mu_{\dagger,m}$ and inequality (b) holds is from Hoeffding's inequality. Also, it can be observed that with $n_{k,m} \geq L_{k,m}(T)$,

$$\begin{aligned} &\hat{\mu}'_{k,m}(t-1) + \sqrt{\frac{2\log(t+t_H)}{\tau_{k,m}+n_{k,m}}} \\ &= \frac{\tau_{k,m}\cdot\gamma_1 + n_{k,m}\cdot\gamma_2}{\tau_{k,m}+n_{k,m}} + \sqrt{\frac{2\log(t+t_H)}{\tau_{k,m}+n_{k,m}}} \leq \mu_{\dagger,m}. \end{aligned}$$

Thus, with a union bound, it holds that

$$\begin{aligned} &\mathbb{P}\left(\pi_m(t) = k, N_{k,m}(t-1) > L_{k,m}(T)\right) \\ &\leq \sum_{n_{k,m}=L_{k,m}(T)}^{t} \sum_{n_{\dagger,m}\in[t]} \frac{1}{(t+t_H)^4} \leq \frac{1}{t^2} \end{aligned}$$

It is then indicated that

$$\mathbb{E}\left[N_{k,m}(T)\right] \leq L_{k,m}(T) + \sum_{t=1}^{T} \frac{1}{t^2} \leq L_{k,m}(T) + 2,$$

which proves Lemma 18. □

*Theorem 19:* (TAL with UCB1 clients; Restatement of Theorem 8). For TAL with $\gamma_1 = 1$ and $\gamma_2 = 0$, if all clients run UCB1 locally and $\mu_{\dagger,m} \neq 0$ for all $m \in [M]$, it holds that

$$R_F(T) = O\left(\sum_{m\in[M]}\sum_{k\neq k_\dagger}\left[\frac{\Delta_k\log(KT)}{\mu_{\dagger,m}M\Delta_{\min}^2} + \frac{\Delta_k\log(KT)}{\mu_{\dagger,m}^2}\right]\right);$$

$$\begin{aligned} C_F(T) = O\Bigg(&\sum_{m\in[M]}\sum_{k\in[K]}\frac{(1-\mu_{k,m})\log(KT)}{M\Delta_{\min}^2} \\ &+ \sum_{m\in[M]}\sum_{k\neq k_\dagger}\left[\frac{\mu_{k,m}\log(KT)}{\mu_{\dagger,m}M\Delta_{\min}^2} + \frac{\mu_{k,m}\log(KT)}{\mu_{\dagger,m}^2}\right]\Bigg). \end{aligned}$$

*Proof of Theorem 19:* From Eqn. (3), it holds that $T_1 = O\left(\frac{K\log(KT)}{M\Delta_{\min}^2}\right)$. With Lemmas 14 and 17, it holds that

$$\begin{aligned} R_{F,1}(T) &\leq \sum_{m\in[M]}\sum_{k\neq k_\dagger}\Delta_k\cdot\overline{\eta}_m\left(T_1;\gamma_1,[K]\right) \\ &= O\left(\sum_{m\in[M]}\sum_{k\neq k_\dagger}\frac{\Delta_k\log(KT)}{M\Delta_{\min}^2}\right); \\ C_{F,1}(T) &\leq \sum_{m\in[M]}\sum_{k\in[K]}\delta_{k,m}(\gamma_1)\cdot\overline{\eta}_m\left(T_1;\gamma_1,[K]\right) \\ &\overset{(a)}{=} O\left(\sum_{m\in[M]}\sum_{k\in[K]}\frac{(1-\mu_{k,m})\log(KT)}{M\Delta_{\min}^2}\right), \end{aligned}$$

where equation (a) also utilizes that with $\gamma_1 = 1$, $\delta_{k,m}(\gamma_1) = 1-\mu_{k,m}$.

Then, with Lemmas 15 and 18, it holds that

$$\begin{aligned} R_{F,2}(T) &\leq \sum_{m\in[M]}\max_{H_m\in\mathcal{H}_m}\sum_{k\neq k_\dagger}\Delta_k\cdot\iota_k(T;H_m,\mathcal{B}_m,\Pi_m) \\ &= O\left(\sum_{m\in[M]}\sum_{k\neq k_\dagger}\frac{\Delta_k\log(KMT)}{\mu_{\dagger,m}M\Delta_{\min}^2} + \frac{\Delta_k\log(KMT)}{\mu_{\dagger,m}^2}\right); \\ C_{F,2}(T) &\leq \sum_{m\in[M]}\max_{H_m\in\mathcal{H}_m}\sum_{k\neq k_\dagger}\delta_{k,m}(\gamma_2)\cdot\iota_k(T;H_m,\mathcal{B}_m,\Pi_m) \\ &\overset{(a)}{=} O\left(\sum_{m\in[M]}\sum_{k\neq k_\dagger}\frac{\mu_{k,m}\log(KT)}{\mu_{\dagger,m}M\Delta_{\min}^2} + \frac{\mu_{k,m}\log(KT)}{\mu_{\dagger,m}^2}\right), \end{aligned}$$

where equation (a) uses the fact that with $\gamma_2 = 0$, $\delta_{k,m}(\gamma_2) = \mu_{k,m}$. The theorem is then proved. □

### C. $\varepsilon$-Greedy Clients: Theorem 11

*Lemma 20:* For any $\gamma \in [0,1]$, if ties among arms are broken uniformly at random, with probability at least $1-1/T$, $\varepsilon$-greedy is $([K],\gamma,\underline{\eta},\overline{\eta})$-sufficiently exploring with $\underline{\eta}(\tau;\gamma,[K])$ and $\overline{\eta}(\tau;\gamma,[K]) = O(\tau/K \pm \log(KT))$.

*Proof of Lemma 20:* Since the rewards on each arm are all $\gamma$ (thus the sample means are all $\gamma$) and the ties among arms are broken uniformly at random, the algorithm would pull each arm $k \in [K]$ with equal probability $1/K$. Thus, denoting the number of pulls on an arbitrary arm $k$ by time $\tau$ as $N_k(\tau)$, it holds that $N_k(\tau) = \sum_{t\in[\tau]} \mathbb{1}\{\pi(t) = k\}$. Using Bernstein's inequality and $\mathbb{P}(\pi(t)=k) = 1/K$, we can obtain that

$$\mathbb{P}\left(\left|N_k(\tau) - \frac{\tau}{K}\right| \geq x\right) \leq 2\exp\left(-\frac{x^2}{2\frac{\tau}{K} + \frac{2}{3}x}\right) \leq \frac{1}{KT},$$

where $x := \frac{\tau}{4K} + \frac{8}{3}\log(KT)$.

With a union bound over $k \in [K]$, we can obtain that $\underline{\eta}(\tau;\gamma,[K]) = \frac{3\tau}{4K} - \frac{8}{3}\log(KT)$ and $\overline{\eta}(\tau;\gamma,[K]) = \frac{5\tau}{4K} + \frac{8}{3}\log(KT)$, which concludes the proof. □

Using Lemmas 14 and 20, we can obtain that with probability $1-1/T$, the learning phase of TAL ends at $T_1$ such that

$$\begin{aligned} T_1 &\leq \max_{m\in[M]} \left\{\underline{\eta}_m^{-1}\left(F(\psi_{\max});\gamma_1,[K]\right)\right\} \\ &= \frac{4K}{3}F(\psi_{\max}) + \frac{32K}{9}\log(KMT) \\ &= O\left(\frac{K\log(KT)}{M\Delta_{\min}^2} + K\log(KMT)\right). \end{aligned} \tag{4}$$

*Lemma 21:* If $\Pi_m$ is $\varepsilon$-greedy and $\mu_{\dagger,m} > \gamma_1 = \gamma_2 = 0$, with probability at least $1-1/T$, it holds that $\max_{H_m\in\mathcal{H}_m}\left\{\sum_{k\neq k_\dagger} \iota_{k,m}(T;H_m,\mathcal{B}_m,\Pi_m)\right\} = O\left(\frac{K\log(T)}{\mu_{\dagger,m}^2}\right)$.

*Proof of Lemma 21:* Since $\gamma_1 = 0$, the reward sequence $H_m \in \mathcal{H}_m$ are all zeros. Further with $\gamma_2 = 0$, once a non-zero reward is received on arm $k_\dagger$, it will immediately have the highest sample mean. First, if arm $k_\dagger$ has been played at least $n' = \left\lceil \frac{\log(2T)}{2\mu_{\dagger,m}^2}\right\rceil$ times, where Hoeffding's inequality indicates that with a probability of at least $1-\frac{1}{2T}$, there is at least one non-zero reward collected from arm $k_\dagger$.

Furthermore, when there are no non-zero rewards collected on arm $k_\dagger$, the arms are pulled with equal probabilities (since they all have zero as sample means due to $\gamma_2 = 0$). With Bernstein's inequality, it further holds that

$$\begin{aligned} &\mathbb{P}\left(\sum_{t\in[\tau']} \mathbb{1}\{\pi_m(t) = k_\dagger\} - \frac{\tau'}{K} \leq n' - \frac{\tau'}{K}\right) \\ &\leq \exp\left(-\left(\frac{\tau'}{K} - n'\right)^2 \Big/ \left(2\frac{\tau'}{K} + \frac{2}{3}\left(\frac{\tau'}{K} - n'\right)\right)\right) \leq \frac{1}{2T}, \end{aligned}$$

where $\tau' = \frac{4K\log(2T)}{3\mu_{\dagger,m}^2} + \frac{32K\log(T)}{9} = O\left(\frac{K\log(T)}{\mu_{\dagger,m}^2}\right)$.

Thus, with at most $\tau'$ steps, the arm $k_\dagger$ would have the highest sample mean. Afterward, the other arms will only be pulled during exploration, i.e., with probability $\varepsilon(t) = O(K/t)$, which would only result in $O(K\log(T))$ pulls in expectation. The lemma is then proved. □

*Theorem 22:* (TAL with $\varepsilon$-greedy clients; Restatement of Theorem 11). For TAL with $\gamma_1 = \gamma_2 = 0$, if clients run $\varepsilon$-greedy and break ties uniformly at random, and $\mu_{\dagger,m} \neq 0, \forall m \in [M]$, it holds that $R_F(T) = O\left(\left[\frac{K\Delta_{\max}}{\Delta_{\min}^2} + \sum_{m\in[M]} \frac{K\Delta_{\max}}{\mu_{\dagger,m}^2}\right]\log(KMT)\right)$ and $C_F(T) = O\left(\sum_{m\in[M]}\left[\frac{K\mu_{*,m}}{M\Delta_{\min}^2} + \frac{K\mu_{*,m}}{\mu_{\dagger,m}^2}\right]\log(KMT)\right)$.

*Proof of Theorem 22:* From Eqn. (4), with probability $1-1/T$, it holds that $T_1 = O\left(\frac{K\log(KT)}{M\Delta_{\min}^2} + K\log(KMT)\right)$. Using $\overline{\eta}(T_1;\gamma_1,[K])$ from Lemma 20, it holds that

$$\begin{aligned} R_{F,1}(T) &\leq \sum_{m\in[M]}\sum_{k\neq k_\dagger} \Delta_k \cdot \overline{\eta}_m\left(T_1;\gamma_1,[K]\right) \\ &= O\left(\frac{K\Delta_{\max}\log(KT)}{\Delta_{\min}^2} + MK\Delta_{\max}\log(KMT)\right); \\ C_{F,1}(T) &\leq \sum_{m\in[M]}\sum_{k\in[K]} \delta_{k,m}(\gamma_1)\cdot\overline{\eta}_m\left(T_1;\gamma_1,[K]\right) \\ &\overset{(a)}{=} O\left(\frac{K\mu_{*,m}\log(KT)}{\Delta_{\min}^2} + MK\mu_{*,m}\log(KMT)\right), \end{aligned}$$

where step (a) leverages the fact that with $\gamma_1 = 0$, $\delta_{k,m}(\gamma_1) = \mu_{k,m} \leq \mu_{*,m}$.

Furthermore, combining Lemmas 15 and 21, with probability $1-1/T$, it holds that

$$\begin{aligned} R_{F,2}(T) &= O\left(\sum_{m\in[M]} \frac{K\Delta_{\max}\log(MT)}{\mu_{\dagger,m}^2}\right); \\ C_{F,2}(T) &\overset{(a)}{=} O\left(\sum_{m\in[M]} \frac{K\mu_{*,m}\log(MT)}{\mu_{\dagger,m}^2}\right), \end{aligned}$$

where step (a) leverages the fact that with $\gamma_2 = 0$, $\delta_{k,m}(\gamma_2) = \mu_{k,m} \leq \mu_{*,m}$. Putting these two observations into Theorem 16, the theorem is then proved. □

## Appendix B
## TWL: Performance Analysis

*Lemma 23:* (Arm Elimination in TWL). Denote event $\mathcal{E}_G$ as

$$\mathcal{E}_G = \{\text{each arm} k \neq k_\dagger \text{ is eliminated from the active arm set } \Upsilon \text{ in TWL by the end of epoch } \psi_k\},$$

where $\psi_k := \lceil \log_2(1/\Delta_k)\rceil$, it holds that $\mathbb{P}(\mathcal{E}_G) \geq 1-1/T$.

*Proof of Lemma 23:* First, similar to Lemma 13, we can establish that with probability at least $1-1/T$, it holds that

$$|\hat{\nu}_k(\psi) - \nu_k| \leq \mathrm{CB}(\psi) = 2^{-\psi-2}, \qquad \forall \psi \leq T, \forall k \in \Upsilon_\psi,$$

where $\Upsilon_\psi$ denotes the active arm set in epoch $\psi$. Based on this event, we can first observe that arm $k_\dagger$ would not be eliminated. Furthermore, at the end of epoch $\psi_k$, if arm $k \neq k_\dagger$ is not eliminated, both arm $k_\dagger$ and arm $k$ would be active. However, we can observe that

$$\begin{aligned} \mathrm{LCB}_\dagger(\psi_k) &= \hat{\nu}_\dagger(\psi_k) - \mathrm{CB}(\psi_k) \geq \nu_\dagger - 2\mathrm{CB}(\psi_k) \\ &\geq \nu_\dagger - \frac{\Delta_k}{2} \geq \nu_k + \frac{\Delta_k}{2} \geq \nu_k + 2\mathrm{CB}(\psi_k) \\ &\geq \hat{\nu}_k(\psi_k) + \mathrm{CB}(\psi_k) = \mathrm{UCB}_k(\psi_k), \end{aligned}$$

which means arm $k$ should already be eliminated. □

*Lemma 24:* (Active Arms in TWL). If $\Pi_m$ is UCB1, for any $\gamma_1$ and $\gamma_2$, with probability at least $1-1/T$, for all $k \neq k_\dagger$, it holds that $N^1_{k,m}(T) := \sum_{t\in[T]} \mathbb{1}\{\pi_m(t) = k, k \in \Upsilon(t)\} = O\left(\frac{\log(KT)}{M\Delta_k^2}\right)$, where $\Upsilon(t)$ denotes the active arm set at time step $t$ and $\psi(t)$ denotes the epoch index at time step $t$.

*Proof of Lemma 24:* Using the same procedure in Lemma 17, with constant reward $\gamma_1$'s for active arms and constant reward $\gamma_2$'s for inactive arms, it can be observed that at the end of epoch $\psi$, each client $m$ pulls each active arm $k \in \Upsilon(\psi)$ the same $F(\psi)$ times. As arm $k \neq k_\dagger$ is eliminated from the active arm set by the end of phase $\psi_k$ based on event $\mathcal{E}_G$ introduced in Lemma 23, it holds that $N^1_{k,m}(T) \leq F(\psi_k) = O\left(\frac{\log(KT)}{M\Delta_k^2}\right)$, which concludes the proof. □

*Lemma 25:* (Inactive Arms in TWL). If $\Pi_m$ is UCB1, for any $\gamma_1$ and $\gamma_2$ such that $\gamma_1 \geq \mu_{\dagger,m} > \gamma_2$, with probability at least $1-1/T$, for all $k \neq k_\dagger$, it holds that $N^2_{k,m}(T) := \sum_{t\in[T]} \mathbb{1}\{\pi_m(t) = k, k \notin \Upsilon(t)\} = O\left(\frac{(\gamma_1-\gamma_2)N^1_{k,m}(T)}{(\mu_{\dagger,m}-\gamma_2)} + \frac{\log(KT)}{(\mu_{\dagger,m}-\gamma_2)^2}\right)$.

*Proof of Lemma 25:* This lemma can be established following the same procedure as Lemma 18. □

*Theorem 26:* (TWL with UCB1 clients; Restatement of Theorem 12). For TWL with $\gamma_1 = 1$ and $\gamma_2 = 0$, if all clients run UCB1 locally and $\mu_{\dagger,m} \neq 0$ for all $m \neq [M]$, it holds that

$$R_F^{\mathrm{TWL}}(T) = O\Bigg(\sum_{m\in[M]}\sum_{k\neq k_\dagger}\left[\frac{\log(T)}{\mu_{\dagger,m}M\Delta_k} + \frac{\Delta_k\log(KT)}{\mu^2_{\dagger,m}}\right]\Bigg),$$

$$C_F^{\mathrm{TWL}}(T) = O\Bigg(\sum_{m\in[M]}\sum_{k\in[K]}\frac{(1-\mu_{k,m})\log(KT)}{M\Delta_k^2} + \sum_{m\in[M]}\sum_{k\neq k_\dagger}\left[\frac{\mu_{k,m}\log(KT)}{\mu_{\dagger,m}M\Delta_k^2} + \frac{\mu_{k,m}\log(KT)}{\mu^2_{\dagger,m}}\right]\Bigg).$$

*Proof of Theorem 26:* From Lemma 24, with probability at least $1-1/T$, it holds that $N^1_{k,m}(T) = O\left(\frac{\log(KT)}{M\Delta_k^2}\right)$, thus we can specify $N^2_{k,m}(T) = O\left(\frac{\log(KT)}{\mu_{\dagger,m}M\Delta_k^2} + \frac{\log(KT)}{\mu^2_{\dagger,m}}\right)$ with Lemma 25. The overall regret and cost can then be bound as

$$R_F^{\mathrm{TWL}}(T) \leq \sum_{m\in[M]}\sum_{k\neq k_\dagger}\left(N^1_{k,m}(T) + N^2_{k,m}(T)\right)\Delta_k + \frac{MT}{T}$$

$$= O\left(\sum\nolimits_{m\in[M]}\sum\nolimits_{k\neq k_\dagger}\frac{\log(KT)}{\mu_{\dagger,m}M\Delta_k} + \frac{\Delta_k\log(KT)}{\mu^2_{\dagger,m}}\right);$$

$$C^{\mathrm{TWL}}_{F,k,1}(T) \leq \sum\nolimits_{m\in[M]}\sum\nolimits_{k\in[K]} N^1_{k,m}(T)\cdot\delta_{k,m}(1) + \sum\nolimits_{m\in[M]}\sum\nolimits_{k\neq k_\dagger} N^1_{k,m}(T)\cdot\delta_{k,m}(0) + \frac{MT}{T}$$

$$= O\left(\sum\nolimits_{m\in[M]}\sum\nolimits_{k\in[K]}\frac{(1-\mu_{k,m})\log(KT)}{M\Delta_k^2} + \sum\nolimits_{m\in[M]}\sum\nolimits_{k\neq k_\dagger}\left[\frac{\mu_{k,m}\log(KT)}{\mu_{\dagger,m}M\Delta_k^2} + \frac{\mu_{k,m}\log(KT)}{\mu^2_{\dagger,m}}\right]\right),$$

which concludes the proof. □

## REFERENCES


[1] C. Shi, W. Xiong, C. Shen, and J. Yang, "Reward teaching for federated multi-armed bandits," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2023, pp. 1454–1459.

[2] Z. Zhu, J. Zhu, J. Liu, and Y. Liu, "Federated bandit: A gossiping approach," in *Proc. ACM Meas. Anal. Comput. Syst.*, 2021, vol. 5, no. 1, pp. 1–29.

[3] C. Shi and C. Shen, "Federated multi-armed bandits," in *Proc. 35th AAAI Conf. Artif. Intell.*, Feb. 2021, pp. 9603–9611.

[4] K. S. Reddy, P. Karthik, and V. Y. Tan, "Almost cost-free communication in federated best arm identification," 2022, *arXiv:2208.09215*.

[5] 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.

[6] A. Dubey and A. Pentland, "Differentially-private federated linear bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 6003–6014.

[7] C. Li and H. Wang, "Asynchronous upper confidence bound algorithms for federated linear bandits," in *Proc. Int. Conf. Artif. Intell. Statist.*, PMLR, 2022, pp. 6529–6553.

[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS)*, Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.

[9] X. Zhang, M. Hong, S. Dhople, W. Yin, and Y. Liu, "FedPD: A federated learning framework with adaptivity to non-IID data," *IEEE Trans. Signal Process.*, vol. 69, pp. 6055–6070, 2021.

[10] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[11] C. Tekin and M. Van Der Schaar, "Distributed online learning via cooperative contextual bandits," *IEEE Trans. Signal Process.*, vol. 63, no. 14, pp. 3700–3714, Jul. 2015.

[12] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5667–5681, Nov. 2010.

[13] T. Li and L. Song, "Privacy-preserving communication-efficient federated multi-armed bandits," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 3, pp. 773–787, Mar. 2022.

[14] N. Karpov and Q. Zhang, "Collaborative best arm identification with limited communication on non-IID data," 2022, *arXiv:2207.08015*.

[15] I. Demirel, Y. Yildirim, and C. Tekin, "Federated multi-armed bandits under Byzantine attacks," 2022, *arXiv:2205.04134*.

[16] C. Shi, C. Shen, and J. Yang, "Federated multi-armed bandits with personalization," in *Proc. Int. Conf. Artif. Intell. Statist.*, PMLR, 2021, pp. 2917–2925.

[17] L. Yang, Y.-Z. J. Chen, S. Pasteris, M. Hajiesmaili, J. Lui, and D. Towsley, "Cooperative stochastic bandits with asynchronous agents and constrained feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 8885–8897.

[18] Z. Chen, P. Karthik, V. Y. Tan, and Y. M. Chee, "Federated best arm identification with heterogeneous clients," 2022, *arXiv:2210.07780*.

[19] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, nos. 2–3, pp. 235–256, 2002.

[20] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, nos. 3/4, pp. 285–294, 1933.

[21] Y. Gai and B. Krishnamachari, "Distributed stochastic online learning policies for opportunistic spectrum access," *IEEE Trans. Signal Process.*, vol. 62, no. 23, pp. 6184–6193, Dec. 2014.

[22] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Multi-armed bandits in multi-agent networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Piscataway, NJ, USA: IEEE Press, 2017, pp. 2786–2790.

[23] C. Shi and C. Shen, "Multi-player multi-armed bandits with collision-dependent reward distributions," *IEEE Trans. Signal Process.*, vol. 69, pp. 4385–4402, 2021.

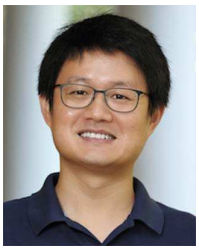
[24] A. Mitra, H. Hassani, and G. Pappas, "Exploiting heterogeneity in robust federated best-arm identification," 2021, *arXiv:2109.05700*.

[25] Y. Ma, K.-S. Jun, L. Li, and X. Zhu, "Data poisoning attacks in contextual bandits," in *Proc. Int. Conf. Decis. Game Theory Secur.*, Berlin, Germany: Springer-Verlag, 2018, pp. 186–204.

[26] E. Garcelon et al., "Adversarial attacks on linear contextual bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 14362–14373.
[27] L. Yang, M. H. Hajiesmaili, M. S. Talebi, J. C. Lui, and W. S. Wong, "Adversarial bandits with corruptions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 19943–19952.
[28] K.-S. Jun, L. Li, Y. Ma, and X. Zhu, "Adversarial attacks on stochastic bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3644–3653.
[29] F. Liu and N. Shroff, "Data poisoning attacks on stochastic bandits," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2019, pp. 4042–4050.
[30] A. Rangi, L. Tran-Thanh, H. Xu, and M. Franceschetti, "Saving stochastic bandits from poisoning attacks via limited data verification," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 7, pp. 8054–8061.
[31] S. Zuo, "Near optimal adversarial attack on UCB bandits," 2020, *arXiv:2008.09312*.
[32] Z. Feng, D. Parkes, and H. Xu, "The intrinsic robustness of stochastic bandits to strategic manipulation," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2020, pp. 3092–3101.
[33] G. Liu and L. Lai, "Action-manipulation attacks against stochastic bandits: Attacks and defense," *IEEE Trans. Signal Process.*, vol. 68, pp. 5152–5165, 2020.
[34] S. Kapoor, K. K. Patel, and P. Kar, "Corruption-tolerant bandit learning," *Mach. Learn.*, vol. 108, no. 4, pp. 687–715, 2019.
[35] Z. Guan et al., "Robust stochastic bandit algorithms under probabilistic unbounded adversarial attack," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4036–4043.
[36] C. Shi and C. Shen, "On no-sensing adversarial multi-player multi-armed bandits with collision communications," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 2, pp. 515–533, Jun. 2021.
[37] T. Lykouris, V. Mirrokni, and R. Paes Leme, "Stochastic bandits robust to adversarial corruptions," in *Proc. 50th Annu. ACM SIGACT Symp. Theory Comput.*, 2018, pp. 114–122.
[38] A. Gupta, T. Koren, and K. Talwar, "Better algorithms for stochastic bandits with adversarial corruptions," in *Proc. Conf. Learn. Theory*, PMLR, 2019, pp. 1562–1578.
[39] J. Liu, S. Li, and D. Li, "Cooperative stochastic multi-agent multi-armed bandits robust to adversarial corruptions," 2021, *arXiv:2106.04207*.
[40] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proc. Int. Conf. Mach. Learn.*, 1999, vol. 99, pp. 278–287.
[41] A. D. Laud, *Theory and Application of Reward Shaping in Reinforcement Learning*. Urbana, IL, USA: Univ. Illinois at Urbana-Champaign, 2004.
[42] O. Simsek and A. G. Barto, "An intrinsic reward mechanism for efficient exploration," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 833–840.
[43] X. Zhang, S. Bharti, Y. Ma, A. Singla, and X. Zhu, "The sample complexity of teaching by reinforcement on Q-learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10939–10947.
[44] M. Ghavamzadeh, S. Mahadevan, and R. Makar, "Hierarchical multi-agent reinforcement learning," *Auton. Agents Multi-Agent Syst.*, vol. 13, no. 2, pp. 197–229, 2006.
[45] S. Pateria, B. Subagdja, A. Tan, and C. Quek, "Hierarchical reinforcement learning: A comprehensive survey," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 5, pp. 1–35, 2021.
[46] Y. Mansour, A. Slivkins, and V. Syrgkanis, "Bayesian incentive-compatible bandit exploration," in *Proc. 16th ACM Conf. Econ. Comput.*, 2015, pp. 565–582.
[47] Y. Mansour, A. Slivkins, V. Syrgkanis, and Z. S. Wu, "Bayesian exploration: Incentivizing exploration in Bayesian games," in *Proc. ACM Conf. Econ. Comput.*, 2016, pp. 661–661.
[48] A. Slivkins, "Exploration and persuasion," in *Online and Matching-Based Market Design*, F. Echenique, N. Immorlica, and V. V. Vazirani, Eds., Cambridge, U.K.: Cambridge Univ. Press, 2023, pp. 655–675.
[49] P. Frazier, D. Kempe, J. Kleinberg, and R. Kleinberg, "Incentivizing exploration," in *Proc. 15th ACM Conf. Econ. Comput.*, 2014, pp. 5–22.
[50] L. Han, D. Kempe, and R. Qiang, "Incentivizing exploration with heterogeneous value of money," in *Proc. Int. Conf. Web Internet Econ.*, Berlin, Germany: Springer-Verlag, 2015, pp. 370–383.
[51] S. Wang and L. Huang, "Multi-armed bandits with compensation," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 5119–5128.
[52] C. Shi, H. Xu, W. Xiong, and C. Shen, "(Almost) free incentivized exploration from decentralized learning agents," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 560–571.
[53] C. Shen, T. Liu, and M. P. Fitz, "On the average rate performance of hybrid-ARQ in quasi-static fading channels," *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3339–3352, Nov. 2009.
[54] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, Dec. 2012.
[55] H. Wang, H. Xu, and H. Wang, "When are linear stochastic bandits attackable?" in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2022, pp. 23254–23273.
[56] J.-Y. Audibert, S. Bubeck et al., "Minimax policies for adversarial and stochastic bandits," in *Proc. 22nd Annu. Conf. Learn. Theory (COLT)*, 2009, vol. 7, pp. 1–122.
[57] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proc. 24th Annu. Conf. Learn. Theory*, 2011, pp. 359–376.
[58] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.
[59] A. Kalvit and A. Zeevi, "A closer look at the worst-case behavior of multi-armed bandit algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 8807–8819.
[60] K. Jamieson and R. Nowak, "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting," in *Proc. 48th Annu. Conf. Inf. Sci. Syst. (CISS)*, Piscataway, NJ, USA: IEEE Press, 2014, pp. 1–6.
[61] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and Markov decision processes," in *Proc. Int. Conf. Comput. Learn. Theory*, Berlin, Germany: Springer-Verlag, 2002, pp. 255–270.
[62] F. M. Harper and J. A. Konstan, "The MovieLens datasets: History and context," *ACM Trans. Interactive Intell. Syst. (TIIS)*, vol. 5, no. 4, pp. 1–19, 2015.

**Chengshuai Shi** received the B.E. degree in electrical engineering from the School of the Gifted Young, University of Science and Technology of China, in 2019. He is currently working toward the Ph.D. degree with Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia. His current research focuses on multiarmed bandits, federated learning, and reinforcement learning.

**Wei Xiong** received the B.S. degree in mathematics from the University of Science and Technology of China, in 2021, and the master's degree in mathematics from The Hong Kong University of Science and Technology, in 2023. He is currently working toward the Ph.D. degree with the Department of Computer Science, University of Illinois Urbana-Champaign. His current research interests focus on reinforcement learning and alignment for foundation generative models.



**Cong Shen** (Senior Member, IEEE) received the B.E. and M.E. degrees from the Department of Electronic Engineering, Tsinghua University, China, and the Ph.D. degree in electrical engineering from the University of California, Los Angeles. He is currently an Assistant Professor with Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia. He has an extensive industry experience, after having worked for Qualcomm, SpiderCloud Wireless, Silvus Technologies, and Xsense.ai, in various full time and consulting roles. He currently serves as an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, and IEEE TRANSACTIONS ON MACHINE LEARNING IN COMMUNICATIONS AND NETWORKING. He received the NSF CAREER award in 2022, and the Best Paper Award in 2021 IEEE International Conference on Communications (ICC). His general research interests are in the area of communications, wireless networks, and machine learning.

**Jing Yang** (Senior Member, IEEE) received the B.S. degree from the University of Science and Technology of China (USTC), and the M.S. and Ph.D. degrees from the University of Maryland, College Park, all in electrical engineering. She is an Associate Professor of electrical engineering with the Pennsylvania State University. She received the National Science Foundation CAREER award in 2015 and the WICE Early Achievement Award in 2020, and was selected as one of the 2020 N2Women: Stars in Computer Networking and Communications. She served as a Symposium/Track/Workshop Co-Chair for Asilomar 2023, ICC 2021, INFOCOM 2021-AoI Workshop, WCSP 2019, CTW 2015, PIMRC 2014, a TPC Member of several conferences, and an Editor for IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, from 2017 to 2020. She is now an Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. Her research interests lie in multiarmed bandits and reinforcement learning, federated learning, and wireless communications and networking.