Collaborative Linear Bandits with Adversarial Agents: Near-Optimal Regret Bounds

Aritra Mitra* **Arman Adibi*** **George J. Pappas** **Hamed Hassani**
Department of Electrical and Systems Engineering
{amitra20, aadibi, hassani, pappasg}@seas.upenn.edu

Abstract

We consider a linear stochastic bandit problem involving M agents that can collaborate via a central server to minimize regret. A fraction α of these agents are adversarial and can act arbitrarily, leading to the following tension: while collaboration can potentially reduce regret, it can also disrupt the process of learning due to adversaries. In this work, we provide a fundamental understanding of this tension by designing new algorithms that balance the exploration-exploitation trade-off via carefully constructed robust confidence intervals. We also complement our algorithms with tight analyses. First, we develop a robust collaborative phased elimination algorithm that achieves $\tilde{O}\left(\alpha + 1/\sqrt{M}\right)\sqrt{dT}$ regret for each good agent; here, d is the model-dimension and T is the horizon. For small α, our result thus reveals a clear benefit of collaboration despite adversaries. Using an information-theoretic argument, we then prove a matching lower bound, thereby providing the first set of tight, near-optimal regret bounds for collaborative linear bandits with adversaries. Furthermore, by leveraging recent advances in high-dimensional robust statistics, we significantly extend our algorithmic ideas and results to (i) the generalized linear bandit model that allows for non-linear observation maps; and (ii) the contextual bandit setting that allows for time-varying feature vectors.

1 Introduction

One of the primary challenges in modern large-scale computing systems is that of *security*. Given that the individual agents in such large systems are often vulnerable to attacks, it is important to understand how the overall system behaves in the face of adversarial corruptions. This observation has spurred a line of research dedicated to the design and analysis of distributed algorithms that are provably robust to a small fraction of adversarial agents; notably, motivated by emerging learning paradigms such as federated learning [1–3], this body of work has focused primarily on empirical risk minimization/stochastic optimization [4–14]. However, when it comes to multi-agent sequential decision-making problems under uncertainty (e.g., bandits and reinforcement learning), our understanding of analogous questions is quite limited. Our goal in this paper is to bridge the above gap by studying a collaborative linear bandit [15, 16] problem in the presence of adversaries.

In our model, M agents interact with the same linear bandit characterized by a d-dimensional unknown parameter θ_*, and a finite set of K arms. These agents can collaborate via a central server to improve performance, as measured by cumulative regret. As examples, consider (i) a team of robots exploring actions (arms) in a common environment and interacting with a central controller; and (ii) a group of people exploring restaurants (arms) and writing reviews for a web-recommendation server. In the absence of adversaries, there is a clear reason to collaborate in either case: by exchanging information, each agent can reduce its uncertainty about the arms faster than it could when it acts

*The first two authors contributed equally.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

alone, and thereby incur lesser regret. The situation becomes murkier and more delicate when certain agents *misbehave*: What if certain robots get attacked or certain people deliberately write spam reviews? More generally, the main question we ask in this paper is the following.

In a multi-agent linear stochastic bandit problem, can we still hope for benefits of collaboration when a fraction α of the agents are adversarial? If so, what are the fundamental limits of such benefits?

As far as we are aware, the answers to these questions have thus far remained elusive, motivating our current study. The main technical hurdle we must overcome is to delicately balance the exploration-exploitation trade-off in the presence of both statistical uncertainties due to the environment, and *worst-case* adversarial behavior. Importantly, the above trade-off - intrinsic to sequential decision-making - is absent in static optimization problems. Thus, the ideas used to guarantee robustness for distributed optimization do not apply to our problem, making our task quite non-trivial.

Our Contributions. In this paper, we contribute to a principled study of several canonical *structured* linear bandit settings with adversarial agents. Our specific contributions are as follows.

• **Robust Collaborative Linear Bandit Algorithm.** We propose `RCLB` - a phased elimination algorithm that relies on distributed exploration, and balances the exploration-exploitation dilemma in the presence of adversaries via carefully constructed *robust confidence intervals*. We prove that `RCLB` guarantees $\tilde{O}\left(\left(\alpha + 1/\sqrt{M}\right)\sqrt{dT}\right)$ regret for each good agent; see Theorem 1. This result is both novel, and significant in that it reveals a clear benefit of collaboration (despite adversaries) for small values of α. In particular, when $\alpha = 0$, the regret bound of `RCLB` is minimax-optimal in all relevant parameters: the model-dimension d, the horizon T, and the number of agents M.

• **Fundamental Limits.** At this stage, one may ask: *Is the additive $\alpha\sqrt{T}$ term in Theorem 1 simply an artifact of our analysis?* In Theorem 2, we establish a fundamental lower bound, revealing that such a term is in fact unavoidable; it is the price one must pay due to adversarial corruptions. A key implication of this result is that *our work is the first to provide tight, near-optimal regret bounds for collaborative linear bandits with adversaries.* The proof of Theorem 2 relies on a novel connection between the information-theoretic arguments in [17], and ideas from the robust mean estimation literature [18, 19]. As such, our proof technique may be relevant for related settings.

In our next set of contributions, we significantly extend our results to more general bandit models.

• **Generalized Linear Bandit Setting.** In Theorem 3, we prove that one can achieve bounds akin to that in Theorem 1 for the generalized linear bandit model (GLM) [20, 21] that accounts for non-linear observation maps. To achieve this result, we propose a variant of `RCLB` that leverages very recently developed tools from high-dimensional robust Gaussian mean estimation [22]. Deriving robust confidence intervals for this setting requires some work: we exploit regularity properties of the non-linear observation model along with error bounds from [22] for this purpose. As far as we are aware, Theorem 3 is the first result to establish adversarial-robustness for GLMs, allowing our framework to be applicable to a broad class of problems (e.g., logistic and probit regression models).

• **Contextual Bandit Setting.** Finally, we turn our attention to the contextual bandit setting where the feature vectors of the arms can change over time. This setting is practically quite relevant as web-recommendation systems are often modeled as contextual bandits [23]. The main challenge here arises from the need to simultaneously contend with time-varying optimal arms and adversaries. To handle this scenario, we develop a robust variant of the `SUPLINREL` algorithm [24] that guarantees a near-optimal regret bound identical to that of Theorem 1; see Theorem 4.

Overall, via the proposal of new robust algorithms complemented with tight analyses, our work takes an important step towards multi-agent sequential decision-making in the presence of adversaries.

Related Work. There is a growing strand of literature that studies the effect of reward corruption in stochastic bandits (for a single-agent setting) where the adversary has a fixed corruption budget [25–32]. The techniques in these papers do not apply to our work as our setting is very different: the adversarial agents in our model can act *arbitrarily*, and have no budget constraints. Several papers study multi-agent bandit problems in the absence of adversaries [33–52]. A few very recent ones [53–56] also look at the effect of attacks, but for the simpler unstructured multi-armed bandit problem. Accounting for adversarial agents in the structured linear bandit setting we consider here requires significantly different techniques that we develop in this paper.

2 Problem Formulation

We consider a setting comprising of a central server and M agents; the agents can communicate only via the server. Each agent $i \in [M]$ interacts with the same linear bandit model characterized by an unknown parameter θ_* that belongs to a known compact set $\Theta \subset \mathbb{R}^d$. We assume $\|\theta_*\| \leq 1$.[2] The set of actions $\mathcal{A}$ for each agent is given by K distinct vectors in $\mathbb{R}^d$, i.e., $\mathcal{A} = \{a_1, \ldots, a_K\}$, where K is a finite, positive integer. Based on all the information acquired by an agent i up to time $t-1$, it takes an action $a_{i,t} \in \mathcal{A}$ at time t, and receives a reward $y_{i,t}$ given by the following observation model:

$$y_{i,t} = \langle \theta_*, a_{i,t} \rangle + \eta_{i,t}. \tag{1}$$

Here, $\{\eta_{i,t}\}$ is a sequence of independent Gaussian random variables with zero mean and unit variance. Thus far, we have essentially described a distributed/multi-agent linear stochastic bandit model. Departing from this standard model, we focus on a setting where a fraction $\alpha \in [0, 1/2)$ of the agents are adversarial; the adversarial set is denoted by $\mathcal{B}$, where $|\mathcal{B}| = \alpha M$. In particular, we consider a *worst-case* attack model, where each adversarial agent $i \in \mathcal{B}$ is assumed to have complete knowledge of the system, and is allowed to act arbitrarily. Under this attack model, an agent $i \in \mathcal{B}$ can transmit arbitrarily corrupted messages to the central server.

Our performance measure of interest is the following group regret metric R_T defined w.r.t. the non-adversarial agents:

$$R_T = \mathbb{E}\left[\sum_{i \in [M] \setminus \mathcal{B}} \sum_{t=1}^{T} \langle \theta_*, a_* - a_{i,t} \rangle \right], \tag{2}$$

where $a_* = \text{argmax}_{a \in \mathcal{A}} \langle \theta_*, a \rangle$ is the optimal arm, and T is the time horizon.[3] We will work under a regime where the horizon T is large, satisfying $T \geq Md$. The goal of the good (non-adversarial) agents is to collaborate via the server and play a sequence of actions that minimize the group regret R_T. Let us now make a few key observations. In principle, each good agent can choose to act independently throughout (i.e., not talk to the server at all), and achieve $\tilde{O}(\sqrt{dT})$ regret by playing a standard bandit algorithm. Clearly, the group regret R_T would scale as $\tilde{O}\left((1-\alpha)M\sqrt{dT}\right)$ in such a case. In the absence of adversaries however, one can achieve a significantly better group regret bound of $\tilde{O}\left(\sqrt{MdT}\right)$ via collaboration, i.e., the regret per good agent can be reduced by a factor of $\sqrt{M}$ relative to the case when it acts independently (see, for instance, [41] and [45]). Our specific interest in this paper is to investigate whether, and to what extent, one can retain the benefits of collaboration despite the worst-case attack model described above. Said differently, we ask: *Can we improve upon the trivial per agent regret bound of $\tilde{O}(\sqrt{dT})$ in the presence of adversaries?*

Throughout the rest of the paper, we will answer the above question in the affirmative by deriving novel robust algorithms for several canonical bandit models, and then establishing near-optimal regret bounds for each such model.

3 Robust Collaborative Phased Elimination Algorithm for Linear Bandits

In this section, we develop a robust phased elimination algorithm that achieves the near-optimal regret bound of $\tilde{O}\left(\left(\alpha + \sqrt{1/M}\right)\sqrt{dT}\right)$ per good agent. This is non-trivial as we must account for the worst-case attack model described in Section 2. To highlight the challenges that we need to overcome, consider the following scenario. During the initial stages of the learning process, when the arms in $\mathcal{A}$ have not been adequately sampled by the agents, even a good agent may have "poor estimates" of the true payoffs associated with each arm, i.e., the variance associated with such estimates may be large. This statistical uncertainty can be exploited by the adversarial agents to their benefit. In particular, we need to devise an approach that can distinguish between benign stochastic perturbations (due to the noise in our model) and deliberate adversarial behavior. In what follows, we describe our proposed algorithm - `Robust Collaborative Phased Elimination for Linear Bandits (RCLB)` - that precisely does so in a principled way.

[2] We will use $\|\cdot\|$ to represent the Euclidean norm, and a' to denote the transpose of a vector a.

[3] For ease of exposition, we assume that there is an unique optimal arm.

Algorithm 1 Robust Collaborative Phased Elimination for Linear Bandits (RCLB)

Input: Action set $\mathcal{A} = \{a_1, \ldots, a_K\}$, confidence parameter δ, and corruption fraction α.
Initialize: $\ell = 1$ and $\mathcal{A}_1 = \mathcal{A}$.

1: Let $V_\ell(\pi) \triangleq \sum_{a \in \mathcal{A}_\ell} \pi(a) aa'$ and $g_\ell(\pi) \triangleq \max_{a \in \mathcal{A}_\ell} \|a\|^2_{V_\ell(\pi)^{-1}}$. Server solves an approximate G-optimal design problem to compute a distribution π_ℓ over $\mathcal{A}_\ell$ such that $g_\ell(\pi_\ell) \leq 2d$ and $|\text{Supp}(\pi_\ell)| \leq 48d \log\log d$.

2: For each $a \in \mathcal{A}_\ell$, server computes $m_a^{(\ell)}$ via Eq. (5), and broadcasts $\{m_a^{(\ell)}\}_{a \in \mathcal{A}_\ell}$ to all agents.

3: **for** $i \in [M] \setminus \mathcal{B}$ **do**

4: For each arm $a \in \mathcal{A}_\ell$, pull it $m_a^{(\ell)}$ times. Let $r_{i,a}^{(\ell)}$ be the average of the rewards observed by agent i for arm a during phase ℓ.

5: Compute local estimate $\hat{\theta}_i^{(\ell)}$ of θ_* as follows.[4]

$$\hat{\theta}_i^{(\ell)} = \tilde{V}_\ell^{-1} Y_{i,\ell}, \text{ where } \tilde{V}_\ell = \sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} aa' \; ; \; Y_{i,\ell} = \sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} r_{i,a}^{(\ell)} a. \tag{3}$$

6: Transmit $\hat{\theta}_i^{(\ell)}$ to server. Adversarial agents can transmit arbitrary vectors at this stage.

7: **end for**

8: Server computes robust mean pay-offs for each active arm: for each $a \in \mathcal{A}_\ell$, estimate $\mu_a^{(\ell)}$ as

$$\mu_a^{(\ell)} = \texttt{Median}\left(\{\langle \hat{\theta}_i^{(\ell)}, a \rangle, i \in [M]\}\right).$$

9: Define robust confidence threshold $\gamma_\ell \triangleq \sqrt{2} C \left(1 + \alpha\sqrt{M}\right) \epsilon_\ell$, where C is as in Lemma 1. Server performs phased elimination with the robust means and γ_ℓ to update active arm set:

$$\mathcal{A}_{\ell+1} = \{a \in \mathcal{A}_\ell : \max_{b \in \mathcal{A}_\ell} \mu_b^{(\ell)} - \mu_a^{(\ell)} \leq 2\gamma_\ell\}. \tag{4}$$

10: $\ell = \ell + 1$ and **Goto** line 1.

Description of RCLB (Algorithm 1). The RCLB algorithm we propose is inspired by the phased elimination algorithm in [57, Chapter 22], but features some key differences due to the distributed and adversarial nature of our problem. The algorithm proceeds in epochs/phases, and in each phase ℓ, the server maintains an active candidate set $\mathcal{A}_\ell$ of potential optimal arms. The exploration of arms in $\mathcal{A}_\ell$ is distributed among the agents. Upon such exploration, each agent i reports back a local estimate $\hat{\theta}_i^{(\ell)}$ of the unknown parameter θ_*; adversarial agents can transmit arbitrarily corrupted messages at this stage. Using the local parameter estimates $\{\hat{\theta}_i^{(\ell)}\}_{i \in [M]}$, the server then constructs (i) a *robust* estimate of the true mean payoff $\langle \theta_*, a \rangle$ for each active arm $a \in \mathcal{A}_\ell$, and (ii) an *inflated* confidence interval that captures both statistical and adversarial uncertainties associated with such robust mean estimates. This step is crucial to our scheme and requires a lot of care as we explain shortly. With the robust mean payoffs and associated confidence intervals in hand, the server eliminates arms with rewards far away from that of the optimal arm a_*. We now elaborate on the above ideas.

• *Optimal Experimental Design.* To minimize the regret incurred in each phase ℓ, we need to minimize the number of arm-pulls made to arms in $\mathcal{A}_\ell$. To that end, we appeal to a concept from statistics known as optimal experimental design. The concept is as follows. Let $\pi : \mathcal{A} \to [0, 1]$ be a distribution on $\mathcal{A}$ such that $\sum_{a \in \mathcal{A}} \pi(a) = 1$. Now define $V(\pi) \triangleq \sum_{a \in \mathcal{A}} \pi(a) aa'$; $g(\pi) \triangleq \max_{a \in \mathcal{A}} \|a\|^2_{V(\pi)^{-1}}$. The so-called *G-optimal design problem* seeks to find a distribution (design) π_* that minimizes g. In essence, sampling each arm $a \in \mathcal{A}$ in proportion to $\pi_*(a)$ minimizes the number of samples/arm-pulls needed to achieve a desired level of precision in the estimates of the arm means $\langle \theta_*, a \rangle$, $a \in \mathcal{A}$. For our purpose, we only need to solve an approximate G-optimal design problem: using the Frank-Wolfe algorithm and an appropriate initialization, one can *efficiently* find an approximate optimal design π such that $g(\pi) \leq 2d$, and $|\text{Supp}(\pi)| \leq 48d \log\log d$ [58, Chapter 3]. Accordingly, the server computes such an approximate optimal distribution π_ℓ over $\mathcal{A}_\ell$ in each epoch ℓ (line 1 of Algo. 1).

[4] Throughout, we assume that $\tilde{V}_\ell^{-1}$ is invertible in each epoch ℓ; this is the case when $\mathcal{A}_\ell$ spans $\mathbb{R}^d$. If $\mathcal{A}_\ell$ does not span $\mathbb{R}^d$, we can consider the lower dimensional subspace given by span($\mathcal{A}_\ell$).

• *Construction of Robust Arm-Payoff Estimates and Confidence Intervals.* In each phase ℓ, the server computes $T_a^{(\ell)}$ and $m_a^{(\ell)}$ for every $a \in \mathcal{A}_\ell$ as follows:

$$T_a^{(\ell)} = \left\lceil \frac{\pi_\ell(a) d}{\epsilon_\ell^2} \log\left(\frac{1}{\delta_\ell}\right) \right\rceil \text{ and } m_a^{(\ell)} = \left\lceil \frac{T_a^{(\ell)}}{M} \right\rceil, \tag{5}$$

where $\pi_\ell(a)$ is obtained from the approximate G-optimal design problem, $\epsilon_\ell = 2^{-\ell}$, $\delta_\ell = \bar{\delta}/(K\ell^2)$, and $\bar{\delta}$ is a design variable to be chosen later. The idea is to explore an arm a $T_a^{(\ell)}$ times to estimate $\langle \theta_*, a \rangle$ up to a precision that scales linearly with ϵ_ℓ; in later phases, we require progressively finer precision (hence, ϵ_ℓ decays exponentially with ℓ). The task of exploration is distributed among the agents, with each agent i being assigned $m_a^{(\ell)}$ arm-pulls for every $a \in \mathcal{A}_\ell$. Using the rewards that it observes during phase ℓ, each good agent $i \in [M] \setminus \mathcal{B}$ computes a local estimate $\hat{\theta}_i^{(\ell)}$ of θ_*, and transmits it to the server (lines 3-7 of Algo. 1). The key question now is as follows: *How should the server use the local estimates* $\{\hat{\theta}_i^{(\ell)}\}_{i \in [M]}$? Let us consider two natural strategies.

Candidate Strategies. One option could be to use the local estimates $\{\hat{\theta}_i^{(\ell)}\}_{i \in [M]}$ along with a high-dimensional robust mean estimation algorithm to compute a robust estimate of θ_*. Yet another strategy could be for the server to query the raw observations (i.e., the $y_{i,t}$'s) from the agents, use an univariate robust mean estimation algorithm (e.g., trimmed mean or median) to generate a "clean" version of each observation, and then use these clean observations to compute a robust estimate of θ_*. Although feasible, each of the above strategies can unfortunately lead to an additional $\sqrt{d}$ factor in the regret bound; we discuss this point in detail in the Appendix. The main message we want to convey here is that certain natural candidate solutions can lead to sub-optimal regret bounds.

Main Ideas. Our main insight is the following: to achieve near-optimal optimal regret bounds, one need not go through the route of first computing a robust estimate of θ_*. As our analysis will soon reveal, it suffices to instead compute robust estimates of the arm pay-offs $\{\langle \theta_*, a \rangle\}_{a \in \mathcal{A}_\ell}$ *directly* by employing the estimator in line 8 of Algo. 1. The key statistical property that we exploit here is that for each $a \in \mathcal{A}_\ell$ and $i \in [M] \setminus \mathcal{B}$, the quantity $\langle \hat{\theta}_i^{(\ell)}, a \rangle$ is conditionally-Gaussian with mean $\langle \theta_*, a \rangle$. This observation informs the choice of the median operator in line 8 of Algo. 1. Our next task is to compute appropriate confidence intervals for the robust arm-mean-estimates in order to eliminate sub-optimal arms. This is a delicate task as such confidence intervals need to account for both statistical uncertainties *and* adversarial perturbations. Indeed, if the confidence intervals are too tight, then they can lead to elimination of the optimal arm a_*; if they are too loose, then they can lead to large regret. The confidence threshold γ_ℓ in line 9 of Algo. 1 strikes just the right balance; the choice of such intervals is justified in Lemma 1.

We now state and discuss our main result concerning the performance of `RCLB`.

Theorem 1. ***(Performance of*** `RCLB`***)*** *Suppose $\alpha \in [0, 0.5)$. Given any $\delta \in (0, 1)$, set $\bar{\delta} = \delta/(10K)$. Then,* `RCLB` *guarantees that with probability at least $1 - \delta$, the following holds for each good agent $i \in [M] \setminus \mathcal{B}$:*

$$\sum_{t=1}^{T} \langle \theta_*, a_* - a_{i,t} \rangle = \tilde{O}\left(\left(\alpha + \sqrt{1/M}\right)\sqrt{dT}\right). \tag{6}$$

Main Takeaways. From Theorem 1, we note that `RCLB` guarantees sublinear regret despite the presence of adversarial agents. More importantly, the regret bound in Eq. (6) has optimal dependence on the model-dimension d, the horizon T, and also on the number of agents M when $\alpha = 0$ (i.e., in the absence of adversaries). When α is small, Eq. (6) reveals that one can indeed retain the benefits of collaboration, and improve upon the trivial per agent regret of $\tilde{O}(\sqrt{dT})$. Interestingly, this result mirrors that of a similar flavor for distributed stochastic optimization under attacks: given M machines, α fraction of which are corrupt, the authors in [6] showed that no algorithm can achieve statistical error lower than $\tilde{\Omega}\left(\alpha/\sqrt{T} + 1/\sqrt{MT}\right)$ for strongly convex loss functions; here, T is the number of samples on each machine. *As far as we are aware, this is the first work to establish an analogous result for collaborative linear bandits*. Inspired by the lower bound in [6], one may ask: *Is the additive $\alpha\sqrt{T}$ term in Eq. (6) unavoidable?* We will provide a definitive answer in Section 4. Finally, we note that Theorem 1 immediately implies a bound of $\tilde{O}\left(\left(\alpha M + \sqrt{M}\right)\sqrt{dT}\right)$ on the group regret R_T. We provide a formal proof of this fact in Appendix C.

Proof Outline of Theorem 1. The first main step in our analysis of Theorem 1 is to provide guarantees on the estimates $\{\mu_a^{(\ell)}\}_{a\in\mathcal{A}_\ell}$ computed in line 8 of RCLB. This is achieved as follows.

Lemma 1. ***(Robust Confidence Intervals)*** *Fix any epoch ℓ. There exists an universal constant $C > 0$ such that for each active arm $b \in \mathcal{A}_\ell$, the following holds with probability at least $1 - \delta_\ell$:*

$$|\mu_b^{(\ell)} - \langle \theta_*, b\rangle| \leq \gamma_\ell, \quad \textit{where} \quad \gamma_\ell = \sqrt{2}C\left(1 + \alpha\sqrt{M}\right)\epsilon_\ell. \tag{7}$$

Equipped with the above result, we argue that with high-probability, (i) the optimal arm a_* is never eliminated by RCLB (Lemma 4 in App. C); and (ii) in each epoch ℓ, an active arm in $\mathcal{A}_\ell$ can contribute to at most $O(\gamma_\ell)$ per-time-step regret (Lemma 5 in App. C). Putting these pieces together in a careful manner yields the desired result; we defer a detailed proof of Theorem 1 to Appendix C.

In the next section, we derive a lower bound that provides fundamental insights into the impact of the adversarial agents for the multi-agent sequential decision-making problem considered in this paper.

4 Lower Bounds

In this section, we assess the optimality of the regret bound obtained in Theorem 1. To do so, we consider a slightly different attack model: we assume that each of the agents is adversarial with probability α, independently of the other agents. Thus, the expected fraction of adversaries is α. Next, to provide a clean argument, we will focus on a class of policies Π where at each time-step t, the server assigns the *same* action to every agent, i.e., $a_{i,t} = a_t, \forall i \in [M]$. We note that all the algorithms developed in this paper adhere to policies in Π. Moreover, policies within the class Π yield the minimax optimal per-agent regret bound of order $\tilde{O}(\sqrt{dT}/\sqrt{M})$ when $\alpha = 0$. Since the standard multi-armed bandit setting [59] is a special case of the structured linear bandit setting considered here, a lower bound for the former implies one for the latter. With this in mind, let us denote by $\mathcal{E}_{\mathcal{N}}^{(K)}(1)$ the class of multi-armed bandits with K arms, where the reward distribution of each arm is Gaussian with unit variance. An instance $\nu_\mu \in \mathcal{E}_{\mathcal{N}}^{(K)}(1)$ is characterized by the mean vector $\mu \in \mathbb{R}^K$ associated with the K arms. Finally, let $R_T^{(s)}(\nu_\mu)$ denote the expected cumulative regret of the server (which is the same as that of a good agent $i \in [M] \setminus \mathcal{B}$) when it interacts with the instance ν_μ. We can now state the following result which establishes a fundamental lower bound for our problem.

Theorem 2. ***(Fundamental Lower Bound)*** *Given any policy in Π, there exist two distinct instances $\nu_\mu, \nu_{\mu'} \in \mathcal{E}_{\mathcal{N}}^{(2)}(1)$, and an universal constant $c > 1$, such that*

$$\max\{R_T^{(s)}(\nu_\mu), R_T^{(s)}(\nu_{\mu'})\} \geq c\alpha\sqrt{T}, \tag{8}$$

irrespective of the number of agents M.

Main Takeaways. Observe from Eq. (6) that even when M is arbitrarily large, the additive $\alpha\sqrt{T}$ term due to the adversaries remains unaffected - *Is this term truly unavoidable or just an artifact of our analysis?* Theorem 2 settles this question by revealing a fundamental performance limit: *every policy* in Π has to suffer the additive $\alpha\sqrt{T}$ regret, regardless of the number of agents. Thus, taken together, *Theorems 1 and 2 provide the first set of tight, near-optimal regret guarantees for the setting considered in this paper. We consider this to be a significant contribution of our work.*

Proof Idea for Theorem 2. For our setting, the standard techniques to prove lower bounds for non-adversarial bandits do not directly apply. The lower-bound proofs used for reward-corruption models [60], and attacks with a fixed budget [31], are not applicable either. This motivates us to use a new proof technique that combines information-theoretic arguments in [17] with ideas from the robust mean estimation literature [18, 19]. Specifically, for a two-armed bandit setting, we carefully construct two instances and attack strategies such that the joint distribution of rewards seen by the server is *identical* for both instances. Moreover, the instances are constructed such that (i) the optimal arm in one instance is sub-optimal for the other; and (ii) the per-time-step regret for selecting a sub-optimal arm in either instance is $\Omega(\alpha/\sqrt{T})$. For a detailed proof, see Appendix D.

Having established tight bounds for the linear bandit model in Section 2, in the sequel, we will show how our algorithmic ideas and results can be significantly extended to more general settings.

5 Extension to Generalized Linear Models with Adversaries

In this section, we will show how to achieve a regret bound akin to that in Theorem 1 for the non-linear observation model shown below [20, 21], known as the generalized linear model (GLM):

$$y_{i,t} = \mu\left(\langle \theta_*, a_{i,t}\rangle\right) + \eta_{i,t}, \tag{9}$$

where $\mu : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable function typically referred to as the (inverse) *link* function, and $\eta_{i,t} \sim \mathcal{N}(0,1)$ is as before. Our goal is to now control the following notion of regret:

$$R_T^{\text{GLM}} = \mathbb{E}\left[\sum_{i\in[M]\setminus\mathcal{B}} \sum_{t=1}^{T} \left(\mu\left(\langle \theta_*, a_*\rangle\right) - \mu\left(\langle \theta_*, a_{i,t}\rangle\right)\right)\right], \tag{10}$$

where $a_* = \text{argmax}_{a\in\mathcal{A}}\, \mu(\langle \theta_*, a\rangle)$. The main technical challenge relative to the setting considered in Section 2 pertains to the construction of the robust confidence intervals. In particular, the non-linearity of the map $\mu(\cdot)$ makes it hard to apply the technique adopted in line 8 of RCLB, necessitating a different approach that we describe next. We start with the following standard assumption [20, 21].

Assumption 1. *The function $\mu : \mathbb{R} \to \mathbb{R}$ is continuously differentiable, Lipschitz with constant $k_2 \geq 1$, and such that*

$$k_1 = \min\{1, \inf_{\theta\in\Theta, a\in\mathcal{A}} \dot{\mu}(\langle \theta, a\rangle)\} > 0.$$

Here, $\dot{\mu}(\cdot)$ is used to represent the derivative of $\mu(\cdot)$.

Next, for any $\theta \in \mathbb{R}^d$, we define $h_\ell(\theta) \triangleq \sum_{a\in\text{Supp}(\pi_\ell)} m_a^{(\ell)} \mu(\langle \theta, a\rangle) a$. We now describe a variant of RCLB - dubbed RC-GLM - for generalized linear models. Notably, unlike the standard bandit algorithms [20, 21] for GLM's that build on LinUCB, RC-GLM is based on phased elimination.

Description of RC-GLM. Our algorithm uses as a sub-routine the recently proposed Iteratively Reweighted Mean Estimator for computing a robust estimate of the mean of high-dimensional Gaussian random variables with adversarial outliers [22]. Specifically, suppose we are given M d-dimensional samples $x_1, \ldots, x_M$, such that $(1-\alpha)M$ of these samples are drawn i.i.d. from $\mathcal{N}(v, \Sigma)$, where $v \in \mathbb{R}^d$ is an unknown mean vector, and $\Sigma \in \mathbb{R}^{d\times d}$ is a known covariance matrix. The remaining αM samples are adversarial outliers and can be arbitrary. The estimator in [22] takes as input the M samples, the corruption fraction α, and the covariance matrix Σ. It then outputs an estimate $\hat{v}$ of v such that with high probability, $\|\hat{v} - v\| = \tilde{O}\left(\|\Sigma\|_2^{1/2}\left(\sqrt{d/M} + \alpha\sqrt{\log(1/\alpha)}\right)\right)$. Importantly, the estimator in [22] runs in polynomial-time, and is minimax-rate-optimal. Let us now see how this estimator - described in Appendix E - can be applied to our setting.

We only describe the key differences of RC-GLM relative to RCLB here, and defer a detailed description of RC-GLM to Appendix E. To get around the difficulty posed by the non-linear link function, our main idea is to first compute a robust estimate $\hat{\theta}^{(\ell)}$ of θ_* at the server, and then use it to develop a phased elimination strategy. To that end, instead of computing a local estimate $\hat{\theta}_i^{(\ell)}$ as in RCLB, each good agent $i \in [M] \setminus \mathcal{B}$ transmits $Y_{i,\ell}$ to the server, and the server computes a vector X_ℓ as follows: $X_\ell = \texttt{ITW}(\{\tilde{V}_\ell^{-1/2} Y_{i,\ell}, i \in [M]\})$. Here, $\tilde{V}_\ell^{-1/2}$ and $Y_{i,\ell}$ are as in Eq. (3), and we used $\texttt{ITW}(\cdot)$ to denote the output of the robust estimator in [22]. Our key observation here is that for each good agent i, $\tilde{V}_\ell^{-1/2} Y_{i,\ell}$ is a d-dimensional Gaussian random variable with mean $\tilde{V}_\ell^{-1/2} h_\ell(\theta_*)$, and covariance matrix $\Sigma = I_d$, justifying the use of the robust Gaussian mean estimator in [22]. The fact that $\Sigma = I_d$ is crucial in our algorithm design as the error-bound in [22] scales with the 2-norm of Σ. Essentially, the above steps enable us to extract a statistic X_ℓ that captures information about the agents' observations during epoch ℓ. Using this statistic, the server next computes an estimate $\hat{\theta}^{(\ell)}$ of θ_* by solving $h_\ell(\hat{\theta}^{(\ell)}) = \tilde{V}_\ell^{1/2} X_\ell$,[5] and employs the following phased elimination strategy:

$$\mathcal{A}_{\ell+1} = \{a \in \mathcal{A}_\ell : \max_{b\in\mathcal{A}_\ell} \mu(\langle \hat{\theta}^{(\ell)}, b\rangle) - \mu(\langle \hat{\theta}^{(\ell)}, a\rangle) \leq 2\bar{\gamma}_\ell\}; \;\; \bar{\gamma}_\ell = \bar{C}(k_2/k_1)\left(\sqrt{d} + \alpha\sqrt{M\log(1/\alpha)}\right)\epsilon_\ell,$$

where $\bar{C}$ is an universal constant known to the server. Deriving an analogue of Lemma 1 to compute the robust confidence threshold $\bar{\gamma}_\ell$ requires some work. This is achieved by exploiting the regularity properties of the link function in tandem with the confidence bounds in [22]; see Appendix E for details and a proof of our main result for RC-GLM stated below.

[5] We argue in Appendix E that this equation admits a unique solution.

Algorithm 2 Robust BaseLinUCB (at Server)

Input: Confidence parameter $\bar{\delta}$, corruption fraction α, and index set $\psi_t \subseteq [t-1]$.

$A_t \leftarrow \frac{I_d}{M} + \sum_{\tau \in \psi_t} x_{\tau,a_\tau} x'_{\tau,a_\tau}$.

for $i \in [M]$ **do** ▷ Compute local parameters for each agent

  $b_{i,t} \leftarrow \sum_{\tau \in \psi_t} r_{i,\tau} x_{\tau,a_\tau}; \hat{\theta}_{i,t} \leftarrow A_t^{-1} b_{i,t}$.

end for

for $a \in [K]$ **do** ▷ Compute robust estimates for each feature vector

  $\hat{r}_{t,a} \leftarrow \texttt{Median}\left(\{\langle \hat{\theta}_{i,t}, x_{t,a}\rangle, i \in [M]\}\right); w_{t,a} \leftarrow \left(\alpha + 2C\sqrt{\frac{\log(\frac{1}{\delta})}{M}}\right) \|x_{t,a}\|_{A_t^{-1}}$, where C is as in Lemma 1.

end for

Theorem 3. ***(Performance of Algorithm*** `RC-GLM`***)*** *Suppose $\alpha < (5-\sqrt{5})/10$, and $M = \Omega(\log(KT))$. Given any $\delta \in (0,1)$,* `RC-GLM` *guarantees that with probability at least $1-\delta$, the following holds for each good agent $i \in [M] \setminus \mathcal{B}$:*

$$\sum_{t=1}^{T} \left(\mu\left(\langle \theta_*, a_*\rangle\right) - \mu\left(\langle \theta_*, a_{i,t}\rangle\right)\right) = \tilde{O}\left((k_2/k_1)\left(\alpha\sqrt{\log(1/\alpha)} + \sqrt{d/M}\right)\sqrt{dT}\right). \tag{11}$$

Main Takeaways. Theorem 3 significantly generalizes Theorem 1 and shows that even for general non-linear observation maps, one can reap the benefits of collaboration in the presence of adversaries. The additional $\sqrt{d}$ factor in the bound of (11) relative to that in (6) is inherited from the error-rate guarantees in [22]; the task of tightening this bound is left as future work. Nonetheless, Theorem 3 is the only result we are aware of that provides adversarial-robustness guarantees for GLMs.

Remark 1. ***(Communication Complexity of*** `RCLB` ***and*** `RC-GLM`***)*** *It is not hard to see that the number of epochs/phases in* `RCLB` *and* `RC-GLM` *is $O(\log(MT))$. Since communication between the server and the agents occurs only once in every epoch, we note that the communication complexity of these algorithms scale logarithmically with the horizon T. Thus, our proposed algorithms not only lead to near-optimal regret bounds in the face of worst-case adversarial attacks, they are also communication-efficient by design. This is an important point to take note of as communication-efficiency is a key consideration in large-scale computing paradigms such as federated learning.*

6 Robust Collaborative Contextual Bandits with Adversaries

In this section, we will consider a collaborative contextual bandit setting where at each time-step $t \in [T]$, the server and the agents observe K d-dimensional feature vectors, $\{x_{t,a} | a \in [K]\}$, with $\|x_{t,a}\| \leq 1, \forall a \in [K]$ and $\forall t \in [T]$. We assume that the adversarial agents have no control over the generation of the feature vectors. Associated with each arm $a \in [K]$, the stochastic reward observed by an agent $i \in [M]$ comes from the following observation model:

$$y_{i,t}(a) = \langle \theta_*, x_{t,a}\rangle + \eta_{i,t}(a), \tag{12}$$

where $\{\eta_{i,t}(a)\}$ are drawn i.i.d. from $\mathcal{N}(0,1)$. At each time step t, a good agent $i \in [M] \setminus \mathcal{B}$ plays an action $a_{i,t} \in [K]$, and receives the corresponding reward $r_{i,t} \triangleq y_{i,t}(a_{i,t})$ based on the observation model in (12). The main difference of the setting considered here relative to the one in Section 2 is that the feature vectors for each arm can change over time. As a result, the optimal action $a_t^* = \text{argmax}_{a \in [K]} \langle \theta_*, x_{t,a}\rangle$ can change over time, making it particularly challenging to compete with a time-varying optimal action in the presence of adversaries. This dictates the need for a different algorithmic strategy compared to the one we developed in Section 1. Before we develop such a strategy, let us first formally define the performance metric of interest to us in this setting:

$$R_T^{\text{Context}} = \mathbb{E}\left[\sum_{i \in [M] \setminus \mathcal{B}} \sum_{t=1}^{T} \langle \theta_*, x_{t,a_t^*} - x_{t,a_{i,t}}\rangle\right]. \tag{13}$$

The main question we ask is: *For the contextual bandit setting described above, can one continue to hope for benefits of collaboration in the presence of adversaries?* In what follows, we will answer this question in the affirmative by developing a variant of the `SUPLINREL` algorithm in [24].

Algorithm 3 Robust Collaborative SupLinUCB for Contextual Bandits (at Server)

Input: Confidence parameter δ, corruption fraction α, and horizon T.

1: $S \leftarrow \lceil \ln T \rceil; \psi_1^{(s)} \leftarrow \emptyset, \forall s \in [S]$.
2: **for** $t \in [T]$ **do**
3: $s \leftarrow 1$ and $\mathcal{A}_1 \leftarrow [K]$.
4: **repeat**
5: Use Algorithm 2 with $\bar{\delta} = \delta/(KST)$, and index set $\psi_t^{(s)}$ to compute robust estimates $\{\hat{r}_{t,a}^{(s)}, w_{t,a}^{(s)}\}$ of the means and variances of the payoffs associated with each arm $a \in \mathcal{A}_s$.
6: If $w_{t,a}^{(s)} > 2^{-s}/\sqrt{M}$ for some $a \in \mathcal{A}_s$, then choose this arm, i.e., set $a_t = a$. Store the corresponding phase: $\psi_{t+1}^{(s)} = \psi_t^{(s)} \cup \{t\}, \psi_{t+1}^{(\ell)} = \psi_t^{(\ell)}\ \forall \ell \neq s$.
7: Else if $w_{t,a}^{(s)} \leq 1/\sqrt{MT}\ \forall a \in \mathcal{A}_s$, then select the arm with the *highest robust upper confidence bound*: $a_t = \text{argmax}_{a \in \mathcal{A}_s} \left(\hat{r}_{t,a}^{(s)} + w_{t,a}^{(s)}\right)$. Do not store this phase, i.e., $\psi_{t+1}^{(\ell)} = \psi_t^{(\ell)}\ \forall \ell \in [S]$.
8: Else if $w_{t,a}^{(s)} \leq 2^{-s}/\sqrt{M}\ \forall a \in \mathcal{A}_s$, then update active arm-set as

$$\mathcal{A}_{s+1} = \{a \in \mathcal{A}_s | \max_{b \in \mathcal{A}_s} \left(\hat{r}_{t,b}^{(s)} + w_{t,b}^{(s)}\right) - \left(\hat{r}_{t,a}^{(s)} + w_{t,a}^{(s)}\right) \leq 2^{(1-s)}/\sqrt{M}\}.$$

9: $s \leftarrow s + 1$.
10: **until** an action a_t is chosen.
11: Broadcast the chosen action a_t to every agent (i.e., $a_{i,t} = a_t, \forall i \in [M]$), and receive corresponding rewards $\{r_{i,t}\}_{i \in [M]}$. Adversarial agents can transmit arbitrary reward values.
12: **end for**

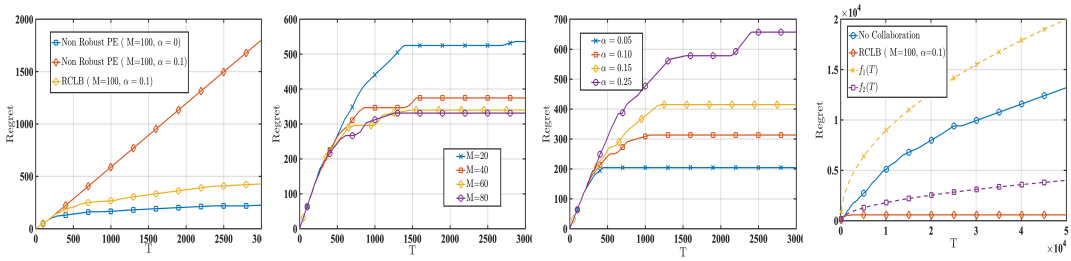
Description of Algorithm 3. At each time-step t, our proposed algorithm, namely Algorithm 3, scans through the set $\mathcal{A}$ of arms to determine a suitable action a_t. This scanning process (lines 4-10 of Algo. 3) is done at the server over S phases. Corresponding to each phase $s \in [S]$, the server maintains a set $\psi_t^{(s)}$; the set $\psi_t^{(s)}$ stores all the time-steps in $[t-1]$ where an action is chosen in phase s of the scanning process based on line 6 on Algo. 3. The scanning process itself relies on Algorithm 2 as a sub-routine. Specifically, in each phase s, the server first invokes Algorithm 2 to obtain a robust estimate $\hat{r}_{t,a}^{(s)}$ of $\langle \theta_*, x_{t,a} \rangle$ for each arm $a \in \mathcal{A}_s$, along with an associated *inflated* confidence width $w_{t,a}^{(s)}$ (line 5 of Algo. 3). If the confidence width is too large for a particular arm (as in line 6), then such an arm requires exploration and is accordingly chosen to be a_t. If, on the other hand, the confidence widths of all arms are sufficiently small (as in line 7), then a_t is chosen to be the arm with the highest upper-confidence bound. Thus, we follow the principle of *optimism in the face of uncertainty here, while exercising caution to account for the presence of adversaries (via the use of inflated confidence intervals)*. If the conditions in lines 6 and 7 both fail, then the arms in $\mathcal{A}_s$ require further screening. Accordingly, we move to the next phase $s+1$, retaining only those arms that are sufficiently close to the optimal arm a_t^*; see line 8 of Algo. 3. Our main innovation lies in (i) the construction of the robust arm-estimates in Algo. 2 that account for both statistical and adversarial behavior, and (ii) the careful use of such estimates in lines 6-8 of Algo. 3 to pick the action a_t. The next result reveals that the combination of these ideas yields near-optimal regret bounds.

Theorem 4. ***(Performance of Algo. 3)*** *Suppose $\alpha \in (0, 0.5)$. Given any $\delta \in (0,1)$, Algo. 3 guarantees that with probability at least $1-\delta$, the following holds for each good agent $i \in [M] \setminus \mathcal{B}$:*

$$\sum_{t=1}^{T} \langle \theta_*, x_{t,a_t^*} - x_{t,a_{i,t}} \rangle = \tilde{O}\left(\left(\alpha + \sqrt{1/M}\right)\sqrt{dT}\right). \tag{14}$$

Main Takeaways. For the contextual bandit setting considered here, the single-agent minimax optimal regret in the absence of adversaries is $\tilde{O}(\sqrt{dT})$ [24]. In light of the lower bound in Theorem 2, we see that Theorem 4 provides a near-optimal regret guarantee, just as Theorem 1.

Remark 2. *For ease of exposition, we have considered Gaussian noise (in the observation model) throughout the paper. However, both our algorithms and results can be extended with slight modifications to sub-Gaussian noise sequences. We elaborate on this point in Appendix B.*



(a) (b) (c) (d)

Figure 1: Plots of per-agent regret for the linear bandit experiment. (a) Comparison between RCLB and a vanilla non-robust phased elimination algorithm. (b) Performance of RCLB for varying number of agents M, with $\alpha = 0.1$. (c) Performance of RCLB for varying corruption fraction α, with $M = 100$. For (d), we set $\alpha = 0.1$, $M = 100$, and compare RCLB to a phased elimination algorithm where the agents do not collaborate. We also plot theoretical upper-bounds: $f_1(T) = 40\sqrt{dT}$ and $f_2(T) = 40(\alpha + \sqrt{(1/M)})\sqrt{dT}$.

7 Simulation Results

We report simulation results on synthetic data to corroborate our developed theory. Additional simulations on contextual bandits and alternate attack models are presented in Appendix H.

Experimental Setup. We consider a setting with 50 actions in $\mathbb{R}^5$; we describe how these actions and θ^* are generated in Appendix H. The rewards are generated based on the observation model in Eq. (1). We now describe the attack model. To manipulate the server into selecting sub-optimal arms, each adversarial agent i employs the simple strategy of reducing the rewards of the good arms and increasing the rewards of the bad arms. More precisely, in each epoch ℓ, upon pulling an arm a and observing the corresponding reward y_a, an adversarial agent i does the following: if $y_a > p\langle\theta_*, a_*\rangle$, then this reward is corrupted to $\tilde{y}_a = y_a - \beta$; and if $y_a \leq p\langle\theta_*, a_*\rangle$, then the reward is corrupted to $\tilde{y}_a = y_a + \beta$. For this experiment, we fix $p = 0.6$ and $\beta = 5$. Agent $i \in \mathcal{B}$ then uses all the corrupted rewards in epoch ℓ to generate the local model estimate $\hat{\theta}_i^{(\ell)}$ that is transmitted to the server.

Discussion of Simulation Results. Fig. 1 summarizes our experimental results. In Fig. 1(a), we compare our proposed algorithm RCLB to a vanilla distributed phased elimination (PE) algorithm that does not account for adversarial agents. Specifically, the latter is designed by replacing the median operation in line 8 of Algorithm 1 with a mean operation, and setting the threshold γ_ℓ in line 9 to be ϵ_ℓ. Fig. 1(a) shows that even a small fraction $\alpha = 0.1$ of adversaries can cause the non-robust PE algorithm to incur linear regret. In contrast, RCLB continues to guarantee sub-linear regret bounds despite adversarial corruptions. Furthermore, the regret bound of RCLB in the presence of a small fraction of adversarial agents is close to that of the non-robust algorithm in the absence of adversaries. This goes on to establish the robustness of RCLB.

Fig. 1(b) depicts the performance of RCLB for varying values of the number of agents M, at a fixed corruption level $\alpha = 0.1$. We observe that increasing M results in lower regret, indicating a clear benefit of collaboration despite the presence of adversaries. In Fig. 1(c), we vary the corruption fraction α, keeping M fixed at 100. As expected, increasing α leads to higher (albeit sub-linear) regret. Importantly, the trends observed in both Fig. 1(b) and Fig. 1(c) are consistent with the theoretical upper-bound of $O((\alpha + 1/\sqrt{M})\sqrt{dT})$ predicted by Theorem 1.

A trivial way to avoid adversarial corruption is for a good agent to not participate in any collaboration at all, and run a standard single-agent bandit algorithm. This would result in such an agent incurring $O(\sqrt{dT})$ regret. The purpose of Fig. 1(d) is to drive home the point that RCLB *can lead to significant improvements over a trivial non-collaborative strategy.* To make this point clear, we compare RCLB to a standard single-agent phased elimination algorithm that does not involve any collaboration, and observe that despite adversarial corruption, RCLB leads to considerably lower regret bounds as compared to the non-collaborative strategy. This highlights the importance of our approach.

Acknowledgments and Disclosure of Funding. This work was supported by NSF Award 1837253, NSF CAREER award CIF 1943064, the AFOSR-YIP award FA9550-20-1-0111, and EnCORE.

References

[1] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[2] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, H Brendan McMahan, et al. Towards federated learning at scale: System design. *arXiv preprint arXiv:1902.01046*, 2019.

[3] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[4] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017.

[5] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 118–128, 2017.

[6] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR, 2018.

[7] Lingjiao Chen, Zachary Charles, Dimitris Papailiopoulos, et al. Draco: Robust distributed training via redundant gradients. *arXiv preprint arXiv:1803.09877*, 2018.

[8] Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. *arXiv preprint arXiv:1803.08917*, 2018.

[9] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Generalized byzantine-tolerant sgd. *arXiv preprint arXiv:1802.10116*, 2018.

[10] Liping Li, Wei Xu, Tianyi Chen, Georgios B Giannakis, and Qing Ling. Rsa: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1544–1551, 2019.

[11] Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*, 2019.

[12] Avishek Ghosh, Raj Kumar Maity, Swanand Kadhe, Arya Mazumdar, and Kannan Ramachandran. Communication efficient and Byzantine tolerant distributed learning. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2545–2550. IEEE, 2020.

[13] Avishek Ghosh, Raj Kumar Maity, and Arya Mazumdar. Distributed newton can communicate less and resist Byzantine workers. *arXiv preprint arXiv:2006.08737*, 2020.

[14] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *International Conference on Machine Learning*, pages 5311–5319. PMLR, 2021.

[15] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

[16] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

[17] Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134. PMLR, 2013.

[18] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance matrix estimation via matrix depth. *arXiv preprint arXiv:1506.00691*, 2015.

[19] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

[20] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *NIPS*, volume 23, pages 586–594, 2010.

[21] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

[22] Arnak S. Dalalyan and Arshak Minasyan. All-in-one robust estimator of the gaussian mean. *The Annals of Statistics*, 2022.

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

[24] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[25] Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Xiaojin Zhu. Adversarial attacks on stochastic bandits. *arXiv preprint arXiv:1810.12188*, 2018.

[26] Fang Liu and Ness Shroff. Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning*, pages 4042–4050. PMLR, 2019.

[27] Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

[28] Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR, 2019.

[29] Ilija Bogunovic, Andreas Krause, and Jonathan Scarlett. Corruption-tolerant gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1071–1081. PMLR, 2020.

[30] Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Jean Tarbouriech, Olivier Teytaud, Alessandro Lazaric, and Matteo Pirotta. Adversarial attacks on linear contextual bandits. *arXiv preprint arXiv:2002.03839*, 2020.

[31] Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 991–999. PMLR, 2021.

[32] Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2205.06811*, 2022.

[33] Keqin Liu and Qing Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, 58(11):5667–5681, 2010.

[34] Dileep Kalathil, Naumaan Nayyar, and Rahul Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

[35] Soummya Kar, H Vincent Poor, and Shuguang Cui. Bandit problems in networks: Asymptotically efficient distributed allocation rules. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 1771–1778. IEEE, 2011.

[36] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 167–172. IEEE, 2016.

[37] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision making in multi-agent multi-armed bandits. *Automatica*, 125:109445, 2021.

[38] Shahin Shahrampour, Alexander Rakhlin, and Ali Jadbabaie. Multi-armed bandits in multi-agent networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2786–2790. IEEE, 2017.

[39] Swapna Buccapatnam, Jian Tan, and Li Zhang. Information sharing in distributed stochastic bandits. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 2605–2613. IEEE, 2015.

[40] Ravi Kumar Kolla, Krishna Jagannathan, and Aditya Gopalan. Collaborative learning of stochastic bandits over a social network. *IEEE/ACM Transactions on Networking*, 26(4):1782–1795, 2018.

[41] Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. *arXiv preprint arXiv:1904.06309*, 2019.

[42] Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–35, 2019.

[43] David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. *arXiv preprint arXiv:1810.04468*, 2018.

[44] Abhimanyu Dubey et al. Kernel methods for cooperative multi-agent contextual bandits. In *International Conference on Machine Learning*, pages 2740–2750. PMLR, 2020.

[45] Abhimanyu Dubey and Alex Pentland. Differentially-private federated linear bandits. *arXiv preprint arXiv:2010.11425*, 2020.

[46] Anusha Lalitha and Andrea Goldsmith. Bayesian algorithms for decentralized stochastic bandits. *arXiv preprint arXiv:2010.10569*, 2020.

[47] Ronshee Chawla, Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. The gossiping insert-eliminate algorithm for multi-agent bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3471–3481. PMLR, 2020.

[48] Ronshee Chawla, Abishek Sankararaman, and Sanjay Shakkottai. Multi-agent low-dimensional linear bandits. *arXiv preprint arXiv:2007.01442*, 2020.

[49] Avishek Ghosh, Abishek Sankararaman, and Kannan Ramchandran. Collaborative learning and personalization in multi-agent stochastic linear bandits. *arXiv preprint arXiv:2106.08902*, 2021.

[50] Mridul Agarwal, Vaneet Aggarwal, and Kamyar Azizzadenesheli. Multi-agent multi-armed bandits with limited communication. *arXiv preprint arXiv:2102.08462*, 2021.

[51] Zhaowei Zhu, Jingxuan Zhu, Ji Liu, and Yang Liu. Federated bandit: A gossiping approach. In *Abstract Proceedings of the 2021 ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, pages 3–4, 2021.

[52] Chengshuai Shi, Cong Shen, and Jing Yang. Federated multi-armed bandits with personalization. In *International Conference on Artificial Intelligence and Statistics*, pages 2917–2925. PMLR, 2021.

[53] Abhimanyu Dubey and Alex Pentland. Private and byzantine-proof cooperative decision-making. In *AAMAS*, pages 357–365, 2020.

[54] Daniel Vial, Sanjay Shakkottai, and R Srikant. Robust multi-agent multi-armed bandits. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 161–170, 2021.

[55] Daniel Vial, Sanjay Shakkottai, and R Srikant. Robust multi-agent bandits over undirected graphs. *arXiv preprint arXiv:2203.00076*, 2022.

[56] Aritra Mitra, Hamed Hassani, and George Pappas. Exploiting heterogeneity in robust federated best-arm identification. *arXiv preprint arXiv:2109.05700*, 2021.

[57] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[58] Michael J Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

[59] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[60] Sayash Kapoor, Kumar Kshitij Patel, and Purushottam Kar. Corruption-tolerant bandit learning. *Machine Learning*, 108(4):687–715, 2019.

[61] Zixin Zhong, Wang Chi Cheung, and Vincent Tan. Probabilistic sequential shrinking: A best arm identification algorithm for stochastic bandits with corruptions. In *International Conference on Machine Learning*, pages 12772–12781. PMLR, 2021.

[62] Shreyas Sundaram and Bahman Gharesifard. Distributed optimization under adversarial nodes. *IEEE Transactions on Automatic Control*, 64(3):1063–1076, 2018.

[63] Lili Su and Nitin H Vaidya. Byzantine-resilient multiagent optimization. *IEEE Transactions on Automatic Control*, 66(5):2227–2233, 2020.

[64] Kananart Kuwaranancharoen, Lei Xin, and Shreyas Sundaram. Byzantine-resilient distributed optimization of multi-dimensional functions. In *2020 American Control Conference (ACC)*, pages 4399–4404. IEEE, 2020.

[65] Nirupam Gupta, Thinh T Doan, and Nitin H Vaidya. Byzantine fault-tolerance in decentralized optimization under 2f-redundancy. In *2021 American Control Conference (ACC)*, pages 3632–3637. IEEE, 2021.

[66] Arman Adibi, Aritra Mitra, George J Pappas, and Hamed Hassani. Distributed statistical min-max learning in the presence of Byzantine agents. *arXiv preprint arXiv:2204.03187*, 2022.

[67] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[68] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[69] Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proc. of the thirtieth annual ACM-SIAM symp. on discrete algorithms*, pages 2755–2771. SIAM, 2019.

[70] Stanislav Minsker. Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*, 2018.

[71] Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] While explaining the main takeaways from Theorem 3 in Section 5, we mention that our regret bound features an additional $\sqrt{d}$ term relative to Theorem 1. Tightening this bound is part of our future work.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We could not think of any negative societal impacts of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] We provide complete proofs of all our results in the supplemental material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Detailed Discussion of Related Work

Below, we provide a detailed discussion of relevant work.

• **Reward Corruption Attacks in Stochastic Bandits.** In the single-agent setting, there is a rich body of work that studies the effect of reward-corruption in stochastic bandits, both for the unstructured multi-armed bandit problem [25–28], and also for structured linear bandits [29–32]. In these works, an adversary can modify the true stochastic reward/feedback on certain rounds; a corruption budget C captures the total corruption injected by the adversary over the horizon T. The attack model we study is fundamentally different: the adversaries in our setting can inject corruptions of *arbitrary* magnitude in *all* rounds, i.e., there are no budget constraints. As such, the algorithmic techniques in [25–31] do not apply to our model.

Continuing with this point, we note that in [28], the authors proved an *algorithm-independent* lower bound of $\Omega(C)$ on the regret. This lower bound suggests that for the reward-corruption attack model, when the attacker's budget C scales linearly with the horizon T, there is no hope for achieving sub-linear regret. In [60], the authors studied a reward-corruption model closely related to those in [27–29], where in each round, with probability η (independently of the other rounds), the attacker can bias the reward seen by the learner. Similar to the lower bound in [28], the authors in [60] proved a lower bound of $\Omega(\eta T)$ on the regret for their model. In sharp contrast to the fundamental limits established in [28, 60], for our setup, as long as the corruption fraction α is strictly less than half, we prove that with high-probability it is in fact possible to achieve sub-linear regret. The key is that for our setting, the server can leverage "clean" information from the good agents in every round; of course, the identities of such good agents are not known to the server. We finally note that beyond the task of minimizing cumulative regret, the impact of fixed-budget reward-contamination has also been explored for the problem of best-arm identification in [61].

• **Multi-Agent Bandits.** There is a growing literature that studies multi-agent multi-armed bandit problems in the absence of adversaries, both over peer-to-peer networks, and also for the server-client architecture model [33–52]. The main focus in these papers is the design of coordination protocols among the agents that balance communication-efficiency with performance. A few very recent works [53–56] also look at the effect of attacks, but for the simpler unstructured multi-armed bandit problem [59]. Accounting for adversarial agents in the structured linear bandit setting we consider here requires significantly different ideas that we develop in this paper.

• **Security in Distributed Optimization and Federated Learning.** As we mentioned earlier, several papers have studied the problem of accounting for adversarial agents in the context of supervised learning [4–10, 62, 11–13, 63–65, 14, 66]. One of the primary applications of interest here is the emerging paradigm of federated learning [1–3]. Different from the sequential decision-making setting we investigate in our paper, the aforementioned works essentially abstract out the supervised learning task as a *static* distributed optimization problem, and then apply some form of secure aggregation on either gradient vectors or parameter estimates.

• **Robust Statistics.** The algorithms that we develop in this paper borrow tools from the literature on robust statistics, pioneered by Huber [67, 68]. We point the reader to [18, 19, 69–71, 22], and the references therein, to get a sense of some of the main results in this broad area of research. In a nutshell, given multiple samples of a random variable - with a small fraction of samples corrupted by an adversary - the essential goal of this line of work is to come up with statistically optimal and computationally efficient robust estimators of the mean of the random variable. Notably, unlike both the sequential bandit setting and the iterative optimization setting, the robust statistics literature focuses on one-shot estimation. In other words, the adversary gets to corrupt the batch of samples *only once*, and the effect of such corruption does not compound over time or iterations.

B Further Comments on our Algorithms

In what follows, we comment on certain key aspects of our proposed algorithms.

• **On the knowledge of the corruption fraction.** In practice, if we do not know the corruption fraction α, but have access to an upper bound on α, say $\tilde{\alpha}$, we can essentially use $\tilde{\alpha}$ as a proxy for α, both in our algorithms and their analyses. In this case, all our results go through identically by simply replacing α with $\tilde{\alpha}$ in the bounds. However, if we have no idea at all about α, then it is not clear whether one can design a robust algorithm that achieves minimax-optimal rates for the setting we consider in this paper. The main reason is as follows. Note that although computing a median does not require knowledge of α, the robust confidence threshold we use in line 9 of Algorithm 1 critically relies on the knowledge of the corruption fraction α (or an upper-bound on it). Moreover, the correctness of our overall approach, and the fact that it is minimax-optimal, relies heavily on the tightness of the robust confidence thresholds in Lemma 1. Thus, at the moment, we do not know of a way that can achieve tight regret bounds without any knowledge whatsoever of α; investigating this aspect further is an interesting topic of future research.

• **Beyond Gaussian Noise.** In what follows, we explain in detail that *both our algorithms and analysis apply, with slight modifications, to both sub-Gaussian noise, and more generally, noise with bounded variance. The reason why we chose Gaussian noise was primarily for ease of exposition.* Suppose the noise samples are i.i.d. with zero-mean (a standard assumption) and unit variance (the argument we present next trivially extends to bounded variance σ^2). To deal with such general noise distributions, we need to make two minor changes to Algorithm 1. First, we replace the Median operation in line 8 of Algorithm 1 with the scalar univariate trimmed mean estimator recently proposed in [71]. The estimator in [71] is *easily implementable and minimax-optimal*. Second, we set the robust threshold γ_ℓ in line 9 to be $\gamma_\ell = C(1 + \sqrt{\alpha M})\epsilon_\ell$, where C is a suitably large universal constant; note that the only thing that has changed from before is the replacement of α by $\sqrt{\alpha}$ in γ_ℓ. We will justify the choice of this threshold shortly, but before that, we mention the implications.

Implications. With the two minor modifications to Algorithm 1 described above, we can establish an analogue of Theorem 1 where for each good agent, with probability at least $1 - \delta$, the regret is bounded above by $\tilde{O}\left(\left(\sqrt{\alpha} + \sqrt{1/M}\right)\sqrt{dT}\right)$. Compared to the bound in Theorem 1, we note that our new bound is worsened by the replacement of α with $\sqrt{\alpha}$ - this is the price paid to account for general non-Gaussian distributions. For robust mean estimation with general noise distributions (with bounded fourth moments), the additive $\sqrt{\alpha}$ factor is *unavoidable*; see [71], for instance. Thus, we conjecture that the bound on regret we mentioned above is also minimax-optimal.

For sub-Gaussian distributions, we can achieve a tighter regret bound: by setting the robust confidence threshold to be $\gamma_\ell = C(1 + \alpha\sqrt{\log(1/\alpha)M})\epsilon_\ell$, we can achieve a per-agent regret bound of $\tilde{O}\left(\left(\alpha\sqrt{\log(1/\alpha)} + \sqrt{1/M}\right)\sqrt{dT}\right)$. *Hence, with sub-Gaussian noise, we essentially recover the same bounds as with Gaussian noise (up to logarithmic factors).* It is very likely that with sub-Gaussian noise, the median will continue to yield the bound above, i.e., we may not even need the univariate trimmed mean estimator from [71]. However, we do not have a concrete proof of this fact yet.

Changes in Analysis. We now go over the minor changes that need to be made to the analysis of Theorem 1 in view of replacing the median operator by the trimmed mean estimator in [71]. We first need an analog of Lemma 2 in Appendix C. This is supplied by Theorem 1 in [71] that provides guarantees on the univariate trimmed mean estimator. Suppose we are given M data samples, where all the good samples are i.i.d. with mean μ and variance σ^2. Moreover, suppose the fraction of bad samples is at most $1/2$. With probability at least $1 - \delta$, we then have

$$|\hat{\mu} - \mu| \leq C\left(\sqrt{\alpha} + \sqrt{\frac{\log(1/\delta)}{M}}\right)\sigma,$$

where $\hat{\mu}$ is the output of the univariate trimmed mean estimator. Equipped with this result, one can follow the *exact same steps* as in the proof of Lemma 1 to justify the choice of the robust confidence threshold $\gamma_\ell = C(1 + \sqrt{\alpha M})\epsilon_\ell$. When the noise is additionally sub-Gaussian, the above bound can be tightened by replacing $\sqrt{\alpha}$ with $\alpha\sqrt{\log(1/\alpha)}$. The rest of the proof goes through *identically*.

To sum up, we have argued that with minor modifications to both our algorithm and analysis, our overall approach can handle both sub-Gaussian noise, and noise with bounded variance.

• **Is it possible to tolerate a corruption fraction** $\alpha > 0.5$? The key question here is: *Who is the learner? The agent or the server?* As we explain shortly, the answer to this question has significant implications for whether or not we can tolerate the $\alpha > 0.5$ case. Suppose the learners are the agents, in that each good agent is capable of taking their own decisions (actions). If it is known ahead of time that $\alpha > 0.5$, then acting alone is the most natural thing to do. This is because there is nothing to be gained by collaborating: no robust aggregation scheme can provide any guarantees when $\alpha > 0.5$; indeed, $\alpha = 0.5$ is a *fundamental breakdown point* when one considers an *arbitrary* corruption model. On the other hand, when $\alpha < 0.5$, one can significantly improve upon the trivial per-agent regret of $O(\sqrt{dT})$ by using the approach developed in our paper. This is revealed not only by our theory, but the simulations that we report in Section 7.

Now suppose the learner is the server, i.e., *the actions are decided by the server*, and the agents interact with an unknown environment and report certain relevant statistics to the server. As a concrete example, consider a web-advertising example where the server displays ads (actions) to a group of people (agents), assesses their (potentially corrupted) feedback, and then decides upon subsequent ad displays to maximize click-through rates. Crucially, unlike the previous setting (where the learners were the agents), the only way the server can acquire feedback about the environment is by interacting with the agents (some of whom might be adversarial). *In other words, the server does not have at its disposal the trivial option of not interacting with the agents, and acting alone*. Since in this setting, the server is forced to interact with potentially corrupted agents, $\alpha < 0.5$ is the only scenario that can lead to meaningful regret bounds. More precisely, when $\alpha > 0.5$, there is no hope of achieving sub-linear regret, let alone benefiting from collaboration.

C Analysis of RCLB: Proof of Theorem 1

In this section, we will prove Theorem 1. We start with a standard result from robust statistics on the guarantees afforded by the median operator for robust mean estimation of univariate Gaussian random variables; see, for instance, [19].

Lemma 2. *Consider a set $\mathcal{S} = \{x_1, \ldots, x_M\}$ of M samples partitioned as $\mathcal{S} = \mathcal{S}_g \cup \mathcal{S}_b$, such that (i) all the samples in $\mathcal{S}_g$ are drawn i.i.d. from $\mathcal{N}(\mu, \sigma^2)$, where $\mu, \sigma^2 \in \mathbb{R}$; (ii) the samples in $\mathcal{S}_b$ are chosen by an adversary, and can be arbitrary; and (iii) $|\mathcal{S}_b| < \alpha|\mathcal{S}|$, where $\alpha < 1/2$. Let $\hat{\mu} = \texttt{Median}(\{x_i\}, i \in [M])$. Given any $\delta \in (0,1)$, we then have that with probability at least $1-\delta$,*

$$|\hat{\mu} - \mu| \leq C\left(\alpha + \sqrt{\frac{\log(\frac{1}{\delta})}{M}}\right)\sigma, \tag{15}$$

where C is a suitably large universal constant.

The next key lemma - a restatement of Lemma 1 in the main body of the paper - informs us about the quality of the robust mean payoffs computed in line 8 of Algorithm 1. Before proceeding to prove this result, we define by $\mathcal{F}_\ell$ the σ-algebra generated by all the actions and rewards up to the beginning of epoch ℓ.

Lemma 3. ***(Robust Confidence Intervals)*** *Fix any epoch ℓ. For each active arm $b \in \mathcal{A}_\ell$, the following holds with probability at least $1 - \delta_\ell$:*

$$|\mu_b^{(\ell)} - \langle \theta_*, b\rangle| \leq \gamma_\ell, \quad \textit{where} \quad \gamma_\ell = \sqrt{2}C\left(1 + \alpha\sqrt{M}\right)\epsilon_\ell, \tag{16}$$

where C is as in Lemma 2.

Proof. Fix an epoch ℓ, an active arm $b \in \mathcal{A}_\ell$, and a good agent $i \in [M] \setminus \mathcal{B}$. We start by analyzing the statistics of the quantity $\langle \hat{\theta}_i^{(\ell)}, b\rangle$. From the definition of $\hat{\theta}_i^{(\ell)}$ and $\tilde{V}_\ell$ in Eq.(3), we have

$$\begin{aligned}
\hat{\theta}_i^{(\ell)} &= \tilde{V}_\ell^{-1} Y_{i,\ell} \\
&= \tilde{V}_\ell^{-1}\left(\sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} r_{i,a}^{(\ell)} a\right) \\
&= \tilde{V}_\ell^{-1}\left(\sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} \left(\langle \theta_*, a\rangle + \bar{\eta}_{i,a}^{(\ell)}\right) a\right) \\
&= \theta_* + \tilde{V}_\ell^{-1}\left(\sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} \bar{\eta}_{i,a}^{(\ell)} a\right).
\end{aligned} \tag{17}$$

For the third equality above, we used the observation model (1), and denoted by $\bar{\eta}_{i,a}^{(\ell)}$ the average of the noise terms associated with the rewards observed by agent i during phase ℓ for arm a. From (17), we then have

$$\langle \hat{\theta}_i^{(\ell)}, b\rangle = \langle \theta_*, b\rangle + \sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} \bar{\eta}_{i,a}^{(\ell)} \langle \tilde{V}_\ell^{-1} a, b\rangle. \tag{18}$$

Now conditioned on $\mathcal{F}_\ell$, the only randomness in the above equation corresponds to the noise terms $\{\bar{\eta}_{i,a}^{(\ell)}\}_{a \in \text{Supp}(\pi_\ell)}$. Furthermore, based on our noise model, it is clear that $\bar{\eta}_{i,a}^{(\ell)} \sim \mathcal{N}(0, 1/m_a^{(\ell)})$ for each $a \in \text{Supp}(\pi_\ell)$. It then follows that

$$\mathbb{E}\left[\langle \hat{\theta}_i^{(\ell)}, b\rangle | \mathcal{F}_\ell\right] = \langle \theta_*, b\rangle.$$

We also have

$$\begin{aligned}\mathbb{E}\left[\left(\langle\hat{\theta}_i^{(\ell)}-\theta_*,b\rangle\right)^2|\mathcal{F}_\ell\right] &= \sum_{a\in\mathsf{Supp}(\pi_\ell)} m_a^{(\ell)}\left(\langle\tilde{V}_\ell^{-1}a,b\rangle\right)^2 \\ &= b'\tilde{V}_\ell^{-1}\left(\sum_{a\in\mathsf{Supp}(\pi_\ell)} m_a^{(\ell)}aa'\right)\tilde{V}_\ell^{-1}b \\ &= \|b\|^2_{\tilde{V}_\ell^{-1}},\end{aligned} \tag{19}$$

where we used the fact that the noise terms are independent across arms. We conclude that conditioned on $\mathcal{F}_\ell$,

$$\langle\hat{\theta}_i^{(\ell)},b\rangle \sim \mathcal{N}\left(\langle\theta_*,b\rangle, \|b\|^2_{\tilde{V}_\ell^{-1}}\right).$$

In each epoch ℓ, the server has access to a set $\mathcal{S}_b^{(\ell)} = \{\langle\hat{\theta}_i^{(\ell)},b\rangle\}_{i\in[M]}$, where the samples corresponding to agents in $[M]\setminus\mathcal{B}$ are independent and identically distributed as per the distribution above. Moreover, at most $\alpha\in[0,1/2)$ fraction of the samples in $\mathcal{S}_b^{(\ell)}$ are corrupted. Recalling that $\mu_b^{(\ell)} = \texttt{Median}\left(\{\langle\hat{\theta}_i^{(\ell)},b\rangle, i\in[M]\}\right)$, and using Lemma 2, we immediately observe that conditioned on $\mathcal{F}_\ell$, with probability at least $1-\delta_\ell$,

$$|\mu_b^{(\ell)} - \langle\theta_*,b\rangle| \le C\left(\alpha + \sqrt{\frac{\log(\frac{1}{\delta_\ell})}{M}}\right)\|b\|_{\tilde{V}_\ell^{-1}}. \tag{20}$$

We now proceed to bound the term $\|b\|_{\tilde{V}_\ell^{-1}}$. To that end, let us start by noting that

$$\begin{aligned}\tilde{V}_\ell &= \sum_{a\in\mathsf{Supp}(\pi_\ell)} m_a^{(\ell)}aa' \\ &= \sum_{a\in\mathsf{Supp}(\pi_\ell)} \left\lceil\frac{T_a^{(\ell)}}{M}\right\rceil aa' \\ &\succcurlyeq \frac{1}{M}\sum_{a\in\mathsf{Supp}(\pi_\ell)} T_a^{(\ell)}aa' \\ &\succcurlyeq \frac{d}{M\epsilon_\ell^2}\log\left(\frac{1}{\delta_\ell}\right)\sum_{a\in\mathsf{Supp}(\pi_\ell)}\pi_\ell(a)aa' \\ &= \frac{d}{M\epsilon_\ell^2}\log\left(\frac{1}{\delta_\ell}\right)\sum_{a\in\mathcal{A}_\ell}\pi_\ell(a)aa' \\ &= \frac{d}{M\epsilon_\ell^2}\log\left(\frac{1}{\delta_\ell}\right)V_\ell(\pi_\ell).\end{aligned} \tag{21}$$

Thus, we have

$$\tilde{V}_\ell^{-1} \preccurlyeq \frac{M\epsilon_\ell^2}{d\log\left(\frac{1}{\delta_\ell}\right)}V_\ell^{-1}(\pi_\ell).$$

Using the above bound, we proceed as follows.

$$
\begin{aligned}
\|b\|_{\tilde{V}_\ell^{-1}} &= \sqrt{b' \tilde{V}_\ell^{-1} b} \\
&\leq \epsilon_\ell \sqrt{\frac{M}{d \log\left(\frac{1}{\delta_\ell}\right)}} \sqrt{b' V_\ell^{-1}(\pi_\ell) b} \\
&\leq \epsilon_\ell \sqrt{\frac{M}{d \log\left(\frac{1}{\delta_\ell}\right)}} \sqrt{\max_{a \in \mathcal{A}_\ell} \|a\|^2_{V_\ell^{-1}(\pi_\ell)}} \\
&\overset{(a)}{=} \epsilon_\ell \sqrt{\frac{M}{d \log\left(\frac{1}{\delta_\ell}\right)}} \sqrt{g_\ell(\pi_\ell)} \\
&\overset{(b)}{\leq} \epsilon_\ell \sqrt{\frac{2M}{\log\left(\frac{1}{\delta_\ell}\right)}}.
\end{aligned}
\tag{22}
$$

In the above steps, we used the definition of $g_\ell(\pi_\ell)$ for (a); for (b), we used the fact that based on the approximate G-optimal design problem solved by the server in line 1 of RCLB, $g_\ell(\pi_\ell) \leq 2d$. Plugging the bound from (22) into (20), and using the fact that $\log\left(\frac{1}{\delta_\ell}\right) \geq 1$, we have that

$$
\mathbb{P}\left(|\mu_b^{(\ell)} - \langle \theta_*, b \rangle| \geq \gamma_\ell | \mathcal{F}_\ell\right) \leq \delta_\ell. \tag{23}
$$

Consider the following event $\mathcal{E}_\ell \triangleq \{|\mu_b^{(\ell)} - \langle \theta_*, b \rangle| \geq \gamma_\ell\}$. Now observe that

$$
\begin{aligned}
\mathbb{P}(\mathcal{E}_\ell) &= \mathbb{E}[\mathbf{1}_{\mathcal{E}_\ell}] \\
&= \mathbb{E}\left[\mathbb{E}[\mathbf{1}_{\mathcal{E}_\ell} | \mathcal{F}_\ell]\right] \\
&= \mathbb{E}\left[\mathbb{P}(\mathcal{E}_\ell | \mathcal{F}_\ell)\right] \\
&\leq \delta_\ell,
\end{aligned}
\tag{24}
$$

where we used $\mathbf{1}_{\mathcal{E}_\ell}$ to denote an indicator random variable associated with the event $\mathcal{E}_\ell$; also, for the last line, we used (23). □

In the following two results, we use the robust confidence intervals from Lemma 3 to construct clean events that hold with high probability on which (i) the optimal arm a_* is never eliminated (Lemma 4); and (ii) any arm retained in epoch ℓ contributes at most $O(\gamma_\ell)$ regret in each time-step within epoch ℓ (Lemma 5). To proceed, for each $a \in \mathcal{A}$, define the arm-gap $\Delta_a = \langle \theta_*, a_* - a \rangle$.

Lemma 4. *Define the event $\mathcal{G}_1 \triangleq \{a_* \in \mathcal{A}_\ell, \forall \ell \in [L]\}$, where L is the total number of epochs. It then holds that $\mathbb{P}(\mathcal{G}_1) \geq 1 - 4\bar{\delta}$.*

Proof. Based on the arm-elimination criterion in line 9 of Algorithm 1, it follows that $\{a_* \in \mathcal{A}_\ell, a_* \notin \mathcal{A}_{\ell+1}\} \implies \{\exists b \in \mathcal{A}_\ell : \mu_b^{(\ell)} - \mu_{a_*}^{(\ell)} > 2\gamma_\ell\}$. Now for any fixed $b \in \mathcal{A}_\ell$, we have

$$
\begin{aligned}
&\mu_b^{(\ell)} - \mu_{a_*}^{(\ell)} > 2\gamma_\ell \\
&\implies \left(\mu_b^{(\ell)} - \langle \theta_*, b \rangle\right) + \left(\langle \theta_*, a_* \rangle - \mu_{a_*}^{(\ell)}\right) > 2\gamma_\ell + \Delta_b \\
&\implies \left(\mu_b^{(\ell)} - \langle \theta_*, b \rangle\right) + \left(\langle \theta_*, a_* \rangle - \mu_{a_*}^{(\ell)}\right) > 2\gamma_\ell,
\end{aligned}
\tag{25}
$$

where for the second step, we used the fact that $\Delta_b \geq 0$. Thus, the event $\{\mu_b^{(\ell)} - \mu_{a_*}^{(\ell)} > 2\gamma_\ell\}$ implies the occurrence of either $\{\mu_b^{(\ell)} - \langle \theta_*, b \rangle > \gamma_\ell\}$ or $\{\langle \theta_*, a_* \rangle - \mu_{a_*}^{(\ell)} > \gamma_\ell\}$. From Lemma 3, we further know that the probability of each of these latter events is at most δ_ℓ. Putting these pieces together,

and using an union bound, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{G}_1^c) &\leq \sum_{\ell \in [L]} \mathbb{P}(a^* \in \mathcal{A}_\ell, a^* \notin \mathcal{A}_{\ell+1}) \\
&\leq \sum_{\ell \in [L]} \mathbb{P}(\exists b \in \mathcal{A}_\ell : \mu_b^{(\ell)} - \mu_{a_*}^{(\ell)} > 2\gamma_\ell) \\
&\leq 2K \sum_{\ell \in [L]} \delta_\ell \\
&= 2K \sum_{\ell \in [L]} \frac{\bar{\delta}}{K\ell^2} \\
&\leq 2 \sum_{\ell=1}^{\infty} \frac{\bar{\delta}}{\ell^2} \\
&\leq 2\bar{\delta} \int_{x=1}^{\infty} \frac{1}{x^2}\, dx \leq 4\bar{\delta}.
\end{aligned} \tag{26}$$

This completes the proof. □

In our next result, we work towards bounding the regret incurred from playing each active arm in a given epoch.

Lemma 5. *Consider any arm $a \in \mathcal{A} \setminus \{a_*\}$. Let ℓ_a be defined as $\ell_a \triangleq \min\{\ell : \gamma_\ell < \frac{\Delta_a}{4}\}$. It then holds that $\mathbb{P}(a \in \mathcal{A}_{\ell_a+1}) \leq 6\bar{\delta}$.*

Proof. Let us start by observing that

$$\begin{aligned}
\mathbb{P}(a \in \mathcal{A}_{\ell_a+1}) &= \mathbb{P}(a \in \mathcal{A}_{\ell_a+1}, a^* \in \mathcal{A}_{\ell_a}) + \mathbb{P}(a \in \mathcal{A}_{\ell_a+1}, a^* \notin \mathcal{A}_{\ell_a}) \\
&\leq \mathbb{P}(a \in \mathcal{A}_{\ell_a+1}, a^* \in \mathcal{A}_{\ell_a}) + \mathbb{P}(a^* \notin \mathcal{A}_{\ell_a}) \\
&\leq \mathbb{P}(a \in \mathcal{A}_{\ell_a}, a \in \mathcal{A}_{\ell_a+1}, a^* \in \mathcal{A}_{\ell_a}) + 4\bar{\delta},
\end{aligned} \tag{27}$$

where for the last step, we used the fact that $\{a \in \mathcal{A}_{\ell_a+1}\} \implies \{a \in \mathcal{A}_{\ell_a}\}$, and Lemma 4. Now, to bound $\mathbb{P}(a \in \mathcal{A}_{\ell_a}, a \in \mathcal{A}_{\ell_a+1}, a^* \in \mathcal{A}_{\ell_a})$, we note based on line 9 of RCLB that

$$\begin{aligned}
\mathbb{P}(a \in \mathcal{A}_{\ell_a}, a \in \mathcal{A}_{\ell_a+1}, a^* \in \mathcal{A}_{\ell_a}) &\leq \mathbb{P}\left(\max_{b \in \mathcal{A}_{\ell_a}} \mu_b^{(\ell_a)} - \mu_a^{(\ell_a)} \leq 2\gamma_{\ell_a}\right) \\
&\leq \mathbb{P}\left(\mu_{a_*}^{(\ell_a)} - \mu_a^{(\ell_a)} \leq 2\gamma_{\ell_a}\right) \\
&\leq \mathbb{P}\left(\left(\mu_a^{(\ell_a)} - \langle \theta_*, a \rangle\right) + \left(\langle \theta_*, a_* \rangle - \mu_{a_*}^{(\ell_a)}\right) > \Delta_a - 2\gamma_{\ell_a}\right) \\
&\leq \mathbb{P}\left(\mu_a^{(\ell_a)} - \langle \theta_*, a \rangle > \gamma_{\ell_a}\right) + \mathbb{P}\left(\langle \theta_*, a_* \rangle - \mu_{a_*}^{(\ell_a)} > \gamma_{\ell_a}\right) \\
&\leq 2\delta_{\ell_a},
\end{aligned} \tag{28}$$

where we used $\Delta_a > 4\gamma_{\ell_a}$ for the second last step, and Lemma 3 for the last step. Noting that $\delta_{\ell_a} \leq \bar{\delta}$, and combining the bounds in equations (27) and (28) leads to the claim of the lemma. □

We are now in place to prove Theorem 1.

Proof. (**Proof of Theorem 1**) We start by constructing an appropriate clean event $\mathcal{E}$ for our subsequent analysis. Accordingly, let us define:

$$\mathcal{E} = \{a_* \in \mathcal{A}_\ell, \forall \ell \in [L]\} \bigcap \{\cap_{a \in \mathcal{A} \setminus \{a_*\}} \{a \notin \mathcal{A}_{\ell_a+1}\}\}. \tag{29}$$

Based on Lemmas 4 and 5, we then have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}^c) &\leq 4\bar{\delta} + \sum_{a \in \mathcal{A} \setminus \{a_*\}} \mathbb{P}(a \in \mathcal{A}_{\ell_a+1}) \\
&\leq 4\bar{\delta} + 6\bar{\delta}K \\
&\leq 10\bar{\delta}K = \delta,
\end{aligned} \tag{30}$$

as per the choice of $\bar{\delta}$ in Theorem 1. Thus, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$. Throughout the rest of the proof, we will condition on the clean event $\mathcal{E}$. Based on the definition of the event $\mathcal{E}$, it is easy to see that for any epoch $\ell \in [L]$, $a \in \mathcal{A}_\ell \implies \Delta_a \leq 8\gamma_\ell$. Using this key fact, we now proceed to bound the regret of any good agent $i \in [M] \setminus \mathcal{B}$.

$$\begin{aligned}
\sum_{t=1}^{T} \langle \theta_*, a_* - a_{i,t} \rangle &= \sum_{\ell=1}^{L} \sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} \langle \theta_*, a_* - a \rangle \\
&= \sum_{\ell=1}^{L} \sum_{a \in \text{Supp}(\pi_\ell)} \left\lceil \frac{T_a^{(\ell)}}{M} \right\rceil \Delta_a \\
&\leq \underbrace{\sum_{\ell=1}^{L} \sum_{a \in \text{Supp}(\pi_\ell)} \frac{T_a^{(\ell)}}{M} \Delta_a}_{T_1} + \underbrace{\sum_{\ell=1}^{L} \sum_{a \in \text{Supp}(\pi_\ell)} \Delta_a}_{T_2} .
\end{aligned} \tag{31}$$

We now bound T_1 and T_2 separately. For bounding T_1, we have:

$$\begin{aligned}
T_1 &= \sum_{\ell=1}^{L} \sum_{a \in \text{Supp}(\pi_\ell)} \frac{T_a^{(\ell)}}{M} \Delta_a \\
&= \frac{d}{M} \sum_{\ell=1}^{L} \frac{1}{\epsilon_\ell^2} \log\left(\frac{1}{\delta_\ell}\right) \sum_{a \in \text{Supp}(\pi_\ell)} \pi_\ell(a) \Delta_a \\
&\leq \frac{8d}{M} \sum_{\ell=1}^{L} \frac{1}{\epsilon_\ell^2} \log\left(\frac{1}{\delta_\ell}\right) \gamma_\ell \\
&= \frac{8\sqrt{2}C(1+\alpha\sqrt{M})d}{M} \sum_{\ell=1}^{L} \frac{1}{\epsilon_\ell} \log\left(\frac{10K^2\ell^2}{\delta}\right) \\
&\leq \frac{8\sqrt{2}C(1+\alpha\sqrt{M})d}{M} \log\left(\frac{10K^2L^2}{\delta}\right) \sum_{\ell=1}^{L} 2^\ell \\
&= O\left(\frac{(1+\alpha\sqrt{M})d}{M} \log\left(\frac{10K^2L^2}{\delta}\right) 2^L\right).
\end{aligned} \tag{32}$$

In the third step above, we used $\sum_{a \in \text{Supp}(\pi_\ell)} \pi_\ell(a) = 1$. We now need an upper-bound on the term 2^L. To that end, notice that the length of the horizon T is bounded below by the length of the last epoch, i.e., the L-th epoch. Moreover, the duration of the L-th epoch corresponds to the number of arm-pulls made by any single good agent during the L-th epoch. We thus have:

$$\begin{aligned}
T &\geq \sum_{a \in \text{Supp}(\pi_L)} \frac{T_a^{(L)}}{M} \\
&\geq \frac{4^L d}{M} \sum_{a \in \text{Supp}(\pi_L)} \pi_L(a) \log\left(\frac{1}{\delta_L}\right) \\
&= \frac{4^L d}{M} \log\left(\frac{10K^2L^2}{\delta}\right) \sum_{a \in \text{Supp}(\pi_L)} \pi_L(a) \\
&= \frac{4^L d}{M} \log\left(\frac{10K^2L^2}{\delta}\right).
\end{aligned} \tag{33}$$

Thus, $2^L \leq \sqrt{MT/(d \log(10K^2L^2/\delta))}$. Plugging this bound in (32), we obtain

$$T_1 = O\left(\left(\alpha + \frac{1}{\sqrt{M}}\right) \sqrt{\log\left(\frac{KT}{\delta}\right) dT}\right) = \tilde{O}\left(\left(\alpha + \frac{1}{\sqrt{M}}\right) \sqrt{dT}\right).$$

As for the term T_2, we have

$$
\begin{aligned}
T_2 &= \sum_{\ell=1}^{L} \sum_{a \in \text{Supp}(\pi_\ell)} \Delta_a \\
&\leq 8 \sum_{\ell=1}^{L} \gamma_\ell |\text{Supp}(\pi_\ell)| \\
&\overset{(a)}{\leq} 384\sqrt{2} C d \log\log d \left(1 + \alpha\sqrt{M}\right) \sum_{\ell=1}^{L} 2^{-\ell} \\
&= O\left(d \log\log d \left(1 + \alpha\sqrt{M}\right)\right) \\
&\overset{(b)}{=} \tilde{O}\left(\left(\alpha + \frac{1}{\sqrt{M}}\right)\sqrt{dT}\right).
\end{aligned}
\tag{34}
$$

In the above steps, for (a), recall from line 1 of RCLB that $|\text{Supp}(\pi_\ell)| \leq 48 d \log\log d$ based on the approximate G-optimal design computation. For (b), we used the fact that by assumption, $T \geq Md$. Combining the bounds on T_1 and T_2, and recalling that $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$, we have that with probability at least $1 - \delta$,

$$
\sum_{t=1}^{T} \langle \theta_*, a_* - a_{i,t} \rangle = O\left(\left(\alpha + \frac{1}{\sqrt{M}}\right)\sqrt{\log\left(\frac{KT}{\delta}\right) dT}\right) = \tilde{O}\left(\left(\alpha + \frac{1}{\sqrt{M}}\right)\sqrt{dT}\right). \tag{35}
$$

This concludes the proof. □

The following is an immediate corollary of Theorem 1 on the group regret R_T.

Corollary 1. ***(Bound on Group Regret)*** *Under the conditions of Theorem 1, we have:*

$$
R_T = \tilde{O}\left(\left(\alpha M + \sqrt{M}\right)\sqrt{dT}\right). \tag{36}
$$

Proof. (**Proof of Corollary 1**) Recall from the proof of Theorem 1 that there exists a clean event $\mathcal{E}$ of measure at least $1 - \delta$ on which the regret of every good agent is bounded above as per Eq. (35). Let $\tilde{C}$ be an upper bound on the maximum instantaneous regret, i.e.,

$$
\max_{a \in \mathcal{A}} \langle \theta_*, a_* - a \rangle \leq \tilde{C}.
$$

Now set $\delta = \frac{1}{MT}$ and observe that:

$$
\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{i \in [M] \setminus \mathcal{B}} \sum_{t=1}^{T} \langle \theta_*, a_* - a_{i,t} \rangle\right] \\
&= \mathbb{E}\left[\sum_{i \in [M] \setminus \mathcal{B}} \sum_{t=1}^{T} \langle \theta_*, a_* - a_{i,t} \rangle \middle| \mathcal{E}\right] \mathbb{P}(\mathcal{E}) + \mathbb{E}\left[\sum_{i \in [M] \setminus \mathcal{B}} \sum_{t=1}^{T} \langle \theta_*, a_* - a_{i,t} \rangle \middle| \mathcal{E}^c\right] \mathbb{P}(\mathcal{E}^c) \\
&\leq C_1 \left(\alpha M + \sqrt{M}\right) \sqrt{\log\left(KMT\right) dT} + \tilde{C} MT \times \frac{1}{MT} \\
&= O\left(\left(\alpha M + \sqrt{M}\right)\sqrt{\log\left(KMT\right) dT}\right) \\
&= \tilde{O}\left(\left(\alpha M + \sqrt{M}\right)\sqrt{dT}\right).
\end{aligned}
\tag{37}
$$

In the above steps, C_1 is a suitably large universal constant. □

D Lower Bound Analysis: Proof of Theorem 2

In this section, we will prove Theorem 2. Before diving into the technical details, we remind the reader that we consider a slightly different adversarial model from the one considered throughout the paper. In this modified model, with probability α, each agent is adversarial independently of the other agents. We will consider a class of policies Π where at each time-step, the same action (decided by the server) is played by every agent. Thus, the regret incurred by any individual agent is the same as the regret incurred by the server. Finally, to prove the lower bound, we will focus on a class of 2-armed bandits where the reward distribution of each arm is Gaussian with unit variance; such a class of bandits is succinctly denoted by $\mathcal{E}_{\mathcal{N}}^{(2)}(1)$.

We will have occasion to use the following result [57, Theorem 14.2].

Lemma 6. *(**Bretagnolle-Huber Inequality**) Let P and Q be two probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be any arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp\left(-KL(P, Q)\right),$$

where A^c is the complement of the event A, and $KL(P, Q)$ is the Kullback-Leibler distance between P and Q.

Our proof comprises of two main steps. First, we construct two *distinct* bandit instances within the class $\mathcal{E}_{\mathcal{N}}^{(2)}(1)$ such that the two instances - although different - appear *identical* to the server. Second, we devise an attack strategy and argue that regardless of the policy played by the server, it will end up suffering a regret of $\Omega(\alpha\sqrt{T})$ upon interacting with at least one of the two instances; here, T is the horizon for our problem. We now elaborate on these two steps.

• **Step 1. Construction of the two bandit instances.** We first take a detour and describe an idea that is typically used to prove lower bounds for the robust mean estimation literature [18, 19]. It will soon be apparent how such an idea can be exploited to construct the two bandit instances for our problem. We show that there are two univariate Gaussian distributions $P_1 = \mathcal{N}(\mu_1, 1)$, $P_2 = \mathcal{N}(\mu_2, 1)$, and two T-dimensional distributions Q_1, Q_2, such that $\|\mu_1 - \mu_2\|_2 = \Omega(\alpha/\sqrt{T})$, and:

$$(1-\alpha)P_1^T + \alpha Q_1 = (1-\alpha)P_2^T + \alpha Q_2, \tag{38}$$

where P_1^T (resp., P_2^T) is the joint distribution of T i.i.d. samples drawn from P_1 (resp., P_2). Clearly, P_1^T (resp., P_2^T) is equivalent to a T-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_1, I_T)$ (resp., $\mathcal{N}(\boldsymbol{\mu}_2, I_T)$), where $\boldsymbol{\mu}_1$ (resp., $\boldsymbol{\mu}_2$) is a T-dimensional vector with each entry equal to μ_1 (resp., μ_2). Let ϕ_1 be the p.d.f. of P_1^T and ϕ_2 be the p.d.f. of P_2^T. Next, let μ_1 and μ_2 be chosen such that the total variation distance $\delta(P_1^T, P_2^T)$ between P_1^T and P_2^T is

$$\frac{1}{2}\int \|\phi_1 - \phi_2\|_1 dx = \frac{\alpha}{1-\alpha}.$$

Let Q_1 be the distribution with p.d.f. $\frac{1-\alpha}{\alpha}(\phi_2 - \phi_1)\mathbf{1}_{\phi_2 \geq \phi_1}$, and Q_2 be the distribution with p.d.f. $\frac{1-\alpha}{\alpha}(\phi_1 - \phi_2)\mathbf{1}_{\phi_1 \geq \phi_2}$. With such a construction of Q_1 and Q_2, one can verify that the equality in Eq. (38) is satisfied; see, for instance, the arguments in Appendix E of [18]. Now from Pinsker's inequality, we know that:

$$\sqrt{\frac{1}{2}\mathrm{KL}(P_1^T, P_2^T)} \geq \delta(P_1^T, P_2^T) = \frac{\alpha}{1-\alpha},$$

where we used $\mathrm{KL}(P_1^T, P_2^T)$ to denote the Kullback-Leibler distance between P_1^T and P_2^T. We conclude that:

$$\frac{1}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 = \mathrm{KL}(P_1^T, P_2^T) \geq 2\left(\frac{\alpha}{1-\alpha}\right)^2.$$

This, in turn, implies

$$\sqrt{T}\left(|\mu_1 - \mu_2|\right) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2 \geq \frac{2\alpha}{1-\alpha} \geq 2\alpha.$$

We have thus shown that there exist $\mu_1, \mu_2 \in \mathbb{R}$, satisfying $|\mu_1 - \mu_2| \geq 2\alpha/\sqrt{T}$, such that with $P_1 = \mathcal{N}(\mu_1, 1)$, $P_2 = \mathcal{N}(\mu_2, 1)$, one can satisfy Eq. (38) with appropriately chosen T-dimensional distributions Q_1 and Q_2. With these ideas in place, we now return to our bandit problem.

Without loss of generality, suppose $\mu_2 > \mu_1$, where μ_2 and μ_1 are as in the construction described above. Let us construct two bandit instances ν and ν', each involving two arms, i.e., $\mathcal{A} = \{a_1, a_2\}$. Let $r_{a_k}^{(\nu)}$ and $r_{a_k}^{(\nu')}$ denote the reward distribution of arm $a_k, k \in \{1, 2\}$, in instance ν and instance ν', respectively. These reward distributions are chosen as follows.

$$\begin{aligned} &\text{Instance } \nu : r_{a_1}^{(\nu)} \sim \mathcal{N}\left(\frac{\mu_1 + \mu_2}{2}, 1\right); \;\; r_{a_2}^{(\nu)} \sim \mathcal{N}(\mu_1, 1) \\ &\text{Instance } \nu' : r_{a_1}^{(\nu')} \sim \mathcal{N}\left(\frac{\mu_1 + \mu_2}{2}, 1\right); \;\; r_{a_2}^{(\nu')} \sim \mathcal{N}(\mu_2, 1). \end{aligned} \tag{39}$$

Thus, the distribution of the first arm is the same in both instances. However, as $\mu_2 > \mu_1$, the first arm is the best arm in the first instance while the opposite is true for the second instance. The attack strategy for the adversarial agents will be dictated by the distributions Q_1 and Q_2 in a manner to be described shortly.

• **Step 2. The attack strategy and regret analysis.** Inspired by the argument in the proof of [17, Theorem 5], we consider a full information setting where the server has access to t reward samples from *each* arm from *each* agent at time-step t. Since this full information setting is simpler than the bandit setting, a lower bound for the former implies one for the latter.

Here is the attack strategy. Suppose an agent is adversarial (which happens with probability α). In either instance, for arm 1, it reports samples from the true distribution for arm 1 corresponding to that instance. In other words, the reward samples for arm 1 are not corrupted by the agent. As for arm 2, in instance ν (resp., ν'), the first t reward samples (where $t \in [T]$) corresponding to a_2 are generated from Q_1^t (resp., Q_2^t) by the adversarial agent. Here, Q_1^t (resp., Q_2^t) is the marginal of the T-dimensional distribution Q_1 (resp., Q_2) corresponding to the first t components. To sum up, in instance ν, the joint distribution $D_{a_1}^{(\nu)}$ of rewards for a_1 over the horizon T, as seen by the server from any given agent, is the T-dimensional Gaussian distribution $\mathcal{N}(\bar{\boldsymbol{\mu}}, I_T)$, where $\bar{\boldsymbol{\mu}}$ is a T-dimensional vector with each component equal to $\frac{\mu_1+\mu_2}{2}$. Based on our discussion above, $D_{a_1}^{(\nu)} = D_{a_1}^{(\nu')}$. Let $D_{a_2}^{(\nu)}$ and $D_{a_2}^{(\nu')}$ have analogous meanings for arm a_2. Then, we have:

$$D_{a_2}^{(\nu)} = (1-\alpha)P_1^T + \alpha Q_1; \;\; D_{a_2}^{(\nu')} = (1-\alpha)P_2^T + \alpha Q_2. \tag{40}$$

In light of Eq. (38), however, we have $D_{a_2}^{(\nu)} = D_{a_2}^{(\nu')}$. Essentially, what we have established is the following: *the joint distribution of rewards for each arm over the horizon T, as seen by the server from each agent, is identical for both instances*.

In what follows, given any two distributions D_1 and D_2, let $D_1 \otimes D_2$ represent their product distribution. Since the rewards across arms are independent, the joint distribution of rewards for both arms is given by $D^{(\nu)} \triangleq D_{a_1}^{(\nu)} \otimes D_{a_2}^{(\nu)}$ in instance ν, and $D^{(\nu')} \triangleq D_{a_1}^{(\nu')} \otimes D_{a_2}^{(\nu')}$ in instance ν'. Since rewards across agents are independent, the joint distributions of rewards from *all* agents, as seen by the server in each of the two instances, are given by:

$$D_M^{(\nu)} \triangleq \underbrace{D^{(\nu)} \otimes D^{(\nu)} \otimes \cdots \otimes D^{(\nu)}}_{M\text{-fold product distribution}}; \;\; D_M^{(\nu')} \triangleq \underbrace{D^{(\nu')} \otimes D^{(\nu')} \otimes \cdots \otimes D^{(\nu')}}_{M\text{-fold product distribution}}.$$

Let $\mathbb{E}_\nu[\cdot]$ (resp., $\mathbb{E}_{\nu'}[\cdot]$) represent the expectation operation w.r.t. the measure $D_M^{(\nu)}$ (resp., $D_M^{(\nu')}$). Let us use $\mathbb{P}_\nu$ as a shorthand for $D_M^{(\nu)}$, and $\mathbb{P}_{\nu'}$ as a shorthand for $D_M^{(\nu')}$. Furthermore, let $n_k(T)$ be the random variable representing the total number of times arm $a_k, k \in \{1, 2\}$, is chosen by the server over the horizon T. Finally, recall that $R_T^{(s)}(\nu)$ (resp., $R_T^{(s)}(\nu')$) is the regret incurred by the server upon interaction with instance ν (resp., instance ν').

In instance ν, each time arm 2 is chosen by the server, it incurs an instantaneous regret of $(\mu_2 - \mu_1)/2$. We thus have:

$$R_T^{(s)}(\nu) = \left(\frac{\mu_2 - \mu_1}{2}\right)(T - \mathbb{E}_\nu[n_1(T)]) \geq \frac{(\mu_2 - \mu_1)T}{4}\mathbb{P}_\nu\left(n_1(T) \leq \frac{T}{2}\right). \tag{41}$$

To see why the latter inequality is true, observe:

$$\begin{aligned}\mathbb{E}_\nu[n_1(T)] &= \mathbb{E}_\nu\left[n_1(T)|n_1(T) \le \frac{T}{2}\right]\mathbb{P}_\nu\left(n_1(T) \le \frac{T}{2}\right) + \mathbb{E}_\nu\left[n_1(T)|n_1(T) > \frac{T}{2}\right]\mathbb{P}_\nu\left(n_1(T) > \frac{T}{2}\right) \\ &\le \frac{T}{2}\mathbb{P}_\nu\left(n_1(T) \le \frac{T}{2}\right) + T\left(1 - \mathbb{P}_\nu\left(n_1(T) \le \frac{T}{2}\right)\right) \\ &= T - \frac{T}{2}\mathbb{P}_\nu\left(n_1(T) \le \frac{T}{2}\right).\end{aligned} \tag{42}$$

In instance ν', each time arm 1 is chosen by the server, it incurs an instantaneous regret of $(\mu_2 - \mu_1)/2$. We thus have:

$$R_T^{(s)}(\nu') = \left(\frac{\mu_2 - \mu_1}{2}\right)\mathbb{E}_{\nu'}[n_1(T)] \ge \frac{(\mu_2 - \mu_1)T}{4}\mathbb{P}_{\nu'}\left(n_1(T) > \frac{T}{2}\right). \tag{43}$$

Combining Eq. (41) and Eq. (43) yields:

$$\begin{aligned}\max\{R_T^{(s)}(\nu), R_T^{(s)}(\nu')\} &\ge \frac{1}{2}\left(R_T^{(s)}(\nu) + R_T^{(s)}(\nu')\right) \\ &\ge \frac{(\mu_2 - \mu_1)T}{8}\left(\mathbb{P}_\nu\left(n_1(T) \le \frac{T}{2}\right) + \mathbb{P}_{\nu'}\left(n_1(T) > \frac{T}{2}\right)\right) \\ &\overset{(a)}{\ge} \frac{(\mu_2 - \mu_1)T}{16}\exp\left(-\mathrm{KL}\left(\mathbb{P}_\nu, \mathbb{P}_{\nu'}\right)\right) \\ &\overset{(b)}{\ge} \frac{\alpha\sqrt{T}}{8}\exp\left(-\mathrm{KL}\left(\mathbb{P}_\nu, \mathbb{P}_{\nu'}\right)\right) \\ &\overset{(c)}{=} \frac{\alpha\sqrt{T}}{8}.\end{aligned} \tag{44}$$

In the above steps, we used the Bretagnolle-Huber inequality (namely, Lemma 6) for (a); for (b), we used the fact that μ_2 and μ_1 were chosen in Step 1 to satisfy $\mu_2 - \mu_1 \ge 2\alpha/\sqrt{T}$; and for (c), we used $\mathrm{KL}\left(\mathbb{P}_\nu, \mathbb{P}_{\nu'}\right) = 0$. To see why $\mathrm{KL}\left(\mathbb{P}_\nu, \mathbb{P}_{\nu'}\right) = 0$, we use the chain-rule for relative entropies to obtain:

$$\begin{aligned}\mathrm{KL}\left(\mathbb{P}_\nu, \mathbb{P}_{\nu'}\right) &= \mathrm{KL}\left(D_M^{(\nu)}, D_M^{(\nu')}\right) \\ &= M\left(\mathrm{KL}\left(D^{(\nu)}, D^{(\nu')}\right)\right) \\ &= M\left(\mathrm{KL}\left(D_{a_1}^{(\nu)}, D_{a_1}^{(\nu')}\right) + \mathrm{KL}\left(D_{a_2}^{(\nu)}, D_{a_2}^{(\nu')}\right)\right) \\ &= 0,\end{aligned} \tag{45}$$

where the last step is a consequence of the fact that $D_{a_1}^{(\nu)} = D_{a_1}^{(\nu')}$, and $D_{a_2}^{(\nu)} = D_{a_2}^{(\nu')}$. The claim of Theorem 2 follows from noting that the resulting lower bound in Eq. (44) holds regardless of the number of agents M.

■

Remark 3. *The proof of Theorem 2 above reveals a constructive attack strategy for the adversarial agents to follow. As future work, it would be interesting to see if similar ideas can be extended to more general reinforcement learning problems.*

Algorithm 4 Robust Collaborative Phased Elimination for Generalized Linear Bandits (RC-GLM)

Input: Action set $\mathcal{A} = \{a_1, \ldots, a_K\}$, confidence parameter δ, and corruption fraction α.
Initialize: $\ell = 1$ and $\mathcal{A}_1 = \mathcal{A}$.

1: Let $V_\ell(\pi) \triangleq \sum_{a \in \mathcal{A}_\ell} \pi(a) aa'$ and $g_\ell(\pi) \triangleq \max_{a \in \mathcal{A}_\ell} \|a\|^2_{V_\ell(\pi)^{-1}}$. Server solves an approximate G-optimal design problem to compute a distribution π_ℓ over $\mathcal{A}_\ell$ such that $g_\ell(\pi_\ell) \leq 2d$ and $|\text{Supp}(\pi_\ell)| \leq 48d \log\log d$.

2: For each $a \in \mathcal{A}_\ell$, server computes $m_a^{(\ell)}$ via Eq. (5), and broadcasts $\{m_a^{(\ell)}\}_{a \in \mathcal{A}_\ell}$ to all agents.

3: **for** $i \in [M] \setminus \mathcal{B}$ **do**

4: For each arm $a \in \mathcal{A}_\ell$, pull it $m_a^{(\ell)}$ times. Let $r_{i,a}^{(\ell)}$ be the average of the rewards observed by agent i for arm a during phase ℓ.

5: Compute $\tilde{V}_\ell$ and $Y_{i,\ell}$ as follows.

$$\tilde{V}_\ell = \sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} aa' \; ; \; Y_{i,\ell} = \sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} r_{i,a}^{(\ell)} a.$$

6: Transmit $Y_{i,\ell}$ to server. Adversarial agents can transmit arbitrary vectors at this stage.

7: **end for**

8: Server computes a statistic X_ℓ as follows:

$$X_\ell = \texttt{ITW}(\{\tilde{V}_\ell^{-1/2} Y_{i,\ell}, i \in [M]\}),$$

where $\texttt{ITW}(\cdot)$ is the output of the Iteratively Reweighted Mean Estimator from [22].

9: Server computes a robust estimate $\hat{\theta}^{(\ell)}$ of θ_* by solving:

$$h_\ell(\hat{\theta}^{(\ell)}) = \tilde{V}_\ell^{1/2} X_\ell.$$

10: Define robust confidence threshold $\bar{\gamma}_\ell = 4C_1(k_2/k_1)\left(\sqrt{d} + \alpha\sqrt{M \log(1/\alpha)}\right)\epsilon_\ell$, where $\epsilon_\ell = 2^{-\ell}$, and C_1 is as in Lemma 8. Server performs phased elimination with the estimate $\hat{\theta}^{(\ell)}$ and confidence threshold $\bar{\gamma}_\ell$ to update active arm set:

$$\mathcal{A}_{\ell+1} = \{a \in \mathcal{A}_\ell : \max_{b \in \mathcal{A}_\ell} \mu(\langle \hat{\theta}^{(\ell)}, b \rangle) - \mu(\langle \hat{\theta}^{(\ell)}, a \rangle) \leq 2\bar{\gamma}_\ell\}. \tag{46}$$

11: $\ell = \ell + 1$ and **Goto** line 1.

E Algorithms and Analysis for the Generalized Linear Bandit Model

In this section, we first provide a detailed outline of the RC-GLM algorithm introduced in Section 5; see Algorithm 4. We then proceed to analyze RC-GLM, and provide a proof for Theorem 3. Finally, since RC-GLM uses the iteratively reweighted mean estimator from [22] as a sub-routine, we also provide a description of this estimator to keep the paper self-contained; this description, however, is deferred to the end of the section. We start by reminding the reader that the non-linear observation model of interest to us in this section is as follows:

$$y_{i,t} = \mu\left(\langle \theta_*, a_{i,t} \rangle\right) + \eta_{i,t}, \tag{47}$$

where $\mu : \mathbb{R} \to \mathbb{R}$ is the link function. We also recall the definition of $h_\ell(\theta)$:

$$h_\ell(\theta) \triangleq \sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} \mu(\langle \theta, a \rangle) a, \forall \theta \in \Theta.$$

From comparing RC-GLM (Algorithm 4) to RCLB (Algorithm 1), we note that while both algorithms share the same general structure, the key difference between the two stems from the manner in which the robust confidence thresholds are computed. In particular, to tackle the difficulty posed by the non-linearity of the observation map, we first compute a robust estimate $\hat{\theta}^{(\ell)}$ of θ_* at the server - a route that we avoided in RCLB - and then use such an estimate to devise a phased elimination rule.

E.1 Proof of Theorem 3

In this section, we prove Theorem 3. The crux of the analysis lies in deriving a robust confidence bound akin to that in Lemma 1. To work towards such a result, we need to first go through a few intermediate steps; these are as follows.

Step 1. Prove that conditioned on $\mathcal{F}_\ell$, for each $i \in [M] \setminus \mathcal{B}$, $\tilde{V}_\ell^{-1/2} Y_{i,\ell}$ is a d-dimensional Gaussian random variable with mean $\tilde{V}_\ell^{-1/2} h_\ell(\theta_*)$, and covariance matrix $\Sigma = I_d$.[6]

Step 2. Use the result from Step 1, along with the error-bounds of the iteratively reweighted Gaussian mean estimator from [22], to derive a high-probability error-bound on $\|X_\ell - \tilde{V}_\ell^{-1/2} h_\ell(\theta_*)\|$.

Step 3. Exploit regularity properties of the link function in tandem with the bounds from Step 2, to derive high-probability error bounds on $|\mu(\langle \hat{\theta}^{(\ell)}, a\rangle) - \mu(\langle \theta_*, a\rangle)|$, for each $a \in \mathcal{A}_\ell$.

We now proceed to formally establish each of the above steps, starting with step 1.

Lemma 7. *For each epoch ℓ, and each good agent $i \in [M] \setminus \mathcal{B}$, it holds that:*

$$\mathbb{E}\left[\tilde{V}_\ell^{-1/2} Y_{i,\ell} | \mathcal{F}_\ell\right] = \tilde{V}_\ell^{-1/2} h_\ell(\theta_*); \quad \textit{and} \tag{48}$$

$$\mathbb{E}\left[\left(\tilde{V}_\ell^{-1/2} Y_{i,\ell} - \tilde{V}_\ell^{-1/2} h_\ell(\theta_*)\right)\left(\tilde{V}_\ell^{-1/2} Y_{i,\ell} - \tilde{V}_\ell^{-1/2} h_\ell(\theta_*)\right)' | \mathcal{F}_\ell\right] = I_d. \tag{49}$$

Proof. Fix an epoch ℓ, and a good agent $i \in [M] \setminus \mathcal{B}$. We start by observing that:

$$\begin{aligned} Y_{i,\ell} &= \sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} r_{i,a}^{(\ell)} a \\ &= \sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} \left(\mu\left(\langle \theta_*, a\rangle\right) + \bar{\eta}_{i,a}^{(\ell)}\right) a \\ &= h_\ell(\theta_*) + \sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} \bar{\eta}_{i,a}^{(\ell)} a, \end{aligned} \tag{50}$$

where for the last step, we used the definition of $h_\ell(\theta_*)$. Just as in the proof of Theorem 1, we have used $\bar{\eta}_{i,a}^{(\ell)}$ to denote the average of the noise terms associated with the rewards observed by agent i during phase ℓ for arm a. We thus have:

$$\tilde{V}_\ell^{-1/2} Y_{i,\ell} = \tilde{V}_\ell^{-1/2} h_\ell(\theta_*) + \tilde{V}_\ell^{-1/2} \left(\sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} \bar{\eta}_{i,a}^{(\ell)} a\right).$$

Now conditioned on $\mathcal{F}_\ell$, the only randomness in the above equation stems from the noise terms $\{\bar{\eta}_{i,a}^{(\ell)}\}, a \in \text{Supp}(\pi_\ell)$, that are each zero-mean. The claim in Eq. (48) thus follows.

Based on Eq. (50), we have:

$$\begin{aligned} \mathbb{E}\left[\left(Y_{i,\ell} - h_\ell(\theta_*)\right)\left(Y_{i,\ell} - h_\ell(\theta_*)\right)' | \mathcal{F}_\ell\right] &= \mathbb{E}\left[\left(\sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} \bar{\eta}_{i,a}^{(\ell)} a\right)\left(\sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} \bar{\eta}_{i,a}^{(\ell)} a\right)' \Big| \mathcal{F}_\ell\right] \\ &\overset{(a)}{=} \sum_{a \in \text{Supp}(\pi_\ell)} \left(m_a^{(\ell)}\right)^2 \mathbb{E}\left[\left(\bar{\eta}_{i,a}^{(\ell)}\right)^2\right] aa' \\ &\overset{(b)}{=} \sum_{a \in \text{Supp}(\pi_\ell)} m_a^{(\ell)} aa' \\ &\overset{(c)}{=} \tilde{V}_\ell. \end{aligned} \tag{51}$$

[6] Recall that $\mathcal{F}_\ell$ is the σ-algebra generated by all the actions and rewards up to the beginning of epoch ℓ.

In the above steps, (a) follows by observing that the noise terms are independent across different arms; hence, the expectation of each of the cross-terms vanish. For (b), we used the fact that $\bar{\eta}_{i,a}^{(\ell)}$ is the average of $m_a^{(\ell)}$ independent Gaussian noise terms, each with zero-mean and unit variance; hence, $\mathbb{E}\left[\left(\bar{\eta}_{i,a}^{(\ell)}\right)^2\right] = 1/(m_a^{(\ell)})$. For (c), we simply used the definition of $\tilde{V}_\ell$. In light of Eq. (51), it is easy to see why Eq. (49) holds. □

We now state - adapted to our notation - one of the main convergence guarantees from [22] for the iteratively reweighted mean estimator.

Lemma 8. *Suppose we are given M d-dimensional samples $x_1, \ldots, x_M$, such that $(1-\alpha)M$ of these samples are drawn i.i.d. from $\mathcal{N}(v, \Sigma)$, where $v \in \mathbb{R}^d$ is an unknown mean vector, and $\Sigma \in \mathbb{R}^{d\times d}$ is a known covariance matrix. The remaining αM samples can be arbitrary. Let the corruption fraction α satisfy $\alpha < (5-\sqrt{5})/10$, and let $\delta \in (16\exp(-M), 1)$ be a given tolerance level. Then, with probability at least $1-\delta$, we have*

$$\|\hat{v} - v\|_2 \leq C_1 \|\Sigma\|_2^{1/2} \left(\sqrt{\frac{d + \log(16/\delta)}{M}} + \alpha\sqrt{\log\frac{1}{\alpha}}\right), \tag{52}$$

where C_1 is a suitably large universal constant, and $\hat{v}$ is the output of the Iteratively Reweighted Mean Estimator, namely Algorithm 1 in [22], when it takes as input the M samples, the covariance matrix Σ, and the corruption fraction α.

Let us now see how the above bound can assist in our cause. Fix any epoch ℓ, and recall that $\delta_\ell = \bar{\delta}/(K\ell^2)$, where $\bar{\delta} = \delta/(10K)$, and δ is the given confidence parameter. Suppose we want to derive an error-bound based on Lemma 8 that holds with probability at least $1-\delta_\ell$. For this to happen, we need $\delta_\ell > 16\exp(-M)$. Since $\ell \leq T$, one can verify that the aforementioned condition is satisfied as long as M is large enough in the following sense:

$$M > \log\left(\frac{160K^2T^2}{\delta}\right). \tag{53}$$

From now on, we assume that the above condition holds. Next, recall that

$$X_\ell = \texttt{ITW}(\{\tilde{V}_\ell^{-1/2} Y_{i,\ell}, i \in [M]\}).$$

Based on Lemma 7, Lemma 8, and the same line of reasoning as used to arrive at Eq. (24), we have that with probability at least $1-\delta_\ell$:

$$\|X_\ell - \tilde{V}_\ell^{-1/2} h_\ell(\theta_*)\| \leq C_1 \left(\sqrt{\frac{d + \log(16/\delta_\ell)}{M}} + \alpha\sqrt{\log\frac{1}{\alpha}}\right). \tag{54}$$

We will call upon the above bound later in our analysis. For now, this ends Step 2. As for Step 3, we start with the following result.

Lemma 9. *Consider any $\theta_1, \theta_2 \in \Theta$, and any epoch ℓ. There exists a symmetric positive definite matrix $G_\ell(\theta_1;\theta_2)$ satisfying $k_1\tilde{V}_\ell \preccurlyeq G_\ell(\theta_1;\theta_2) \preccurlyeq k_2\tilde{V}_\ell$, such that:*

$$h_\ell(\theta_1) - h_\ell(\theta_2) = G_\ell(\theta_1;\theta_2)(\theta_1 - \theta_2). \tag{55}$$

Proof. For any $\theta \in \Theta$, let us denote by $\nabla h_\ell(\theta)$ the Jacobian matrix of $h_\ell(\cdot)$ at θ. Such a matrix exists based on Assumption 1. Now based on the mean value theorem, $\exists \alpha \in (0,1)$ such that

$$h_\ell(\theta_1) - h_\ell(\theta_2) = \left(\nabla h_\ell\left(\alpha\theta_1 + (1-\alpha)\theta_2\right)\right)(\theta_1 - \theta_2).$$

Let $\bar{\theta} = \alpha\theta_1 + (1-\alpha)\theta_2$, and $G_\ell(\theta_1;\theta_2) = \nabla h_\ell(\bar{\theta})$. To complete the proof, we need to show that the matrix $G_\ell(\theta_1;\theta_2)$ defined above is symmetric, positive definite, and bounded above and below (in the Loewner sense) by scalar multiples of $\tilde{V}_\ell$. To that end, observe that:

$$\begin{aligned}
\nabla h_\ell(\bar{\theta}) &\overset{(a)}{=} \sum_{a\in\text{Supp}(\pi_\ell)} m_a^{(\ell)} \dot{\mu}(\langle\bar{\theta}, a\rangle) aa' \\
&\overset{(b)}{\succcurlyeq} k_1 \left(\sum_{a\in\text{Supp}(\pi_\ell)} m_a^{(\ell)} aa'\right) \\
&= k_1 \tilde{V}_\ell.
\end{aligned} \tag{56}$$

In the above steps, we used the definition of $h_\ell(\cdot)$ for (a); and for (b), we used Assumption 1. From the above steps, it is clear that $G_\ell(\theta_1;\theta_2)$ is symmetric. That it is also positive definite follows from the fact that $\tilde{V}_\ell \succ 0$. Finally, using the fact that $\mu(\cdot)$ is k_2-Lipschitz, and a similar line of reasoning, one can show that $G_\ell(\theta_1;\theta_2) \preccurlyeq k_2\tilde{V}_\ell$. This concludes the proof. □

We now have all the pieces required to establish an analogue of Lemma 1.

Lemma 10. ***(Robust Confidence Intervals for*** `RC-GLM`***)*** *Suppose M satisfies the condition in Eq. (53), and $\alpha < (5-\sqrt{5})/10$. Fix any epoch ℓ. For each arm $a \in \mathcal{A}_\ell$, with probability at least $1-\delta_\ell$, it holds that:*

$$|\mu(\langle\hat{\theta}^{(\ell)}, a\rangle) - \mu(\langle\theta_*, a\rangle)| \le \bar{\gamma}_\ell, \quad \textit{where} \quad \bar{\gamma}_\ell = 4C_1\frac{k_2}{k_1}\left(\sqrt{d} + \alpha\sqrt{M\log(1/\alpha)}\right)\epsilon_\ell, \tag{57}$$

and C_1 is as in Lemma 8.

Proof. Let us start by conditioning on the event of measure at least $1-\delta_\ell$ on which Eq. (54) holds. Invoking Lemma 9, we know that there exists a symmetrix positive definite matrix G_ℓ such that $k_1\tilde{V}_\ell \preccurlyeq G_\ell \preccurlyeq k_2\tilde{V}_\ell$, and:[7]

$$\begin{aligned} G_\ell\left(\hat{\theta}^{(\ell)} - \theta_*\right) &= h_\ell(\hat{\theta}^{(\ell)}) - h_\ell(\theta_*) \\ &= \tilde{V}_\ell^{1/2}X_\ell - h_\ell(\theta_*) \\ &= \tilde{V}_\ell^{1/2}\left(X_\ell - \tilde{V}_\ell^{-1/2}h_\ell(\theta_*)\right). \end{aligned} \tag{58}$$

For the second step above, we used the fact that based on line 9 of `RC-GLM`, $h_\ell(\hat{\theta}^{(\ell)}) = \tilde{V}_\ell^{1/2}X_\ell$. Now fix any arm $a \in \mathcal{A}_\ell$, and observe that:

$$\langle\hat{\theta}^{(\ell)} - \theta_*, a\rangle = \left\langle G_\ell^{-1}\tilde{V}_\ell^{1/2}\left(X_\ell - \tilde{V}_\ell^{-1/2}h_\ell(\theta_*)\right), a\right\rangle.$$

This, in turn, implies:

$$|\langle\hat{\theta}^{(\ell)} - \theta_*, a\rangle| \le \underbrace{\left\|G_\ell^{-1}\tilde{V}_\ell^{1/2}\left(X_\ell - \tilde{V}_\ell^{-1/2}h_\ell(\theta_*)\right)\right\|_{\tilde{V}_\ell}}_{T_1}\|a\|_{\tilde{V}_\ell^{-1}}. \tag{59}$$

We bound T_1 as follows.

$$\begin{aligned} T_1 &= \sqrt{\left(X_\ell - \tilde{V}_\ell^{-1/2}h_\ell(\theta_*)\right)'\tilde{V}_\ell^{1/2}G_\ell^{-1}\tilde{V}_\ell G_\ell^{-1}\tilde{V}_\ell^{1/2}\left(X_\ell - \tilde{V}_\ell^{-1/2}h_\ell(\theta_*)\right)} \\ &\overset{(a)}{\le} \frac{1}{\sqrt{k_1}}\sqrt{\left(X_\ell - \tilde{V}_\ell^{-1/2}h_\ell(\theta_*)\right)'\tilde{V}_\ell^{1/2}G_\ell^{-1}\tilde{V}_\ell^{1/2}\left(X_\ell - \tilde{V}_\ell^{-1/2}h_\ell(\theta_*)\right)} \\ &\overset{(b)}{\le} \frac{1}{k_1}\sqrt{\left(X_\ell - \tilde{V}_\ell^{-1/2}h_\ell(\theta_*)\right)'\left(X_\ell - \tilde{V}_\ell^{-1/2}h_\ell(\theta_*)\right)} \\ &\overset{(c)}{\le} \frac{C_1}{k_1}\left(\sqrt{\frac{d+\log(16/\delta_\ell)}{M}} + \alpha\sqrt{\log\frac{1}{\alpha}}\right). \end{aligned} \tag{60}$$

For both (a) and (b) above, we used $k_1\tilde{V}_\ell \preccurlyeq G_\ell$; for (c), we invoked the bound in Eq. (54). Now recall from Eq. (22) that

$$\|a\|_{\tilde{V}_\ell^{-1}} \le \epsilon_\ell\sqrt{\frac{2M}{\log\left(\frac{1}{\delta_\ell}\right)}}.$$

Combining the above bound with the ones in Eq. (59) and Eq. (60), and using $\log(1/\delta_\ell) \ge 1$, we obtain

$$|\langle\hat{\theta}^{(\ell)} - \theta_*, a\rangle| \le \frac{\sqrt{2}C_1}{k_1}\left(\sqrt{d + \frac{\log(16/\delta_\ell)}{\log(1/\delta_\ell)}} + \alpha\sqrt{M\log\frac{1}{\alpha}}\right)\epsilon_\ell.$$

[7] Here, we have dropped the dependence of G_ℓ on $\hat{\theta}^{(\ell)}$ and θ_* to lighten the notation.

Elementary calculations coupled with the fact that $\log(1/\delta_\ell) \geq 1$ yields:

$$\frac{\log(16/\delta_\ell)}{\log(1/\delta_\ell)} \leq 4.$$

Putting all the pieces together, and simplifying, we arrive at the following bound:

$$|\langle \hat{\theta}^{(\ell)} - \theta_*, a\rangle| \leq \frac{4C_1}{k_1}\left(\sqrt{d} + \alpha\sqrt{M\log\frac{1}{\alpha}}\right)\epsilon_\ell. \tag{61}$$

Using the fact that $\mu(\cdot)$ is k_2-Lipschitz then yields:

$$|\mu(\langle \hat{\theta}^{(\ell)}, a\rangle) - \mu(\langle \theta_*, a\rangle)| \leq k_2|\langle \hat{\theta}^{(\ell)} - \theta_*, a\rangle| \leq 4C_1\frac{k_2}{k_1}\left(\sqrt{d} + \alpha\sqrt{M\log\frac{1}{\alpha}}\right)\epsilon_\ell,$$

which is the desired claim. This completes the proof. □

Having derived the robust confidence bounds for `RC-GLM`, we can complete the proof of Theorem 3.

Proof. (**Proof of Theorem 3**). Using essentially the same arguments as those used to prove Lemmas 4 and 5, we can prove that there exists a clean event, say $\mathcal{E}$, of measure at least $1-\delta$, such that on $\mathcal{E}$, the following hold: (i) $a_* \in \mathcal{A}_\ell, \forall \ell \in [L]$, where L is the total number of epochs; and (ii) for any epoch $\ell \in [L], a \in \mathcal{A}_\ell \implies \tilde{\Delta}_a \leq 8\bar{\gamma}_\ell$. Here, $\tilde{\Delta}_a = \mu(\langle \theta_*, a_*\rangle) - \mu(\langle \theta_*, a\rangle)$. Let us condition on this clean event $\mathcal{E}$. The remainder of the proof follows the same line of reasoning as that of Theorem 1. For any good agent $i \in [M] \setminus \mathcal{B}$, we can bound the regret as follows.

$$\begin{aligned}
\sum_{t=1}^{T}\left(\mu(\langle \theta_*, a_*\rangle) - \mu(\langle \theta_*, a_{i,t}\rangle)\right) &= \sum_{\ell=1}^{L}\sum_{a\in\mathsf{Supp}(\pi_\ell)} m_a^{(\ell)}\left(\mu(\langle \theta_*, a_*\rangle) - \mu(\langle \theta_*, a\rangle)\right) \\
&= \sum_{\ell=1}^{L}\sum_{a\in\mathsf{Supp}(\pi_\ell)} \left\lceil \frac{T_a^{(\ell)}}{M}\right\rceil \tilde{\Delta}_a \\
&\leq \underbrace{\sum_{\ell=1}^{L}\sum_{a\in\mathsf{Supp}(\pi_\ell)} \frac{T_a^{(\ell)}}{M}\tilde{\Delta}_a}_{T_1} + \underbrace{\sum_{\ell=1}^{L}\sum_{a\in\mathsf{Supp}(\pi_\ell)} \tilde{\Delta}_a}_{T_2}.
\end{aligned} \tag{62}$$

As in the proof of Theorem 1, we bound T_1 and T_2 separately. For bounding T_1, we have:

$$\begin{aligned}
T_1 &= \sum_{\ell=1}^{L}\sum_{a\in\mathsf{Supp}(\pi_\ell)} \frac{T_a^{(\ell)}}{M}\tilde{\Delta}_a \\
&= \frac{d}{M}\sum_{\ell=1}^{L}\frac{1}{\epsilon_\ell^2}\log\left(\frac{1}{\delta_\ell}\right)\sum_{a\in\mathsf{Supp}(\pi_\ell)}\pi_\ell(a)\tilde{\Delta}_a \\
&\leq \frac{8d}{M}\sum_{\ell=1}^{L}\frac{1}{\epsilon_\ell^2}\log\left(\frac{1}{\delta_\ell}\right)\bar{\gamma}_\ell \\
&= \frac{32C_1 d\left(\sqrt{d} + \alpha\sqrt{M\log(1/\alpha)}\right)}{M}\left(\frac{k_2}{k_1}\right)\sum_{\ell=1}^{L}\frac{1}{\epsilon_\ell}\log\left(\frac{10K^2\ell^2}{\delta}\right) \\
&\leq \frac{32C_1 d\left(\sqrt{d} + \alpha\sqrt{M\log(1/\alpha)}\right)}{M}\left(\frac{k_2}{k_1}\right)\log\left(\frac{10K^2L^2}{\delta}\right)\sum_{\ell=1}^{L}2^\ell \\
&= O\left(\left(\frac{k_2}{k_1}\right)\frac{d\left(\sqrt{d} + \alpha\sqrt{M\log(1/\alpha)}\right)}{M}\log\left(\frac{10K^2L^2}{\delta}\right)2^L\right).
\end{aligned} \tag{63}$$

Recall the following fact that we proved earlier for Theorem 1:

$$2^L \leq \sqrt{\frac{MT}{d\log\left(10K^2L^2/\delta\right)}}.$$

Plugging this bound in Eq. (63), we obtain:

$$T_1 = O\left(\left(\frac{k_2}{k_1}\right)\left(\alpha\sqrt{\log(1/\alpha)} + \sqrt{\frac{d}{M}}\right)\sqrt{\log\left(\frac{KT}{\delta}\right)dT}\right).$$

One can upper-bound T_2 using the same bound as above using exactly the same steps as in the proof of Theorem 1. Combining the bounds on T_1 and T_2 leads to the desired claim. □

• **Comments on solving for $\hat{\theta}^{(\ell)}$ in** `RC-GLM`. Recall that line 9 of Algorithm `RC-GLM` requires the server to solve for $\hat{\theta}^{(\ell)}$ based on the following equation:

$$h_\ell(\hat{\theta}^{(\ell)}) = \tilde{V}_\ell^{1/2} X_\ell. \tag{64}$$

Based on Lemma 9, we have that for any $\theta_1, \theta_2 \in \Theta$ such that $\theta_1 \neq \theta_2$:

$$(\theta_1 - \theta_2)' \left(h_\ell(\theta_1) - h_\ell(\theta_2)\right) = (\theta_1 - \theta_2)' \, G_\ell(\theta_1; \theta_2) \, (\theta_1 - \theta_2) > 0,$$

since $G_\ell(\theta_1; \theta_2)$ is positive-definite. Thus, the map $h_\ell : \mathbb{R}^d \to \mathbb{R}^d$ is injective, and h_ℓ^{-1} is well-defined. This, in turn, implies that Eq. (64) has a unique solution.

E.2 The Iteratively Reweighted Mean Estimation Algorithm

In this section, we briefly explain the main idea behind the Iteratively Reweighted Mean Estimation Algorithm in [22]. Suppose we are given an α-corrupted set $\mathcal{X}$ of samples, namely, M d-dimensional samples $x_1, \ldots, x_M$, such that $(1-\alpha)M$ of these samples are drawn i.i.d. from $\mathcal{N}(v, \Sigma)$, and the remaining αM samples are arbitrarily corrupted by an adversary. The goal is to recover the unknown mean vector $v \in \mathbb{R}^d$, given knowledge of the samples $\mathcal{X}$, the corruption fraction α, and the covariance matrix Σ. To see how this is done, let us define a couple of quantities for any pair of vectors $w \in [0,1]^d$ and $\mu \in \mathbb{R}^d$:

$$\bar{x}_w = \sum_{i=1}^{M} w_i x_i; \quad G(w, \mu) = \lambda_{\max}\left(\sum_{i=1}^{M} w_i \left(x_i - \mu\right)\left(x_i - \mu\right)' - \Sigma\right). \tag{65}$$

The basic idea is to find a weight vector $\hat{w}$ within the d-dimensional probability simplex such that the weighted average $\bar{x}_{\hat{w}}$ is close to the true mean v. Intuitively, a "small" value of $G(\hat{w}, \bar{x}_{\hat{w}})$ is an indicator of a good candidate for such a weight vector. This is essentially the strategy pursued in Algorithm 5 where one iteratively updates the weight vectors, and the associated weighted averages, so as to minimize the function $G(\cdot, \cdot)$ defined in Eq. (65). At the termination of this algorithm, say after N iterations, the goal is to output a weight vector $\hat{w}_N$ that mimics the ideal weight vector w^* defined by: $w_j^* = \mathbf{1}(j \in \mathcal{I})/|\mathcal{I}|$, where $\mathcal{I}$ is the set of good samples (inliers). The steps of the Iteratively Reweighted Mean Estimator are outlined in Algorithm 5.

Algorithm 5 Iteratively Reweighted Mean Estimator (`ITW`)

Input: α-corrupted set of M samples $\mathcal{X}$, corruption fraction α, and covariance matrix Σ.
Output: Robust estimate of the mean $\hat{v}$.
Initialize: Compute $\hat{v}_0$ as a minimizer of $\operatorname{argmin}_{\mu} \sum_{i=1}^{M} \|x_i - \mu\|$.
Let $N = 0 \vee \left\lceil \frac{\log(4r_\Sigma) - 2\log(\alpha(1-2\alpha))}{2\log((1-2\alpha)) - \log(\alpha) - \log(1-\alpha)} \right\rceil$. Here, $r_\Sigma = \text{Trace}(\Sigma)/\|\Sigma\|_2$.

for $k = 1$ to N **do**
Compute current weights:

$$w \in \operatorname*{argmin}_{(M-M\epsilon)\|w\|_\infty \leq 1} \lambda_{\max}\left(\sum_{i=1}^{M} w_i (x_i - \hat{v}_{k-1})(x_i - \hat{v}_{k-1})' - \Sigma\right) \vee 0.$$

Update the estimator:

$$\hat{v}_k = \sum_{i=1}^{M} w_i x_i.$$

end for
Return $\hat{v} = \hat{v}_K$.

F Analysis for the Contextual Bandit Setting: Proof of Theorem 4

The proof of Theorem 4 proceeds in multiple steps. We start with an analysis of the Robust BaseLinUCB subroutine, namely Algorithm 2.

Lemma 11. ***(Bounds for Robust BaseLinUCB)*** *Suppose the input index set Ψ_t is constructed so that for fixed $x_{\tau,a_\tau}, \tau \in \Psi_t$, the rewards $\{r_{i,\tau}\}_{\tau\in\Psi_t}$ are independent random variables for each good agent $i \in [M] \setminus \mathcal{B}$. Then, for each $a \in [K]$, with probability at least $1-\bar{\delta}$, it holds that:*

$$|\hat{r}_{t,a} - \langle \theta_*, x_{t,a}\rangle| \leq \left(\alpha + 2C\sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}}\right) \|x_{t,a}\|_{A_t^{-1}}, \tag{66}$$

where C is as in Lemma 2.

Proof. Fix any agent $i \in [M] \setminus \mathcal{B}$. Now from the expression for $\hat{\theta}_{i,t}$ in Algorithm 2, observe that:

$$\begin{aligned}
\hat{\theta}_{i,t} &= A_t^{-1} b_{i,t} \\
&= A_t^{-1}\left(\sum_{\tau\in\Psi_t} r_{i,\tau} x_{\tau,a_\tau}\right) \\
&= A_t^{-1}\left(\sum_{\tau\in\Psi_t} \left(\langle \theta_*, x_{\tau,a_\tau}\rangle + \eta_{i,\tau}\right) x_{\tau,a_\tau}\right) \\
&= A_t^{-1}\left(\sum_{\tau\in\Psi_t} x_{\tau,a_\tau} x'_{\tau,a_\tau}\right)\theta_* + A_t^{-1}\left(\sum_{\tau\in\Psi_t} \eta_{i,\tau} x_{\tau,a_\tau}\right) \\
&= \left(I_d - \frac{A_t^{-1}}{M}\right)\theta_* + A_t^{-1}\left(\sum_{\tau\in\Psi_t} \eta_{i,\tau} x_{\tau,a_\tau}\right),
\end{aligned} \tag{67}$$

where we used $\eta_{i,\tau}$ as a shorthand for $\eta_{i,\tau}(a_\tau)$. Now consider any $a \in [K]$. Using the above expression, we will now decompose the error in estimation of $\langle \theta_*, x_{t,a}\rangle$ into a bias term and a variance term:

$$\langle \hat{\theta}_{i,t} - \theta_*, x_{t,a}\rangle = \underbrace{-\frac{1}{M}\langle A_t^{-1}\theta_*, x_{t,a}\rangle}_{\text{bias term}} + \underbrace{\sum_{\tau\in\Psi_t} \langle A_t^{-1} x_{\tau,a_\tau}, x_{t,a}\rangle \eta_{i,\tau}}_{\text{variance term}}. \tag{68}$$

For a fixed set of feature vectors, under our assumption that $\{r_{i,\tau}\}_{\tau\in\Psi_t}$ are independent random variables, the variance term is a sum of independent zero-mean Gaussian noise variables. Thus,

$$\mathbb{E}\left[\sum_{\tau\in\Psi_t} \langle A_t^{-1} x_{\tau,a_\tau}, x_{t,a}\rangle \eta_{i,\tau}\right] = \sum_{\tau\in\Psi_t} \langle A_t^{-1} x_{\tau,a_\tau}, x_{t,a}\rangle \mathbb{E}\left[\eta_{i,\tau}\right] = 0.$$

Furthermore, we have:

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{\tau\in\Psi_t} \langle A_t^{-1} x_{\tau,a_\tau}, x_{t,a}\rangle \eta_{i,\tau}\right)^2\right] &= x'_{t,a} A_t^{-1}\left(\sum_{\tau\in\Psi_t} \mathbb{E}\left[\left(\eta_{i,\tau}^2\right)\right] x_{\tau,a_\tau} x'_{\tau,a_\tau}\right) A_t^{-1} x_{t,a} \\
&= x'_{t,a} A_t^{-1}\left(\sum_{\tau\in\Psi_t} x_{\tau,a_\tau} x'_{\tau,a_\tau}\right) A_t^{-1} x_{t,a} \\
&\leq x'_{t,a} A_t^{-1} \underbrace{\left(\frac{I_d}{M} + \sum_{\tau\in\Psi_t} x_{\tau,a_\tau} x'_{\tau,a_\tau}\right)}_{A_t} A_t^{-1} x_{t,a} \\
&= \|x_{t,a}\|^2_{A_t^{-1}}.
\end{aligned} \tag{69}$$

From the above arguments, we conclude that for each $i \in [M] \setminus \mathcal{B}$,

$$\langle \hat{\theta}_{i,t}, x_{t,a} \rangle \sim \mathcal{N}\left(\langle \theta_* - \frac{A_t^{-1}\theta_*}{M}, x_{t,a} \rangle, \sigma^2 \right), \quad \text{where} \quad \sigma^2 \leq \|x_{t,a}\|^2_{A_t^{-1}}.$$

Since the noise samples are independent across agents, we also know that $\{\langle \hat{\theta}_{i,t}, x_{t,a} \rangle\}_{i \in [M] \setminus \mathcal{B}}$ are independent. Recalling that $\hat{r}_{t,a} = \texttt{Median}\left(\{\langle \hat{\theta}_{i,t}, x_{t,a} \rangle, i \in [M]\} \right)$, and invoking Lemma 2, we then have that with probability at least $1 - \bar{\delta}$,

$$|\hat{r}_{t,a} - \langle \theta_* - \frac{A_t^{-1}\theta_*}{M}, x_{t,a} \rangle| \leq \left(\alpha + C\sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \right) \|x_{t,a}\|_{A_t^{-1}}. \tag{70}$$

This immediately implies:

$$|\hat{r}_{t,a} - \langle \theta_*, x_{t,a} \rangle| \leq \frac{1}{M} |\langle A_t^{-1}\theta_*, x_{t,a} \rangle| + \left(\alpha + C\sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \right) \|x_{t,a}\|_{A_t^{-1}}. \tag{71}$$

It remains to bound the first term in the above display. To that end, we proceed as follows:

$$\begin{aligned}
\frac{1}{M} |\langle A_t^{-1}\theta_*, x_{t,a} \rangle| &= \frac{1}{M} |\langle A_t^{-1} x_{t,a}, \theta_* \rangle| \\
&\leq \frac{1}{M} \|A_t^{-1} x_{t,a}\| \|\theta_*\| \\
&\overset{(a)}{\leq} \frac{1}{\sqrt{M}} \sqrt{x'_{t,a} A_t^{-1} \frac{I_d}{M} A_t^{-1} x_{t,a}} \\
&\leq \frac{1}{\sqrt{M}} \sqrt{x'_{t,a} A_t^{-1} \left(\frac{I_d}{M} + \sum_{\tau \in \Psi_t} x_{\tau,a_\tau} x'_{\tau,a_\tau} \right) A_t^{-1} x_{t,a}} \\
&= \frac{1}{\sqrt{M}} \|x_{t,a}\|_{A_t^{-1}} \\
&\overset{(b)}{\leq} C\sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \|x_{t,a}\|_{A_t^{-1}},
\end{aligned} \tag{72}$$

where for (a), we used the fact that $\|\theta_*\| \leq 1$ by assumption; and for (b), we used $C \geq 1, \log\left(\frac{1}{\bar{\delta}}\right) \geq 1$. Combining the above bound with the one in Eq.(71) leads to the desired claim. □

For Lemma 11 to hold, the crucial requirement is for the rewards corresponding to indices in Ψ_t to be independent. Our next result shows that this is indeed the case; the proof of this lemma essentially follows the same arguments as that of [24, Lemma 14], we reproduce these arguments here only for completeness.

Lemma 12. ***(Independence of Samples)*** *Fix any agent $i \in [M] \setminus \mathcal{B}$. For each $s \in [S]$ and each $t \in [T]$, given any fixed sequence of feature vectors $\{x_{\tau,a_\tau}\}_{\tau \in \Psi_t^{(s)}}$, the rewards $\{r_{i,\tau}\}_{\tau \in \Psi_t^{(s)}}$ are independent random variables.*

Proof. Let us start by observing that a time-step t can be added to $\Psi_t^{(s)}$ only in line 6 of Algorithm 3. Thus, the event $\{t \in \Psi_t^{(s)}\}$ only depends on all prior phases $\cup_{\ell < s} \Psi_t^{(\ell)}$, and on the confidence width $w_{t,a}^{(s)}$. From the definition of $w_{t,a}^{(s)}$ in line 7 of Algorithm 2, we note that $w_{t,a}^{(s)}$ depends only on the feature vectors $x_{\tau,a_\tau}, \tau \in \Psi_t^{(s)}$, and $x_{t,a}$. Combining the above observations, it is easy to see that $\{t \in \Psi_t^{(s)}\}$ only depends on the feature vectors. Noting that the feature vector sequence is fixed, and cannot be controlled by the adversarial agents, leads to the claim of the lemma. □

The next result tells us that with high-probability, at each time-step $t \in [T]$, (i) the best arm a_t^* is retained in all stages of the screening process; and (ii) an active arm in phase s can contribute to at most $8/(2^s\sqrt{M})$ instantaneous regret.

Lemma 13. *With probability at least $1 - \bar{\delta} KST$, for any $t \in [T]$ and any $s \in [S]$, the following hold:*

(i) $|\hat{r}_{t,a}^{(s)} - \langle \theta_*, x_{t,a} \rangle| \leq w_{t,a}^{(s)}, \forall a \in \mathcal{A}_s$.

(ii) $a_t^* \in \mathcal{A}_s$.

(iii) $\langle \theta_*, x_{t,a_t^*} \rangle - \langle \theta_*, x_{t,a} \rangle \leq 8/(2^s \sqrt{M}), \forall a \in \mathcal{A}_s$.

Proof. Part (i) of the result follows directly from Lemma 11, and an union bound over all time-steps, phases and arms.

For part (ii), let us condition on the clean event, say $\mathcal{E}$, on which part (i) holds. From the rules of Algorithm 3, it holds trivially that $a_t^* \in \mathcal{A}_s$ for $s = 1$. Now suppose there exists some phase $s > 1$ such that $a_t^* \in \mathcal{A}_{s-1}$, but $a_t^* \notin \mathcal{A}_s$. From line 8 in Algorithm 3, we must have $w_{t,a}^{(s-1)} \leq 2^{1-s}/\sqrt{M}, \forall a \in \mathcal{A}_{s-1}$. From the phased elimination strategy in line 8 of Algorithm 3, $a_t^* \notin \mathcal{A}_s$ implies the existence of some arm $a \in \mathcal{A}_s$ such that:

$$\begin{aligned}
&\left(\hat{r}_{t,a}^{(s-1)} + w_{t,a}^{(s-1)}\right) - \left(\hat{r}_{t,a_t^*}^{(s-1)} + w_{t,a_t^*}^{(s-1)}\right) > \frac{2^{2-s}}{\sqrt{M}} \\
\overset{(a)}{\implies} &\langle \theta_*, x_{t,a} \rangle + 2w_{t,a}^{(s-1)} - \left(\hat{r}_{t,a_t^*}^{(s-1)} + w_{t,a_t^*}^{(s-1)}\right) > \frac{2^{2-s}}{\sqrt{M}} \\
\overset{(b)}{\implies} &\langle \theta_*, x_{t,a} - x_{t,a_t^*} \rangle + 2w_{t,a}^{(s-1)} > \frac{2^{2-s}}{\sqrt{M}} \\
\overset{(c)}{\implies} &2w_{t,a}^{(s-1)} > \frac{2^{2-s}}{\sqrt{M}},
\end{aligned} \tag{73}$$

which leads to a contradiction as $w_{t,a}^{(s-1)} \leq 2^{1-s}/\sqrt{M}$. In the above steps, both (a) and (b) follow from the defining property of the clean event $\mathcal{E}$; for (c), we used $\langle \theta_*, x_{t,a} - x_{t,a_t^*} \rangle \leq 0$ from the optimality of a_t^*. This completes the proof of part (ii).

For part (iii), let us once again condition on the clean event $\mathcal{E}$ on which part (i) holds. Now $a \in \mathcal{A}_s \implies a \in \mathcal{A}_{s-1}$. We also know from part (ii) that $a_t^* \in \mathcal{A}_{s-1}$. The retention of arm a in $\mathcal{A}_s$ implies (based on line 8 of Algorithm 3),

$$\begin{aligned}
&\left(\hat{r}_{t,a_t^*}^{(s-1)} + w_{t,a_t^*}^{(s-1)}\right) - \left(\hat{r}_{t,a}^{(s-1)} + w_{t,a}^{(s-1)}\right) \leq \frac{2^{2-s}}{\sqrt{M}} \\
\overset{(a)}{\implies} &\langle \theta_*, x_{t,a_t^*} \rangle - \left(\hat{r}_{t,a}^{(s-1)} + w_{t,a}^{(s-1)}\right) \leq \frac{2^{2-s}}{\sqrt{M}} \\
\overset{(b)}{\implies} &\langle \theta_*, x_{t,a_t^*} - x_{t,a} \rangle \leq \frac{2^{2-s}}{\sqrt{M}} + 2w_{t,a}^{(s-1)} \\
\overset{(c)}{\implies} &\langle \theta_*, x_{t,a_t^*} - x_{t,a} \rangle \leq \frac{8}{2^s \sqrt{M}},
\end{aligned} \tag{74}$$

which is the desired claim. Here, for (a) and (b) we used the defining property of $\mathcal{E}$. As for (c), we used the fact that $w_{t,a}^{(s-1)} \leq 2^{1-s}/\sqrt{M}$. □

The final piece needed in the proof of Theorem 4 is a bound on $|\Psi_T^{(s)}|$ for each $s \in [S]$.

To that end, we will make use of the elliptical potential lemma from [16].

Lemma 14. ***(Elliptical Potential Lemma)*** *Let $\{X_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$, V a $d \times d$ positive definite matrix, and define $\bar{V}_t = V + \sum_{\tau=1}^{t} X_\tau X_\tau'$. If $\|X_t\| \leq L$ for all t, then we have that*

$$\sum_{t=1}^{T} \min\{1, \|X_t\|^2_{\bar{V}_{t-1}^{-1}}\} \leq 2 \log \frac{\det(\bar{V}_T)}{\det(V)} \leq 2(d \log((\mathit{trace}(V) + TL^2)/d) - \log(\det V)).$$

We have the following result.

Lemma 15. ***(Bound on*** $|\Psi_T^{(s)}|$ ***)*** *Fix any* $s \in [S]$*. The following then holds:*

$$|\Psi_T^{(s)}| \leq 2^s \sqrt{M} \left(\alpha + 2C \sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \right) \sqrt{2d|\Psi_T^{(s)}| \log(2M|\Psi_T^{(s)}|)}. \tag{75}$$

Proof. Fix any phase $s \in [S]$. Now consider any time-step $t \in \Psi_T^{(s)}$. Since $t \in \Psi_T^{(s)}$, based on line 6 of Algorithm 3, it must be that:

$$\frac{1}{2^s \sqrt{M}} < w_{t,a_t}^{(s)} = \left(\alpha + 2C \sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \right) \|x_{t,a_t}\|_{A_{s,t}^{-1}}, \quad \text{where} \tag{76}$$

$$A_{s,t} = \left(\frac{I_d}{M} + \sum_{\tau \in \Psi_t^{(s)}} x_{\tau,a_\tau} x'_{\tau,a_\tau} \right).$$

Also, since $s \geq 1$, we have the following trivial inequality:

$$\frac{1}{2^s \sqrt{M}} < \left(\alpha + 2C \sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \right).$$

Combining the above inequality with the one in (76), we note that for each $t \in \Psi_T^{(s)}$, it holds that:

$$\frac{1}{2^s \sqrt{M}} < \left(\alpha + 2C \sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \right) \min\{1, \|x_{t,a_t}\|_{A_{s,t}^{-1}}\}.$$

Summing the above display over all indices in $\Psi_T^{(s)}$ yields:

$$\begin{aligned}
\frac{|\Psi_T^{(s)}|}{2^s \sqrt{M}} &< \left(\alpha + 2C \sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \right) \sum_{t \in \Psi_T^{(s)}} \min\{1, \|x_{t,a_t}\|_{A_{s,t}^{-1}}\} \\
&\overset{(a)}{\leq} \left(\alpha + 2C \sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \right) \sqrt{|\Psi_T^{(s)}| \sum_{t \in \Psi_T^{(s)}} \min\{1, \|x_{t,a_t}\|_{A_{s,t}^{-1}}^2\}} \\
&\overset{(b)}{\leq} \left(\alpha + 2C \sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \right) \sqrt{2d|\Psi_T^{(s)}| \log\left(1 + \frac{M}{d}|\Psi_T^{(s)}|\right)} \\
&\leq \left(\alpha + 2C \sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \right) \sqrt{2d|\Psi_T^{(s)}| \log\left(2M|\Psi_T^{(s)}|\right)},
\end{aligned} \tag{77}$$

where (a) follows from Jensen's inequality, and (b) follows from an application of Lemma 14. Reorganizing the resulting inequality above leads to the desired claim. □

We are now ready to prove Theorem 4.

Proof. (**Proof of Theorem 4**) Since $S = \lceil \log T \rceil$, we have $1/(2^S \sqrt{M}) \leq 1/(\sqrt{MT})$. Thus, from the rules of Algorithm 3, it is apparent that at every time-step t, an action a_t is *always* chosen, either based on line 6, or on line 7. If we use Ξ_T to store those time steps in $[T]$ where an action is chosen based on line 7 of Algorithm 3, then the above reasoning implies: $[T] = \Xi_T \cup \bigcup_{s \in [S]} \Psi_T^{(s)}$.

Throughout the rest of the proof, we will condition on the clean event on which items (i)-(iii) in Lemma 13 hold. We also recall that this clean event has measure at least $1 - \bar{\delta} KST$. Now fix any

good agent $i \in [M] \setminus \mathcal{B}$, and note that the same action a_t is played by every good agent at time t. The cumulative regret for agent i can thus be decomposed as follows:

$$\sum_{t=1}^{T} \left(\langle \theta_*, x_{t,a_t^*} \rangle - \langle \theta_*, x_{t,a_{i,t}} \rangle \right) = \underbrace{\sum_{t \in \Xi_T} \left(\langle \theta_*, x_{t,a_t^*} \rangle - \langle \theta_*, x_{t,a_t} \rangle \right)}_{T_1} + \underbrace{\sum_{s=1}^{S} \sum_{t \in \Psi_T^{(s)}} \left(\langle \theta_*, x_{t,a_t^*} \rangle - \langle \theta_*, x_{t,a_t} \rangle \right)}_{T_2}. \tag{78}$$

Let us first bound T_2 as follows.

$$\begin{aligned}
T_2 &= \sum_{s=1}^{S} \sum_{t \in \Psi_T^{(s)}} \left(\langle \theta_*, x_{t,a_t^*} \rangle - \langle \theta_*, x_{t,a_t} \rangle \right) \\
&\overset{(a)}{\leq} \sum_{s=1}^{S} \sum_{t \in \Psi_T^{(s)}} \frac{8}{2^s \sqrt{M}} \\
&= \sum_{s=1}^{S} \frac{8}{2^s \sqrt{M}} |\Psi_T^{(s)}| \\
&\overset{(b)}{\leq} \sum_{s=1}^{S} 8 \left(\alpha + 2C \sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \right) \sqrt{2d |\Psi_T^{(s)}| \log(2M |\Psi_T^{(s)}|)} \\
&\overset{(c)}{\leq} 8S \left(\alpha + 2C \sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \right) \sqrt{2dT \log(2MT)} \\
&\leq 8(1 + \log(T)) \left(\alpha + 2C \sqrt{\frac{\log\left(\frac{1}{\bar{\delta}}\right)}{M}} \right) \sqrt{2dT \log(2MT)},
\end{aligned} \tag{79}$$

where for (a), we used item (iii) of Lemma 13; for (b), we invoked Lemma 15 to bound $|\Psi_T^{(s)}|$; and for (c), we used the trivial bound $|\Psi_T^{(s)}| \leq T$. With $\bar{\delta} = \delta/(KST)$, the bound on T_2 reads as follows:

$$T_2 = O\left(\left(\alpha + 2C \sqrt{\frac{\log\left(\frac{KT}{\delta}\right)}{M}} \right) \log(T) \sqrt{2dT \log(2MT)} \right). \tag{80}$$

Now let us turn to bounding T_1. Consider any time-step $t \in \Xi_T$ where the action a_t is chosen based on line 7 of Algorithm 3. Since a_t^* is never eliminated on the clean event (item (ii) of Lemma 13), and since a_t has the highest robust upper confidence bound among all active arms, it must be that:

$$\begin{aligned}
& \hat{r}_{t,a_t^*}^{(s)} + w_{t,a_t^*}^{(s)} \leq \hat{r}_{t,a_t}^{(s)} + w_{t,a_t}^{(s)} \\
\implies & \langle \theta_*, x_{t,a_t^*} - x_{t,a_t} \rangle \leq 2 w_{t,a_t}^{(s)} \\
\implies & \langle \theta_*, x_{t,a_t^*} - x_{t,a_t} \rangle \leq \frac{2}{\sqrt{MT}},
\end{aligned} \tag{81}$$

where for the first implication, we used item (i) of Lemma 13; and for the second, we used the fact for an action to be chosen based on line 7, it's confidence width must be bounded above by $1/(\sqrt{MT})$. We conclude that

$$T_1 = \sum_{t \in \Xi_T} \left(\langle \theta_*, x_{t,a_t^*} \rangle - \langle \theta_*, x_{t,a_t} \rangle \right) = O\left(\sqrt{\frac{T}{M}} \right).$$

Combining the above bound on T_1 with that on T_2 in Eq. (80), we have that with probability at least $1-\delta$, the following is true for each good agent $i \in [M] \setminus \mathcal{B}$:

$$\begin{aligned}\sum_{t=1}^{T}\left(\langle\theta_*, x_{t,a_t^*}\rangle - \langle\theta_*, x_{t,a_t}\rangle\right) &= O\left(\left(\alpha + 2C\sqrt{\frac{\log\left(\frac{KT}{\delta}\right)}{M}}\right)\log(T)\sqrt{2dT\log(2MT)}\right) \\ &= \tilde{O}\left(\left(\alpha + \sqrt{1/M}\right)\sqrt{dT}\right).\end{aligned} \tag{82}$$

This completes the proof. □

G Alternate Strategies for Robust Collaborative Phased Elimination can lead to Sub-Optimal Regret Bounds.

In Section 3, where we introduced RCLB, we alluded to the fact that certain natural candidate strategies may lead to sub-optimal regret bounds. In this section, we elaborate on this point. Note that our end goal is to come up with a phased elimination step akin to line 9 of RCLB. To achieve this, in every epoch ℓ, we need estimates of $\langle \theta_*, a \rangle$ along with associated confidence intervals for each $a \in \mathcal{A}_\ell$. In what follows, we will consider two natural candidate strategies for the same, and demonstrate that they each lead to confidence bounds that are looser than the ones we derived in Lemma 1. As such, using such bounds in the phased elimination step will lead to regret guarantees that are sub-optimal in their dependence on the model dimension d.

• **Candidate Strategy 1.** Suppose in every epoch ℓ, the server collects the local model estimates $\{\hat{\theta}_i^{(\ell)}\}_{i \in [M]}$, and aims to first construct a robust estimate $\hat{\theta}^{(\ell)}$ of θ_*. Subsequently, it uses $\langle \hat{\theta}^{(\ell)}, a \rangle$ as an estimate of $\langle \theta_*, a \rangle$ for each active arm $a \in \mathcal{A}_\ell$. To extract $\hat{\theta}^{(\ell)}$ from the local estimates $\{\hat{\theta}_i^{(\ell)}\}_{i \in [M]}$, we need a high-dimensional robust mean estimator. One natural candidate for this is the Iteratively Reweighted Mean Estimator from [22] since it leads to minimax-optimal error bounds. However, there is an immediate obstacle to directly applying the estimator from [22] on the model estimates $\{\hat{\theta}_i^{(\ell)}\}_{i \in [M]}$. This stems from the observation that although $\hat{\theta}_i^{(\ell)}$ is an unbiased estimate of θ_* for each good agent i, the covariance matrix associated with such an estimate may be *ill-conditioned*. In particular, it is not hard to verify that for each $i \in [M] \setminus \mathcal{B}$:

$$\mathbb{E}\left[\left(\hat{\theta}_i^{(\ell)} - \theta_*\right)\left(\hat{\theta}_i^{(\ell)} - \theta_*\right)' \middle| \mathcal{F}_\ell\right] = \tilde{V}_\ell^{-1}.$$

Thus, if we were to construct $\hat{\theta}^{(\ell)}$ as

$$\hat{\theta}^{(\ell)} = \texttt{ITW}(\{\hat{\theta}_i^{(\ell)},\, i \in [M]\}),$$

then based on Lemma 8, the error bound $\|\hat{\theta}^{(\ell)} - \theta_*\|$ would scale with $\|\tilde{V}_\ell^{-1}\|_2^{1/2}$.[8] This is undesirable as $\|\tilde{V}_\ell^{-1}\|_2^{1/2}$ can potentially take on a large value. To bypass this problem, we can use the same trick as we did for RC-GLM, and compute $\hat{\theta}^{(\ell)}$ as follows:

$$\hat{\theta}^{(\ell)} = \tilde{V}_\ell^{-1/2}\left(\texttt{ITW}(\{\tilde{V}_\ell^{1/2}\hat{\theta}_i^{(\ell)},\, i \in [M]\})\right).$$

The rationale behind the above approach is that the covariance matrix associated with $\tilde{V}_\ell^{1/2}\hat{\theta}_i^{(\ell)}, i \in [M] \setminus \mathcal{B}$, is I_d. Using Lemma 8, and following similar arguments as used to arrive at Lemmas 1 and 10, one can show that for each $a \in \mathcal{A}_\ell$, with probability at least $1 - \delta_\ell$, it holds that:

$$|\langle \hat{\theta}^{(\ell)}, a \rangle - \langle \theta_*, a \rangle| = O\left(\left(\sqrt{d} + \alpha\sqrt{M \log(1/\alpha)}\right)\epsilon_\ell\right). \tag{83}$$

It is instructive to compare the above estimate on $\langle \theta_*, a \rangle$ with the estimate $\mu_a^{(\ell)}$ we used in RCLB. Specifically, recall from Lemma 1 that for each $a \in \mathcal{A}_\ell$, with probability at least $1 - \delta_\ell$, it holds that

$$|\mu_a^{(\ell)} - \langle \theta_*, a \rangle| = O\left(\left(1 + \alpha\sqrt{M}\right)\epsilon_\ell\right).$$

Comparing the above upper bound with the one in Eq. (83), we note that while the former is independent of the dimension d, the latter does exhibit a dependence via the $\sqrt{d}$ term. Now suppose we use the upper-bound from Eq. (83) to construct a robust confidence threshold - say $\tilde{\gamma}_\ell$ - and use it to devise a phased elimination step as the one in line 9 of RCLB. Then, following the reasoning as that used to prove Theorem 1, one can establish a per-agent regret bound of

$$\tilde{O}\left(\left(\alpha\sqrt{\log(1/\alpha)} + \sqrt{\frac{d}{M}}\right)\sqrt{dT}\right),$$

which is unfortunately weaker than the guarantee we have in Theorem 1.

[8] Recall that we use $\texttt{ITW}(\cdot)$ to represent the output of the Iteratively Reweighted Mean Estimator from [22], namely Algorithm 5.

• **Candidate Strategy 2.** The main idea is as follows. In every epoch ℓ, the server queries each agent $i \in [M]$ to report their aggregate observation $r_{i,a}^{(\ell)}$ for each arm $a \in \text{Supp}(\pi_\ell)$. Recall that $r_{i,a}^{(\ell)}$ is the average of the rewards for arm a observed by agent i during epoch ℓ. The server next computes an aggregate "clean" observation $\tilde{r}_a^{(\ell)}$ for each $a \in \text{Supp}(\pi_\ell)$ as follows:

$$\tilde{r}_a^{(\ell)} = \texttt{Median}\left(\{r_{i,a}^{(\ell)}, i \in [M]\}\right).$$

It then uses these clean observations to compute an estimate $\hat{\theta}^{(\ell)}$ of θ_* as follows:

$$\hat{\theta}^{(\ell)} = \bar{V}_\ell^{-1} Y_\ell, \text{ where } \bar{V}_\ell = \sum_{a \in \text{Supp}(\pi_\ell)} T_a^{(\ell)} a a' \ ; \ Y_\ell = \sum_{a \in \text{Supp}(\pi_\ell)} T_a^{(\ell)} \tilde{r}_a^{(\ell)} a,$$

and

$$T_a^{(\ell)} = \left\lceil \frac{\pi_\ell(a) d}{\epsilon_\ell^2} \log\left(\frac{1}{\delta_\ell}\right) \right\rceil.$$

The quantity $\hat{\theta}^{(\ell)}$ obtained above is now used to compute $\langle \hat{\theta}^{(\ell)}, a \rangle$ as an estimate of the true mean payoff $\langle \theta_*, a \rangle$ of each arm $a \in \mathcal{A}_\ell$.[9] As before, our goal is to bound $|\langle \hat{\theta}^{(\ell)}, a \rangle - \langle \theta_*, a \rangle|$ for each $a \in \mathcal{A}_\ell$. To that end, we start by noting that for each good agent i, $r_{i,a}^{(\ell)} \sim \mathcal{N}\left(\langle \theta_*, a \rangle, \frac{1}{m_a^{(\ell)}}\right)$. Invoking Lemma 2 then tells us that with probability at least $1 - \delta_\ell$,

$$|\tilde{r}_a^{(\ell)} - \langle \theta_*, a \rangle| \leq C \left(\alpha + \sqrt{\frac{\log(\frac{1}{\delta_\ell})}{M}}\right) \frac{1}{\sqrt{m_a^{(\ell)}}} \leq C \left(\alpha\sqrt{M} + \sqrt{\log\left(\frac{1}{\delta_\ell}\right)}\right) \frac{1}{\sqrt{T_a^{(\ell)}}}. \tag{84}$$

For our subsequent discussion, let us condition on the event on which the above bound holds for every arm in $\mathcal{A}_\ell$. On this event, we can say that for each $a \in \mathcal{A}_\ell$, $\tilde{r}_a^{(\ell)} = \langle \theta_*, a \rangle + e_a^{(\ell)}$, where $e_a^{(\ell)}$ is an error term satisfying the bound in Eq. (84). Now fix any $b \in \mathcal{A}_\ell$. Simple calculations reveal that:

$$\begin{aligned}
|\langle \hat{\theta}^{(\ell)} - \theta_*, b \rangle| &= \Big| \sum_{a \in \text{Supp}(\pi_\ell)} T_a^{(\ell)} \langle \bar{V}_\ell^{-1} a, b \rangle e_a^{(\ell)} \Big| \\
&\leq \sum_{a \in \text{Supp}(\pi_\ell)} T_a^{(\ell)} |\langle \bar{V}_\ell^{-1} a, b \rangle| |e_a^{(\ell)}| \\
&\overset{(a)}{\leq} C \left(\alpha\sqrt{M} + \sqrt{\log\left(\frac{1}{\delta_\ell}\right)}\right) \underbrace{\sum_{a \in \text{Supp}(\pi_\ell)} \sqrt{T_a^{(\ell)}} |\langle \bar{V}_\ell^{-1} a, b \rangle|}_{T_1} \\
&\overset{(b)}{\leq} C \left(\alpha\sqrt{M} + \sqrt{\log\left(\frac{1}{\delta_\ell}\right)}\right) \sqrt{|\text{Supp}(\pi_\ell)| \left(b' \bar{V}_\ell^{-1} \left(\sum_{a \in \text{Supp}(\pi_\ell)} T_a^{(\ell)} a a'\right) \bar{V}_\ell^{-1} b\right)} \\
&\leq C \left(\alpha\sqrt{M} + \sqrt{\log\left(\frac{1}{\delta_\ell}\right)}\right) \sqrt{|\text{Supp}(\pi_\ell)|} \, \|b\|_{\bar{V}_\ell^{-1}} \\
&\overset{(c)}{\leq} C \left(\alpha\sqrt{M} + \sqrt{\log\left(\frac{1}{\delta_\ell}\right)}\right) \sqrt{48 d \log\log d} \, \|b\|_{\bar{V}_\ell^{-1}} \\
&\overset{(d)}{=} \tilde{O}\left(\sqrt{d}\left(1 + \alpha\sqrt{M}\right)\epsilon_\ell\right).
\end{aligned} \tag{85}$$

[9] Note that the observations obtained from each agent i, namely $r_{i,a}^{(\ell)}, a \in \text{Supp}(\pi_\ell)$, provide direct information about the mean payoffs of arms *only* in $\text{Supp}(\pi_\ell)$. However, for the phased elimination step, we need estimates of the mean payoffs of *all arms in* $\mathcal{A}_\ell$, not just the ones in $\text{Supp}(\pi_\ell) \subseteq \mathcal{A}_\ell$. This is precisely why we need to go through an intermediate regression step to first compute an estimate of θ_*.

In the above steps, for (a) we used the bound from Eq. (84); for (b), we used Jensen's inequality; for (c), we used the fact that $|\text{Supp}(\pi_\ell)| \leq 48d \log\log d$; and for (d), following a similar argument as in the proof of Lemma 1, we used that

$$\|b\|_{\bar{V}_\ell^{-1}} = O\left(\frac{\epsilon_\ell}{\sqrt{\log\left(1/\delta_\ell\right)}}\right).$$

Comparing the bound in Eq. (85) with the one in Lemma 1, we once again note that while the latter bound is d-independent, the former has a clear dependence on $\sqrt{d}$. At the risk of sounding repetitive, if one were to employ the bound in Eq. (85) to construct a confidence threshold for phased elimination, and run through the same arguments as in the proof of Theorem 1, one would end up with a per-agent regret bound of

$$\tilde{O}\left(\left(1 + \alpha\sqrt{M}\right) d\sqrt{T}\right).$$

Unlike the near-optimal guarantee we have in Theorem 1, the above bound is clearly off by a factor of $\sqrt{d}$ from the optimal dependence on the model dimension d. The looseness in the bound mainly stems from the following fact: the error terms $\{e_a^{(\ell)}\}_{a \in \text{Supp}(\pi_\ell)}$ are not necessarily sub-Gaussian random variables that are independent across arms. One can contrast this to the analysis in Lemma 1, where the noise terms $\{\bar{\eta}_{i,a}^{(\ell)}\}_{a \in \text{Supp}(\pi_\ell)}$ were in fact Gaussian, and independent across arms. It is precisely the lack of nice statistical properties for the error terms $\{e_a^{(\ell)}\}_{a \in \text{Supp}(\pi_\ell)}$ that compels us to use Jensen's inequality to bound the term T_1 in Eq. (85). At the moment, it is unclear to us whether one can come up with a tighter bound for this candidate strategy.

Main Takeaway. The main message from this section is that deriving robust confidence intervals that lead to near-optimal bounds (such as the one in Theorem 1) is non-trivial, and requires a lot of care. In particular, the above discussion serves to highlight the significance of our algorithmic approach.

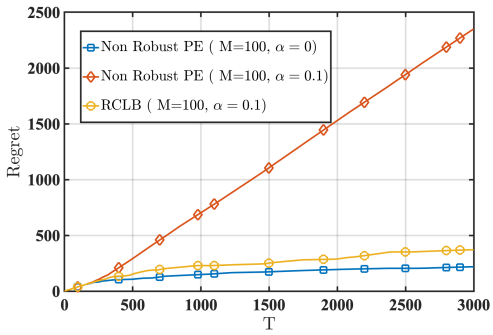


Figure 2: Performance of a vanilla non-robust distributed phased elimination algorithm vs RCLB for the attack model in Eq. (86).

H Experimental Results

In this section, we will provide various simulation results on synthetic data to corroborate the theory developed in our work. We start by describing the experimental setup for the linear bandit setting considered in Section 7.

H.1 Experiments for the Linear Bandit Setting

Linear Bandit Experimental Setup. We generate 50 arms $a_1, a_2, \ldots, a_{50} \in \mathbb{R}^d$, where $d = 5$. Each arm $a_j, j \in [50]$, is generated by drawing each of the arm's coordinates i.i.d from the interval $[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]$. It thus follows that $\|a_j\|_2 \leq 1, \forall j \in [50]$. The model parameter θ_* is chosen to be a 5-dimensional vector with each entry equal to $1/\sqrt{d}$. The rewards are generated based on the observation model in Eq. (1). We now describe the attack model for the linear bandit setting.

Attack Model for Linear Bandit Setting. The collective goal of the adversarial agents is to manipulate the server into selecting sub-optimal arms. To that end, each adversarial agent i employs the simple strategy of reducing the rewards of the good arms and increasing the rewards of the bad arms. More precisely, in each epoch ℓ, upon pulling an arm a and observing the corresponding reward y_a, an adversarial agent i does the following: if $y_a > p\langle\theta_*, a_*\rangle$, then this reward is corrupted to $\tilde{y}_a = y_a - \beta$; and if $y_a \leq p\langle\theta_*, a_*\rangle$, then the reward is corrupted to $\tilde{y}_a = y_a + \beta$. For this experiment, we fix $p = 0.6$ and $\beta = 5$. Agent $i \in \mathcal{B}$ then uses all the corrupted rewards in epoch ℓ to generate the local model estimate $\hat{\theta}_i^{(\ell)}$ that is transmitted to the server.

An Alternate Attack Model. To further test the robustness of RCLB, we consider an attack model different from the one in Section 7. In this attack, in each epoch ℓ, every adversarial agent $i \in \mathcal{B}$ generates and transmits the following corrupted local model estimate to the server:

$$\hat{\theta}_i^{(\ell)} = -\frac{M}{|\mathcal{B}|}\theta_* - \frac{1}{|\mathcal{B}|}\sum_{j \in [M]\setminus\mathcal{B}} \hat{\theta}_j^{(\ell)}. \tag{86}$$

The idea behind the above attack is to trick the server into thinking that the true model estimate is $-\theta_*$, as opposed to θ_*, by shifting the average of the agents' local model estimates towards $-\theta_*$. As we can see from Figure 2, the adversarial agents succeed in doing so when one employs a vanilla non-robust distributed phased elimination algorithm. However, our proposed approach RCLB continues to remain immune to such attacks, and guarantees sub-linear regret as suggested by our theory.

H.2 Experiments for the Contextual Bandit Setting

The goal of this section is to validate our proposed robust collaborative algorithm for the contextual bandit setting, namely Algorithm 3.

Contextual Bandit Experimental Setup. As in the linear bandit experiment, we set the number of arms K to be 50, the model dimension d to be 5, and the true parameter θ_* to be a d-dimensional

vector with each entry equal to $1/\sqrt{d}$. At each time-step t, for each $a \in \mathcal{A}$, we generate the feature vector $x_{t,a}$ by drawing each of its entries i.i.d from the interval $[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]$. The rewards are then generated based on the observation model in Eq. (12).

Attack Model for Contextual Bandit Setting. We use an attack strategy similar in spirit to the first attack model for the linear bandit setting. Specifically, at each time-step t, each adversarial agent $i \in \mathcal{B}$ does the following: if $r_{i,t} > p\langle \theta_*, x_{t,a_t^*} \rangle$, then the attacker sets $\tilde{r}_{i,t} = r_{i,t} - \beta$; if $r_{i,t} < p\langle \theta_*, x_{t,a_t^*} \rangle$, then the attacker sets $\tilde{r}_{i,t} = r_{i,t} + \beta$. The corrupted reward $\tilde{r}_{i,t}$ is then sent to the server. In this experiment, we fixed $p = 0.6$ and $\beta = 5$.

Discussion of Simulation Results. Figure 3 illustrates the results for the contextual linear bandit experiment. In Figure 3(a), we compare our proposed algorithm, namely Algorithm 3, to a naive distributed implementation of Algorithm 3 that does not account for adversarial agents. Similar to what we observed in Figure 1(a), while the non-robust algorithm incurs linear regret in the presence of adversaries, Algorithm 3 continues to guarantee sub-linear regret bounds. The plots in Figures 3(b)-(d) are analogous to the ones in Figures 1(b)-(d). In short, these plots once again indicate a clear benefit of collaboration (for small α) in the presence of adversarial agents, thereby highlighting the importance of Algorithm 3, and validating Theorem 4.

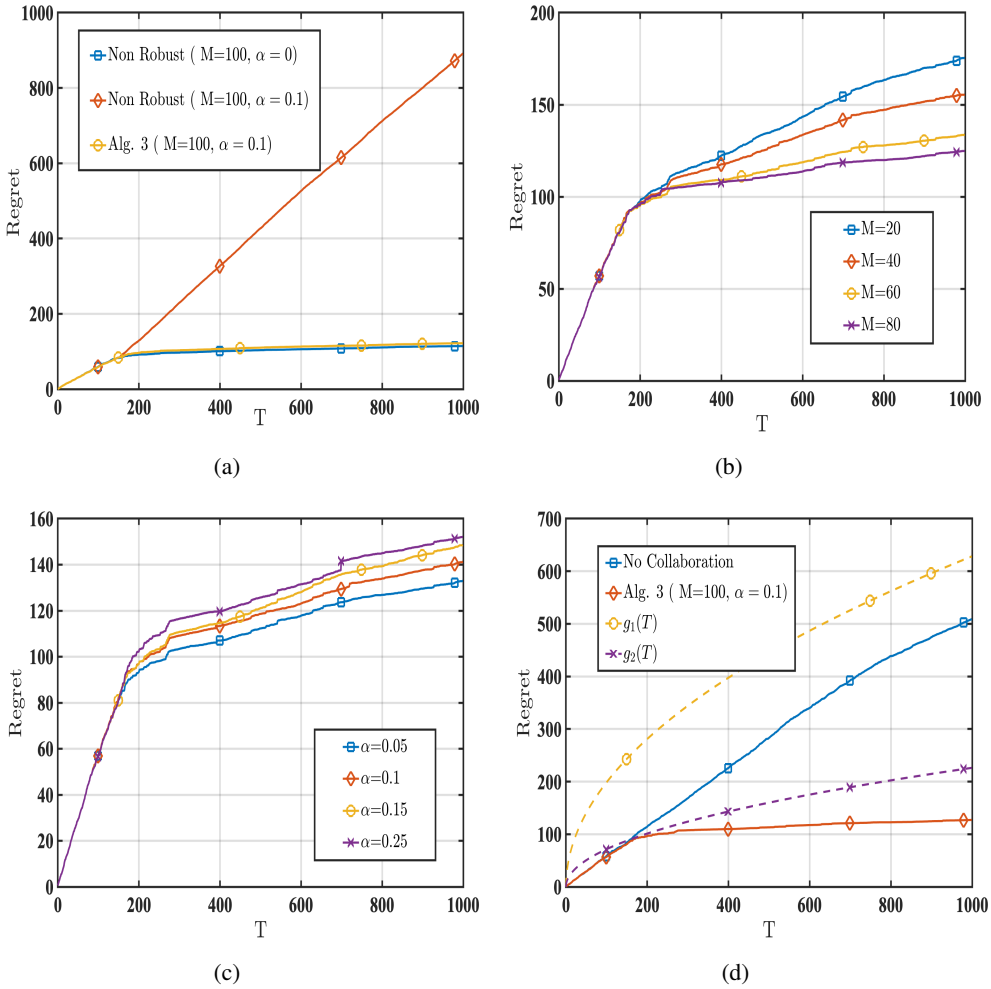


Figure 3: Plots of per-agent regret for the contextual bandit experiment. (a) Comparison between our proposed algorithm, namely Algorithm 3, and a vanilla non-robust distributed contextual bandit algorithm. (b) Performance of Algorithm 3 for varying number of agents M, with $\alpha = 0.1$. (c) Performance of Algorithm 3 for varying corruption fraction α, with $M = 100$. (d) Comparison of Algorithm 3 to a non-robust contextual bandit algorithm where the agents do not collaborate; here, $\alpha = 0.1$ and $M = 100$. We also plotted theoretical upper bounds: $g_1(T) = 3\sqrt{dT}$ and $g_2(T) = 17(\alpha + \sqrt{\frac{1}{M}})\sqrt{dT}$.