
Coordinated Attacks against Contextual Bandits: Fundamental Limits and Defense Mechanisms

Jeongyeol Kwon¹ Yonathan Efroni² Constantine Caramanis¹ Shie Mannor^{3,4}

Abstract

Motivated by online recommendation systems, we propose the problem of finding the optimal policy in multitask contextual bandits when a small fraction $\alpha < 1/2$ of tasks (users) are arbitrary and adversarial. The remaining fraction of good users share the same instance of contextual bandits with S contexts and A actions (items). Naturally, whether a user is good or adversarial is not known in advance. The goal is to robustly learn the policy that maximizes rewards for good users with as few user interactions as possible. Without adversarial users, established results in collaborative filtering show that $O(1/\epsilon^2)$ per-user interactions suffice to learn a good policy, precisely because information can be shared across users. This parallelization gain is fundamentally altered by the presence of adversarial users: unless there are super-polynomial number of users, we show a lower bound of $\tilde{\Omega}(\min(S, A) \cdot \alpha^2/\epsilon^2)$ per-user interactions to learn an ϵ -optimal policy for the good users. We then show we can achieve an $\tilde{O}(\min(S, A) \cdot \alpha/\epsilon^2)$ upper-bound, by employing efficient robust mean estimators for both uni-variate and high-dimensional random variables. We also show that this can be improved depending on the distributions of contexts.

1. Introduction

Online recommendation systems (Adomavicius & Tuzhilin, 2005; Li et al., 2010) are ubiquitous, and used in diverse platform applications including video streaming, online shopping, travel and restaurant recommendations. These are large-scale systems, designed for millions of users and

items. Thus, successful algorithms exploit shared preferences across different users (as in collaborative filtering). This naturally leads to a multitask contextual bandit framework (Sarwar et al., 2001; Maillard & Mannor, 2014; Sen et al., 2017; Chawla et al., 2020; Yang et al., 2020; Ghosh et al., 2021b). Manipulation of collaborative filtering recommendation engines is well documented e.g., (O’Mahony et al., 2004; Mobasher et al., 2007; Burke et al., 2015), and various approaches to building resilient algorithms have been considered e.g., (Olsen, 2002; Van Roy & Yan, 2010; Chen et al., 2015). The ability to learn across users changes dramatically even in the presence of a small fraction of adversarial users, thus posing fundamental new challenges. This is precisely the problem we address.

Problem Setup. We consider a multitask contextual bandit framework where the majority of users share preferences over all contexts and items. We call these the “good” users. Some α -fraction of users ($\alpha < 1/2$), however, may have arbitrary preferences or may even attempt to manipulate the recommendation system; we call these the “adversarial” users.

Formally, our system consists of L users, a set of contexts \mathcal{S} , and a set of actions \mathcal{A} . Contexts are the temporal status of the user such as browsing histories, keywords, categories, etc. and the key assumption is that the good users share preferences across contexts, thus enabling cross-user learning (collaborative filtering). We note that, in particular, the user ID is a unique identifier, and not part of the context (in contrast to, e.g., (Sen et al., 2017; Deshmukh et al., 2017)). At each step, a user $i \in [L]$ arrives and reveals its current status (context) $s \in \mathcal{S}$. Then, based on the context, we select an action $a \in \mathcal{A}$ and get a reward (feedback) from the user. If the user is good, then the mean-reward returned by this user is $\mu(s, a) \in \mathbb{R}$. If the user is adversarial, it can return any reward based on all interaction histories.

Goal of the Paper. Our goal is to find a provably-approximately-correct (PAC) policy for the majority of good users, with the minimum per-user interactions. Without adversarial users, *i.e.*, assuming all users share preferences across items and contexts, the problem is the well-studied contextual bandit problem (Lu et al., 2010; Chu et al., 2011).

¹Department of Electrical and Computer Engineering, University of Texas at Austin, USA ²Microsoft Research, New York ³Technion, Israel ⁴Nvidia Research. Correspondence to: Jeongyeol Kwon <kwonchungli@utexas.edu>, Yonathan Efroni <jonathan.efroni@gmail.com>.

Classical results show that as soon as we have enough users, $L \geq SA$, the common preferences can be exploited, and existing contextual bandit algorithms (Auer et al., 2002; Zhou, 2015) can find a good policy after $O(SA/L) \sim O(1)$ per-user interactions. That is, we can achieve $O(1/L)$ *parallelization gain*, the ratio between required per-user interactions with a single user ($L = 1$) and many users.

However, exploiting common preferences to reduce per-user interactions makes us susceptible to manipulation: a small but constant fraction $\alpha = O(1)$ of adversarial users can easily manipulate the state of the art contextual bandit algorithms. Moreover, as we show, a full parallelization gain of $O(1/L)$ -factor is not possible even with a single context, for any algorithm if $\alpha > 0$. On the other hand, we can always abandon parallelization gain, and learn an optimal policy for each user, with $O(SA/\epsilon^2)$ per-user interactions, thus learning is indeed possible. Our goal is to investigate the extent to which we can beat this naive baseline, while maintaining robustness in the presence of adversarial users.

Main Results. We show that using recent ideas from robust statistics, we can design a robust contextual bandit algorithm that requires $O(\min(S, A) \cdot \alpha/\epsilon^2)$ per-user interactions, i.e., it achieves a $O(\alpha/\max(S, A))$ parallelization gain. As this falls short of the $1/L$ -parallelization gain in the absence of manipulators, a fundamental question is whether this is the best possible we can achieve. We show that the scaling in A and S cannot be improved, and we give a lower bound that matches up to one factor of α . We summarize our results:

- **Robust Algorithm:** We initially consider two regimes: $A \gg S$, and $S \gg A$. In the first setting, we show that partitioning the users into A groups, and within each group playing the same arm $1/\epsilon^2$ times. Then using median estimator, we show we can obtain ϵ -accurate estimates of each arm in each context, thus giving a $O(1/A)$ parallelization gain. When $S \gg A$, we can do better than $(1/A)$ gain. Yet we need a different idea, since we may not see the same user–context pair enough times. Instead, we show we can leverage recent results in high dimensional mean estimation (e.g., (Diakonikolas et al., 2019; Depersin & Lecué, 2022)), to get a $O(1/S)$ parallelization gain.
- **Fundamental Limits:** Can the $\max(S, A)$ parallelization gain be further improved once sufficiently many users $L = \Omega(SA \log(SA))$ are given? We show that, perhaps surprisingly, $\tilde{\Omega}(\min(S, A) \cdot \alpha^2/\epsilon^2)$ per-user sample complexity cannot be improved unless super-polynomial number of users are available. This is the first kind of negative results of learning contextual bandits in multitask settings (Section 4).

- **Problem Dependent Results:** In more practical scenarios, a certain set of contexts might be more frequently observed. For instance, there can be certain genres, keywords or categories that are more popular and frequently searched by users, while less popular contexts are rarely searched and thus less critical for the overall performance. We show that we can unify efficient uni-variate and high-dimensional robust estimators, and obtain the improved problem-dependent per-user sample-complexity (Section 5).

1.1. Comparison to Previous Work

Multitask contextual bandits have attracted significant recent attention, for a variety of applications. We only review the most closely related theoretical results on this topic.

Multitask Contextual Bandits. Multitask contextual bandits have been considered largely in two problem settings: (i) tasks of all user preferences are assumed to be embedded in low dimensional spaces (Sen et al., 2017; Gopalan et al., 2016; Yang et al., 2020; Hu et al., 2021), (ii) users can be clustered into a small number of groups (compared to the number of users) that share the same instance of contextual bandits (Maillard & Mannor, 2014; Gentile et al., 2014; 2017; Ghosh et al., 2021b). In the former setting (i), low-dimensional representations of tasks are the key objective to recover, and this has been done with general-purpose techniques such as tensor-decomposition or low-rank factorization. The adversarial users destroy this low dimensionality, hence these methods are not directly applicable to our setting. The latter setting (ii), parallelization gain leverages preference similarity within clusters (Maillard & Mannor, 2014; Gentile et al., 2014; 2017; Ghosh et al., 2021b). While our problem is more closely related to this setting, again the challenge comes from the adversarial users. As their behavior can be arbitrary, we thus may have $O(L)$ unbalanced clusters among L users, thus destroying our ability to leverage clustering. And indeed, the referenced work requires a small (in fact, a constant) number of clusters. We note that even in our work, we never identify the good vs. adversarial users.

Corruption Robust Bandits. Corruption robust bandit algorithms (Gupta et al., 2019; Lykouris et al., 2018; 2021; Ma, 2021; Liu et al., 2021) consider the setting where the reward feedback can be corrupted at any time by an adversary with limited budget. There is little that can be done when the corruption budget scales linearly with the number of interactions. If we cast our problem in this setting, then though the corruption budget scales linearly, our adversary is limited and cannot corrupt the rewards of the good users. This additional information eventually allows us to learn the best policy for the underlying contextual bandit even with

large total amount of corruptions. Necessarily, the strategies we develop are fundamentally different than those in the works references above.

Exploration in Contextual Bandits. There is a long line of work that studies efficient exploration algorithms for contextual bandits (and reinforcement learning) in both stochastic and adversarial reward settings (Sutton & Barto, 2018; Auer et al., 2002; Abbasi-Yadkori et al., 2011; Even-Dar et al., 2006; Lattimore & Szepesvári, 2020; Audibert et al., 2009; Bubeck & Slivkins, 2012; Gerchinovitz & Lattimore, 2016). In the adversary-free setting, *i.e.*, $\alpha = 0$, we can use any of these algorithms, ignoring the user identifier. However, with $\alpha = O(1)$ fraction of adversarial users, algorithms that ignore user identifiers can be easily manipulated.

Other Related Work. Contextual bandits have also been studied when rewards are represented as a linear combination of features of users and action items (Abbasi-Yadkori et al., 2011; Chu et al., 2011). In this work, we do not assume that any prior information (such as the features of context-actions) is given in advance. Another line of work considers regret-minimization problem in latent bandits/MDPs where we can interact with each user for a fixed short time-horizon (Brunskill & Li, 2013; Kwon et al., 2021b; Zhou et al., 2021; Kwon et al., 2021a). In contrast, we allow as many interactions with each user as needed, while minimizing the number of per-user interactions.

2. Preliminaries

We consider a setting with L users. Of these, a $(1 - \alpha)$ -fraction, called *good* users, have the same preferences, and hence follow a multitask contextual bandit framework $\mathcal{B} := (\mathcal{S}, \mathcal{A}, \nu, \mathcal{R})$. There are S contexts \mathcal{S} drawn according to distribution ν , and A actions \mathcal{A} . The rewards follow the same distribution $\mathcal{R}(s, a)$ under context s and action a . We use $\mu(s, a)$ to denote the vector of mean rewards for the good users, and we assume that the reward distribution has bounded mean and variance: $|\mu(s, a)| \leq 1$ and $\mathbb{E}_{r \sim \mathcal{R}(s, a)}[(r - \mu(s, a))^2] \leq 1$. The remaining α -fraction of users are adversarial, and do not follow model \mathcal{B} : their returns are not bound by any distribution, and can in fact be a function of the history to that point. Our goal is to find a policy $\pi \in \Pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ that maximizes the expected reward for the good users, *i.e.*, \mathcal{B} :

$$V_{\mathcal{B}}^* = \max_{\pi \in \Pi} V_{\mathcal{B}}^{\pi} := \mathbb{E}_{s \sim \nu} [\mu(s, a) | a \sim \pi(s)].$$

Let $\mathcal{I}^* \subseteq [L]$ be the set of good users where $|\mathcal{I}^*| \geq (1 - \alpha)L$. Our interaction model is defined as follows:

1. At each step $t \in \mathbb{N}$, nature selects a user $i_t \in [L]$ according to some unknown process Γ which call the

user-arrival model. The only restriction on Γ is that the difference in user frequency is controlled – see Assumption 2.1 below.

2. If the user is good ($i_t \in \mathcal{I}^*$), then the user samples a context s_t from ν . Then, the algorithm selects an action a_t based on s_t , and gets reward r_t sampled from $\mathcal{R}(s_t, a_t)$.
3. If the user is adversarial ($i_t \notin \mathcal{I}^*$), then the adversary chooses an arbitrary context $s_t \in \mathcal{S}$, and returns an arbitrary reward $r_t \in \mathbb{R}$.

The adversarial users can choose any $s_t \in \mathcal{S}$ and $r_t \in \mathbb{R}$ based on all histories, the underlying bandit problem \mathcal{B} for good users, and the algorithm, and in particular, they can coordinate. Though the user IDs are fixed, we do not know which users are good and which are adversarial. Moreover, we have no prior information on the underlying shared task of good users in advance.

The only assumption we require for the arrival process across all users, is that the difference in user frequency is controlled.

Assumption 2.1. Let $n_i(t)$ be the number of times that the i^{th} user has interacted with the environment up to time step t . Then, there exists a universal constant $c_L = O(1)$ such that for all $t \geq L$, $\max_{i \in [L]} n_i(t) / \min_{i \in [L]} n_i(t) \leq c_L$.

Remark. The exact constant c_L factors into all of our results as a linear multiplier, as we simply need to wait for the slowest user to accumulate the required interactions. To simplify the notation, we simply take it to equal 1. Our goal is to find an (ϵ, δ) provably-approximately-correct (PAC) optimal policy, which we refer as a near-optimal policy, defined as follows:

Definition 2.2. An algorithm is (ϵ, δ) -PAC if it returns a policy $\hat{\pi}$ such that $\mathbb{P}(V_{\mathcal{B}}^* - V_{\mathcal{B}}^{\hat{\pi}} \leq \epsilon) \geq 1 - \delta$.

Note that the optimal policy π^* is a stationary policy that satisfies $\pi^*(s) = \arg \max_{a \in \mathcal{A}} \mu(s, a)$. The performance of the algorithm is measured in terms of the *number of per-user interactions*: $N = T/L$ where T is the total number of steps the system has taken, to return an (ϵ, δ) -PAC policy. We will only consider the case $\epsilon < \alpha$, as otherwise the problem is straightforward (*e.g.*, we can simply ignore adversaries).

Notation. Let $\text{Unif}(\mathcal{A})$ be a uniform distribution over \mathcal{A} . We use $B(p)$ to denote a Bernoulli distribution with parameter $p \in [0, 1]$. For any probability distribution \mathcal{D} , we use $\mathcal{D}^{\otimes n}$ to mean a n -product distribution of \mathcal{D} . $d_{TV}(\mathcal{D}_1, \mathcal{D}_2)$ is a total-variation distance between two probability distributions $\mathcal{D}_1, \mathcal{D}_2$. We interchangeably use ν and μ as probability distributions and vectors of probabilities.

3. Warm Up: Two Base Cases

In this section, we focus on two base cases when either (i) the number of contexts is small, or (ii) the number of actions is small. We develop an (ϵ, δ) -PAC algorithm robust to adversarial users for arbitrary accuracy $\epsilon > 0$ and small failure probability $\delta > 0$. Together, these two results show that $\tilde{O}(\min(S, A) \cdot \alpha/\epsilon^2)$ number of per-user interactions are enough to obtain an ϵ -optimal policy. This result sets the stage for one of the main results of this paper, that establishes a nearly matching lower bound (Section 4).

3.1. Multi-Armed Bandit: $S = O(1)$, $A \gg 1$

We first consider a special case of contextual bandits, with very few ($O(1)$) contexts; thus \mathcal{B} is essentially a multi-armed bandit problem with many arms. For this case, we can use any efficient *univariate robust estimator*, to calculate the mean-reward of each arm. Once we have that, we can play (nearly) optimally.

The details are as follows. First, assume $S = 1$. Divide the L (good and adversarial) users randomly into A groups, \mathcal{I}_a , $a \in \mathcal{A}$. Whenever we see a user from \mathcal{I}_a , we play $a \in \mathcal{A}$. After T total plays, we compute the empirical mean of each user's rewards. Then we take $\hat{\mu}(a)$ to be the α -trimmed-mean of the empirical estimates of each user in \mathcal{I}_a (Lugosi & Mendelson, 2021). Precisely, we have the following result.

Proposition 3.1. *Let the number of users be at least $L = \Omega(A \log A/\alpha)$ and the adversarial rate be $\alpha < 1/3$. Let $\hat{\mu}$ denote the vector of mean estimates, produced by the procedure outlined above. Then after $T/L = O(\alpha/\epsilon^2)$ per-user interactions, with probability at least $9/10$, $|\hat{\mu}(a) - \mu(a)| \leq \epsilon$ for all $a \in \mathcal{A}$.*

Once we obtain a set of estimated mean-rewards, our policy is immediate: pick the arm $\hat{a} = \arg \max_{a \in \mathcal{A}} \hat{\mu}_a$. This is a $(2\epsilon, 0.1)$ -PAC guaranteed policy.

The procedure can be easily extended to constantly many contexts $S = O(1)$: for each context, we play the same procedure independently, i.e., a user arrival i_t with different s_t can be handled in a separate procedure. This would require $T/L = O(\alpha S/\epsilon^2)$ per-user interactions (assuming a uniform distribution over contexts) to obtain an ϵ -optimal policy, therefore achieving the $O(\alpha/A)$ parallelization gain.

The proof of the above proposition requires us to show that the median-of-means estimator can obtain an ϵ -accurate estimate in the face of as much as an α -fraction of corruptions. Thus, it is essentially an immediate corollary of, for instance, the following result from (Lugosi & Mendelson, 2021), which directly applies to the trimmed-mean estimator:

Theorem 3.2 (Theorem 1 in (Lugosi & Mendelson, 2021)). *Let \mathcal{D} be a distribution on \mathbb{R} with unknown mean μ and*

finite variance σ^2 . Let $\alpha < 1/3$. Given an α -corrupted set of $L = \Omega(\log(1/\delta)/\alpha)$ samples drawn from \mathcal{D} , then the α -trimmed mean produces $\hat{\mu}$ such that with probability at least $1 - \delta$, we have $|\hat{\mu} - \mu| \leq O(\sigma\sqrt{\alpha})$.

3.2. Many Contexts, A Few Actions: $S \gg 1$, $A = O(1)$

The second baseline is a contextual bandit case with many contexts and constant number of arms. Let us assume here that $\nu(s) = 1/S$, i.e., uniform over all contexts. The main challenge here is that we cannot estimate a mean-reward accurately enough from a single user for a fixed state-action pair, since the same context is unlikely to be seen more than once unless we interact with the user long enough (that is, unless $T/L \geq S$). Therefore, a univariate robust estimation approach would not work in this case.

In this scenario, we can use a *robust high-dimensional estimator* (e.g., Cheng et al. (2019); Diakonikolas et al. (2019)) to estimate mean-rewards. Specifically, let $n_i(T) := \sum_{t=1}^T \mathbb{1}_{i_t=i}$ be a set of time steps in which the system interacts with the i^{th} user. We collect data by simply playing a randomly selected action at every step. We then estimate the vector $\hat{\mu}_i$ for all users:

$$\hat{\mu}_i = \frac{SA}{n_i(T)} \sum_{t: i_t=i} r_t \cdot e_{(s_t, a_t)}, \quad (1)$$

where $e_{(s,a)}$ is a standard basis vector with 1 at position (s, a) . For good users $i \in \mathcal{I}^*$, we can see that the first- and second-order moments of a random quantity $\hat{\mu}_i$ satisfy: $\mathbb{E}[r_t \cdot e_{(s_t, a_t)}] = \frac{1}{SA}\mu$, $\text{Cov}(r_t \cdot e_{(s_t, a_t)}) \preceq \frac{1}{SA}I$, and thus

$$\mathbb{E}[\hat{\mu}_i] = \mu, \text{Cov}(\hat{\mu}_i) \preceq \frac{L}{T}SA \cdot I. \quad (2)$$

Hence after T steps, we can equivalently consider a set of $\{\hat{\mu}_i\}_{i=1}^L$ as α -corrupted L independent samples from the same mean- μ distributions with bounded second-order moments. Then, we can use an efficient robust mean estimator for a distribution with bounded second moments developed in (Cheng et al., 2019).

The details are as follows. Every time step $t \in [T]$ a user i_t arrives with context s_t , and we play a random action $a_t \sim \text{Unif}(\mathcal{A})$. After T steps, we construct $\hat{\mu}_i$ as in (1) for all users $i \in [L]$, and run the high-dimensional robust estimator whose existence is guaranteed in Theorem 3.4 below, with input $\{\hat{\mu}_i\}_{i=1}^L$. The quality of the estimated mean-rewards $\hat{\mu}$ is guaranteed by the following:

Proposition 3.3. *Let the number of users be at least $L = \Omega(SA \log(SA)/\alpha)$ and the adversarial rate be $\alpha < 1/3$. After T steps, with probability at least $9/10$, $\|\hat{\mu} - \mu\|_2 \leq O(\sqrt{\alpha LSA/T})$.*

Given this estimate we output a policy $\hat{\pi}$ such that $\hat{\pi}(s) = \arg \max_{a \in \mathcal{A}} \hat{\mu}(s, a)$ for all $s \in \mathcal{S}$. Then a simple algebra

shows that (see Appendix B.2),

$$V_B^* - V_B^{\hat{\mu}} \leq \frac{2}{\sqrt{S}} \|\mu - \hat{\mu}\|_2 \leq O\left(\sqrt{\alpha LA/T}\right). \quad (3)$$

Thus after $T/L = O(\alpha A/\epsilon^2)$ per-user interactions, we obtain an $(\epsilon, 0.1)$ -PAC guaranteed policy with the $O(\alpha/S)$ parallelization gain. Note that the failure probability is chosen for the analysis purpose and can be arbitrarily improved to $1 - \delta$ guarantee by collecting $\log(1/\delta)$ times more interactions and/or repeating the procedure $\log(1/\delta)$ times.

The proof of Proposition 3.3 is essentially a corollary of the following key result on robust mean estimation.

Theorem 3.4 (Theorem 1.3 in (Cheng et al., 2019)). *Let \mathcal{D} be a distribution on \mathbb{R}^d with unknown mean μ and bounded covariance Σ such that $\Sigma \preceq \sigma^2 I$. Given an α -corrupted set of $L = \Omega(d \log(d)/\alpha)$ samples drawn from \mathcal{D} with $\alpha < 1/3$, there is an algorithm that runs in time $\tilde{O}(Ld/\alpha^6)$ and outputs a hypothesis vector $\hat{\mu}$ such that with probability at least $9/10$, it holds $\|\hat{\mu} - \mu\|_2 \leq O(\sigma\sqrt{\alpha})$.*

4. Lower Bound

Through the two warm-up cases, we have seen that $\tilde{O}(\min(S, A) \cdot \alpha/\epsilon^2)$ number of per-user interactions are enough to obtain an ϵ -optimal policy. A natural follow-up question is whether we can improve the sample complexity when there are large number of both contexts and actions $S, A \gg 1$. In this section, we show that this is the best sample complexity we can achieve when all context probabilities are uniform, i.e., $\nu(s) = 1/S, \forall s \in S$:

Theorem 4.1. *Suppose $\alpha < 1/3$ and $L \leq \text{poly}(S, A, 1/\epsilon)$. For any constant $\beta > 0$, there exists a set of α -corrupted multi-user systems such that no algorithm with $T/L = O(\min(S, A)^{1-\beta} \cdot \alpha^2/\epsilon^2)$ per-user interactions can output an ϵ -optimal policy with probability more than $2/3$.*

By Theorem 4.1, we need at least $\Omega(\min(S, A)/\epsilon^2)$ per-user interactions (up to inverse logarithmic factors) to obtain an ϵ -optimal policy with a meaningfully large probability. Note that any algorithm that succeeds with a constant probability can be boosted to a high-probability algorithm by repeating the same procedure, and thus our lower bound also implies the non-existence of an algorithm with a constant success probability with $o(\min(S, A))$ per-user interactions. We prove this result by constructing two systems that are statistically indistinguishable: one is a completely random system (A) where all users only generate rewards from $B(1/2)$. Thus no algorithm can do better than a complete random guess of optimal actions in this system. In the other system (B), good users sample a reward from $B(1/2 + \epsilon)$ for the optimal action given a context, while sampling a reward from $B(1/2)$ for all other actions. We show that in system (B), adversaries can generate sequences of rewards

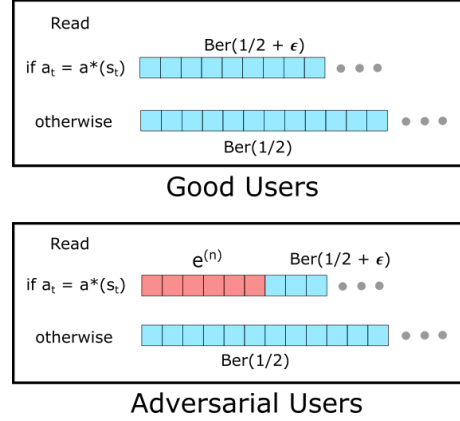


Figure 1: The lower bound construction. The tape describes the data gathered while taking action a . (Blue): indicates an uncorrupted reward, and (Red): indicates corrupted reward for the optimal action. For all non-optimal actions, all users generate rewards from $B(1/2)$.

for the optimal actions such that, without the identity of users (whether they are good or adversarial), the collection of all users' reward sequences are statistically indistinguishable in both systems. We then show that this implies the there cannot be an $(\epsilon/3, 1/3)$ -PAC algorithm.

4.1. Lower Bound for Two-Armed Bandit Case

The proof of Theorem 4.1 starts from a standard construction for the two-armed bandit problem that highlights the necessity of $\Omega(1/\epsilon^2)$ per-user interactions (up to inverse $\log(1/\epsilon)$ factors) for $S = 1, A = 2$. Let $\frac{1}{L} \leq \alpha < \frac{1}{6}$, $L \leq \epsilon^{-O(1)}$, and $\epsilon < \frac{0.01\alpha}{\log L}$ be sufficiently small.

Consider a base system (A) where the reward of all users is sampled from $B(1/2)$ regardless of arms played. Further suppose that a virtual optimal arm $a^* \in \mathcal{A}$ is selected independently and uniformly at random. Clearly, we cannot guess a^* with probability better than $1/2$ since the observed rewards are statistically independent from a^* .

Then, consider the following procedure which generates a corrupted multi-user system (B) as follows (Figure 1):

1. For every $i \in [L]$, with probability $1 - \alpha$, add i to \mathcal{I}^* , i.e., i is a good user. Otherwise, i is an adversarial user.
2. If $i \in \mathcal{I}^*$, then a user i returns a reward $r \sim B(1/2 + \epsilon)$ whenever a^* is played, and $r \sim B(1/2)$ otherwise.
3. If $i \notin \mathcal{I}^*$, then if the chosen action is a^* , it reads the reward sequence $e^{(n)}$ for the first n times, where $e^{(n)}$ is sampled from the distribution $E^{(n)}$ (described in the proof Lemma 4.2). If the action is different than a^* , then it samples a reward $r \sim B(1/2)$.

Note that the corruption by an adversarial user happens only

for the first n times and only for the optimal action. Also note that since the true identity of users are decided following $B(\alpha)$ independently, with high probability, system (B) generates a 2α -corrupted multi-user system.

Now we show that adversaries can play reward sequences such that systems (A) and (B) are not distinguishable. We first need the following lemma on confusing the sequence of product distributions:

Lemma 4.2. *Let $n \leq 0.01\alpha^2/(\epsilon^2 \log L)$. Then there exists a distribution $E^{(n)}$ over $\{0, 1\}^n$ such that*

$$d_{TV}\left(B(1/2)^{\otimes n}, (1 - \alpha) \cdot B(1/2 + \epsilon)^{\otimes n} + \alpha \cdot E^{(n)}\right) \leq 1/L^4. \quad (4)$$

Note that $B(1/2)^{\otimes n}$ corresponds to a distribution of length n reward sequence of any user in the base system. The key fact here is that, if we can only see up to length- n sequences of rewards from all L users, then due to Lemma 4.2 and Le Cam's two-points method (LeCam, 1973), we cannot identify whether a system is (A) or (B) with probability better than $1/2 + 1/L^3 \approx 1/2$.

Now we show by contradiction that no algorithm that interacts with each user less than n times can recover a^* in system (B) with probability more than $2/3$. To this end, suppose there exists an $(\epsilon/3, 1/3)$ -PAC algorithm with at most n per-user interactions. Then using this algorithm, we can conduct a statistical test that recovers the identity of the system (whether it is system (A) or (B)). But this contradicts the information theoretical limit. The test goes as follows. Let $\hat{a} \in \mathcal{A}$ be the optimal action (the one with the most probability assigned in the returned policy) that the algorithm outputs. If $\hat{a} = a^*$, then output (B), otherwise output (A). The success probability of this testing is at least $1/2 \cdot 1/2 + 1/2 \cdot 2/3 \approx 7/12$. This contradicts Le Cam's lower bound. In other words, any $(\epsilon/3, 1/3)$ -PAC algorithm must require at least $\tilde{\Omega}(\alpha^2/\epsilon^2)$ per-user interactions.

4.2. $\min(S, A)$ Lower Bound

We turn our attention to the general case with large S and A . We still consider two systems (A) and (B), with the only difference that now an optimal action $a^*(s) \in \mathcal{A}$ depends on the context $s \in \mathcal{S}$. For all $s \in \mathcal{S}$, let each $a^*(s)$ be decided uniformly at random, independently of any other events. The key idea for getting $\min(S, A)$ lower bound is to check that it is impossible to play better than a random guess on $a^*(\cdot)$ in system (A). Furthermore, if we cannot play $a^*(\cdot)$ more than n times with any user in system (A), then by the similar contradiction argument to Le Cam's fundamental limit, we cannot play the optimal action $a^*(\cdot)$ more than n times for any user either in system (B). Here, following Lemma 4.2, we let $n = o(\alpha^2/\epsilon^2)$.

Algorithm 1 Estimate ν

- 1: **for** $t = 1, 2, \dots, T_0$ **do**
 - 2: An user i_t arrives with a context s_t .
 - 3: Play any action and move to next time step.
 - 4: **end for**
 - 5: Let $\sigma_s^2 := \max(n_s(T_0), O(\log S)) / T_0$ for all $s \in \mathcal{S}$.
 - 6: Compute $b_i = \frac{1}{n_i(T_0)} \sum_{t: i_t=i} \sigma_{s_t}^{-1} e_{s_t}$ for all $i \in [L]$.
 - 7: Get \hat{b} with input $\{\hat{b}_i\}_{i=1}^L$ to high-dimensional robust estimator described in Theorem 3.4.
 - 8: Return $\hat{\nu} := \text{diag}(\{\sigma_s\}_{s \in \mathcal{S}}) \cdot \hat{b}$.
-

Note that $a^*(s)$ can be any action with probability $1/A$ for every context s , therefore a random guess on $a^*(\cdot)$ is correct with probability $1/A$. In the setting where $S \gg A$, within each user we see any context essentially only once, and only with $1/A$ probability the optimal action can be chosen. Hence we need at least $N = O(An)$ interactions to ensure at least n correct actions chosen for any user. On the other hand, if $A \gg S$, then from a random guess, one can show that it is unlikely to guess the optimal actions for more than constant number of contexts within $L = A^{O(1)}$ trials. In other words, for each user we would play the optimal actions for only constant number of contexts, and such contexts are only observed $O(N/S)$ times after N per-user interactions. Thus, we need at least $N = O(Sn)$ interactions to play optimal actions at least n times.

Therefore, we can conclude that we need at least $\Omega(\min(S, A) \cdot \alpha^2/\epsilon^2)$ per-user interactions in order to play $a^*(\cdot)$ more than n times. Furthermore, we show that this is also the case in system (B) due to Le Cam's two point method, which can be translated to impossibility of $(\epsilon/3, 1/3)$ -PAC algorithm. Complete and formal proof of Theorem 4.1 is provided in Appendix C.2.

We comment that our lower bound holds only when L is polynomial in S and A . This is required because, if we allow arbitrarily large number of users, e.g., $L \gg A^S$, then we can divide users into A^S groups and in each group we can evaluate every possible stationary policy with $O(1/\epsilon^2)$ per-user interactions. Theorem 4.1 suggests that such desirable sample-complexity is only possible when more than polynomial number of users are available.

5. Main Algorithm

In Section 3, we showed that the best strategy is simply using either univariate robust estimator if $A \geq S$, or high-dimensional robust estimator $S \geq A$ when the context probability is uniform, i.e., $\nu(s) = 1/S$ for all $s \in \mathcal{S}$. We extend this approach to the case where $\nu(s)$ is a general distribution over contexts. Non-uniform context probability often arises in recommendation systems where a certain set of contexts

(e.g., keywords) are more preferred by users.

Our main idea to exploit non-uniform context probability is quite simple: suppose we can order contexts in probability descending order $s_{[1]}, \dots, s_{[S]}$ such that $\nu(s_{[1]}) \geq \dots \geq \nu(s_{[S]})$. Let $\mathcal{S}_+ := \{s_{[j]} \mid j \in [\min(A, S)]\}$ be a set of top- A frequent contexts. Note that if $A \geq S$, then $\mathcal{S}_+ = \mathcal{S}$. Now for the estimation of rewards under contexts $s \in \mathcal{S}_+$, we use a uni-variate robust estimator (Section 3.1). For the other contexts $s \in \mathcal{S}_- := \mathcal{S}/\mathcal{S}_+$, we estimate μ for the remaining part using a high-dimensional robust estimator (Section 3.2). By combining the two robust estimators in a corrupted multi-user system, we can achieve the improved sample complexity in polynomial time.

One challenge is that we are not given the probability distribution over contexts ν observed by good users. Thus, in the pre-processing step, we need to estimate ν to decide which contexts we consider as top- A contexts, i.e., find $\hat{\nu}$ such that $\|\hat{\nu} - \nu\|_1 \leq \epsilon$. This can be done by the robust high-dimensional estimator with re-scaling of each coordinate after $T_0 = O(\alpha L/\epsilon^2)$ time steps. Let $n_s(T_0) := \sum_{t=1}^{T_0} \mathbb{1}_{s_t=s}$ be the total number of times that a context s is observed, and $\sigma_s^2 := \max(n_s(T_0), O(\log S))/T_0$. Note that σ_s^{-1} scaling serves as a equalizer of context-wise variances. We summarize the procedure to estimate ν in Algorithm 1.

Note that the samples $\{\hat{b}_i\}_{i=1}^L$ are not independent to each other. Nevertheless, robust estimators in Cheng et al. (2019); Lugosi & Mendelson (2021) can still be used when samples are not exactly identical or independent as long as some deterministic conditions hold (see Appendix B.1). We show that with this nice property, we can still recover good enough estimates of $\hat{\nu}$ in Algorithm 1.

We conclude this section with a theoretical guarantee on $(\epsilon, 1/3)$ -optimality of the policy returned by the main Algorithm 2 (Robust MCB) combining all components:

Theorem 5.1. *Let $\alpha < 1/3$ and $L = \Omega(SA \log(SA)/\alpha)$. If we run Algorithm 2 with $T_0 = O(L\alpha/\epsilon^2)$ and $T > L \cdot \alpha/\epsilon^2$, then with probability at least $2/3$, $\hat{\pi}$ satisfies:*

$$V_B^* - V_B^{\hat{\pi}} \leq O\left(K(\mathcal{B}) \cdot \sqrt{\alpha L/T}\right), \quad (5)$$

where $K(\mathcal{B})$ is an instance-dependent quantity given by

$$K(\mathcal{B}) = \sum_{j=1}^A \sqrt{\nu(s_{[j]})} + \sqrt{\sum_{j=A+1}^S \nu(s_{[j]}) \cdot A}. \quad (6)$$

For the uniform context probability, Theorem 5.1 provides a PAC guarantee with $O(\min(S, A) \cdot \alpha/\epsilon^2)$ per-user interactions. The benefit of non-uniform context probabilities is more explicit in the following example: suppose $\nu(s_{[j]}) \propto 1/j^{1+\gamma}$ for $\gamma > 0$, i.e., context probability decays in polynomial rates. Then the sample complexity per-user

Algorithm 2 Robust Multi-task Contextual Bandits

- 1: Pre-Process 1: Get top- A frequent contexts \mathcal{S}_+ from $\hat{\nu}$ returned by Algorithm 1.
 - 2: Pre-Process 2: For each $s \in \mathcal{S}_+$ and $i \in [L]$, assign $a_{i,s} \sim \text{Unif}(\mathcal{A})$. Add i to $\mathcal{I}_{s,a_{i,s}}$.
 - 3: **for** $t = 1, 2, \dots, T$ **do**
 - 4: An user i_t arrives with a context s_t .
 - 5: **if** $s_t \in \mathcal{S}_+$ **then**
 - 6: Play $a_t = a_{i_t, s_t}$ and observe reward r_t .
 - 7: **else**
 - 8: Play $a_t \sim \text{Unif}(\mathcal{A})$ and observe reward r_t .
 - 9: **end if**
 - 10: **end for**
 - 11: Retrieve $\hat{\mu}$ from the procedure in Appendix D.1.
 - 12: Return $\hat{\pi}$ s.t. $\hat{\pi}(s) = \arg \max_{a \in \mathcal{A}} \hat{\mu}(s, a) \forall s \in \mathcal{S}$.
-

guaranteed by Algorithm 2 is $O(\min(S, A)^{1-\gamma} \cdot \alpha/\epsilon^2)$ for $\gamma < 1$. When $\gamma \geq 1$, we have a desired per-user sample-complexity $\tilde{O}(\alpha/\epsilon^2)$.

5.1. Discussion

We remark here a few discussion points that could be of independent and future interest.

Regret Minimization: In the multi-armed bandit case (or $S \leq A$), we can implement a simple regret-minimization algorithm by combining doubling trick and successive elimination techniques. The challenge arises in the other case $S \geq A$. Our main idea for handling this case is to use a high-dimensional robust estimator. However, the guarantee for mean-reward estimators are only given in overall l_2 -distances, which guarantee performance of returned policies *only in expectation*. However, low-regret algorithm should be able to eliminate sub-optimal actions, which needs coordinate-wise accurate estimates of mean-rewards which is not available for high-dimensional robust estimation.

One way to minimize the *overall* regret is to divide user sets into two different groups: one group for exploration, and the other one for exploitation. If the number of users are given $\Omega(S^{3/2}A)$, then we can interact with $O(1/\sqrt{S})$ -fraction of users to find an improved policy, while for others we play a policy obtained in the previous round of epoch. It would be an interesting question to find an algorithm that performs uniformly good for all users in terms of regret.

Tightness on α : Theorem 4.1 suggests $\Omega(\alpha^2)$ lower bound while our upper bound is guaranteed with $O(\alpha)$ samples. This α -gap in the lower and upper bounds results from the fact that our lower bound is built upon Bernoulli reward assumptions, and thus holds for sub-Gaussian type reward distributions, while our upper bound relies only on

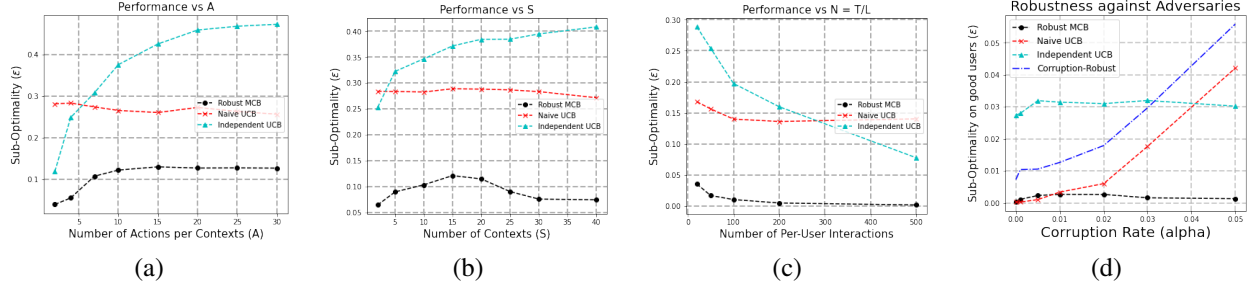


Figure 2: (a-c) Sub-optimality of the returned policy ϵ depending on each parameters $S, A, T/L$. Y-axis illustrates the sub-optimality of the policy, i.e., $\epsilon \propto \sqrt{\min(S, A) \cdot T/L}$. (d) Robustness of algorithms in α -corrupted multi-user systems.

the bounded second-order moment condition for reward distributions. We mention that in the case when the inlier distribution satisfies certain bounded higher-order moment assumptions (satisfied by the average of n i.i.d. Bernoulli variables), near $O(\alpha\sigma)$ -rate robust estimation is information-theoretically possible (Hopkins & Li, 2019). However, to achieve this near-optimal rate, all known efficient algorithms (e.g., (Hopkins & Li, 2018)) require much higher sample complexity, which in our case it means that we need significantly larger number of users L . Tightening a factor of α for sub-Gaussian reward distributions might require a more efficient robust estimators for sub-Gaussian distributions with unknown bounded covariances, and thus we leave it as future work.

Similar Preferences: For personalized recommendations, good users may have similar but not exactly the same preference over all contexts and items (Ghosh et al., 2021b). For instance, suppose that for all $i \neq j$ and $i, j \in \mathcal{I}^*$ with underlying tasks $\mathcal{B}_i, \mathcal{B}_j$ respectively, and we have $\forall \pi \in \Pi, |V_{\mathcal{B}_i}^\pi - V_{\mathcal{B}_j}^\pi| \leq \epsilon_0$. We mention here that the robust estimators we employ here are robust to small perturbations in samples. Therefore, we can first run Algorithm 2 to find a common policy $\bar{\pi}$ such that $|V_{\mathcal{B}_i}^{\bar{\pi}} - V_{\mathcal{B}_j}^{\bar{\pi}}| \leq O(\epsilon_0 + \epsilon)$, after which we can learn for each user separately, e.g., using the algorithm in (Ghosh et al., 2021a).

6. Experiments

We evaluate the proposed algorithm on synthetic data. We set the sub-optimality gap in all contexts approximately 0.3. Our first experiment illustrates the performance of Robust MCB (Algorithm 2) in terms of the sub-optimality ϵ of a returned policy for various numbers of users, contexts, actions, per-user interactions and corruption-rates. Additional experiments on the rate of adversaries and similar preferences are presented in Appendix A.

Sample Complexity. We first check the sample complexity dependence on S and A as stated in Theorem 5.1. We compare our robust multitask contextual bandit (Robust

MCB) algorithm to two primitive algorithms: (i) an UCB algorithm that does not share information across users (Independent UCB) and (ii) the UCB algorithm that ignores user identifiers as if there is no adversary (Naive UCB). The performance of policy is evaluated after a certain number of time steps on a good user. We generate random instances of multitask contextual bandits on various number of contexts, actions, and the number of users. The fraction of adversaries is 20 percent, i.e., $\alpha = 0.2$. The measured sub-optimal gaps ϵ are averaged over 50 independent experiments.

The experimental results are given in Figure 2 (a)-(c). We fix the base parameters as $S = 10, A = 10$, and $T/L = 30$ with $L = O(SA \log(SA))$ users, and measure the accuracy of returned policies for varying parameters. (c) shows how the number of per-user interactions contributes to the performance of returned policies. As shown in the figure, our robust method outperforms two naive approaches. More importantly, we can observe that the increase in S or A does not degrade the performance which confirms $\min(S, A)$ dependency on the sample complexity.

Corruption Robust Algorithms. Although not considered in the framework of multitask learning, it is worth considering corruption-robust algorithms to defend adversaries' plays (Lykouris et al., 2018; Liu et al., 2021). We implement the robust algorithm in (Liu et al., 2021) without communication constraint, and compare the performance as increasing the corruption rate α fixing $S = 2, A = 5, L = 500$ and $T/L = 500$. Since the amount of corruption is linear in T , a small fraction of adversarial users can easily attack the corruption-robust algorithm, while we defend against such attacks by exploiting the side information i_t , the unique identifiers of data source (Figure 2 (d)).

7. Future Work

We believe that our work opens up the prospect of investigating more general problems of multitask learning with adversarial users. In particular, we believe coordinated attacks on Markov decision processes (MDPs) would be an

interesting future problem to explore. Extending our results in the context of function approximation would also be an interesting future direction. In more technical directions, it would be also interesting to study tighter instance-dependent sample-complexity and regret minimization algorithms.

On a deeper level, the type of questions we have asked should be put in the context of responsible AI. Our stylized model provided principled answers to questions such as how many individuals must collude to manipulate a decision? And how to effectively address the possibility of collusion between agents? We show that by hardening a decision algorithm, it is possible to overcome collusion of a much larger portion of the population.

References

- Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.
- Adomavicius, G. and Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749, 2005.
- Audibert, J.-Y., Bubeck, S., et al. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pp. 1–122, 2009.
- Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.
- Brunskill, E. and Li, L. Sample complexity of multi-task reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 122. Citeseer, 2013.
- Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 42–1. JMLR Workshop and Conference Proceedings, 2012.
- Burke, R., O’Mahony, M. P., and Hurley, N. J. Robust collaborative recommendation. In *Recommender systems handbook*, pp. 961–995. Springer, 2015.
- Chawla, R., Sankararaman, A., Ganesh, A., and Shakkottai, S. The gossiping insert-eliminate algorithm for multi-agent bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3471–3481. PMLR, 2020.
- Chen, Y., Xu, H., Caramanis, C., and Sanghavi, S. Matrix completion with column manipulation: Near-optimal sample-robustness-rank tradeoffs. *IEEE Transactions on Information Theory*, 62(1):503–526, 2015.
- Cheng, Y., Diakonikolas, I., and Ge, R. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the thirtieth annual ACM-SIAM symposium on discrete algorithms*, pp. 2755–2771. SIAM, 2019.
- Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.
- Depersin, J. and Lecué, G. Robust sub-gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, 50(1):511–536, 2022.
- Deshmukh, A. A., Dogan, U., and Scott, C. Multi-task learning for contextual bandits. *arXiv preprint arXiv:1705.08618*, 2017.
- Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pp. 999–1008. PMLR, 2017.
- Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.
- Even-Dar, E., Mannor, S., Mansour, Y., and Mahadevan, S. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.
- Gentile, C., Li, S., and Zappella, G. Online clustering of bandits. In *International Conference on Machine Learning*, pp. 757–765, 2014.
- Gentile, C., Li, S., Kar, P., Karatzoglou, A., Zappella, G., and Etrue, E. On context-dependent clustering of bandits. In *International Conference on Machine Learning*, pp. 1253–1262. PMLR, 2017.
- Gerchinovitz, S. and Lattimore, T. Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems*, pp. 1198–1206, 2016.
- Ghosh, A., Sankararaman, A., and Kannan, R. Problem-complexity adaptive model selection for stochastic linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 1396–1404. PMLR, 2021a.
- Ghosh, A., Sankararaman, A., and Ramchandran, K. Collaborative learning and personalization in multi-agent stochastic linear bandits. *arXiv preprint arXiv:2106.08902*, 2021b.

- Gopalan, A., Maillard, O.-A., and Zaki, M. Low-rank bandits with latent mixtures. *arXiv preprint arXiv:1609.01508*, 2016.
- Gupta, A., Koren, T., and Talwar, K. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pp. 1562–1578. PMLR, 2019.
- Hopkins, S. B. and Li, J. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1021–1034, 2018.
- Hopkins, S. B. and Li, J. How hard is robust mean estimation? In *Conference on Learning Theory*, pp. 1649–1682. PMLR, 2019.
- Hu, J., Chen, X., Jin, C., Li, L., and Wang, L. Near-optimal representation learning for linear bandits and linear RL. In *International Conference on Machine Learning*, pp. 4349–4358. PMLR, 2021.
- Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. Reinforcement learning in reward-mixing mdps. *Advances in Neural Information Processing Systems*, 34, 2021a.
- Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. RL for latent mdps: Regret guarantees and a lower bound. *arXiv preprint arXiv:2102.04939*, 2021b.
- Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.
- LeCam, L. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pp. 38–53, 1973.
- Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.
- Liu, J., Li, S., and Li, D. Cooperative stochastic multi-agent multi-armed bandits robust to adversarial corruptions. *arXiv preprint arXiv:2106.04207*, 2021.
- Lu, T., Pál, D., and Pál, M. Contextual multi-armed bandits. In *Proceedings of the Thirteenth international conference on Artificial Intelligence and Statistics*, pp. 485–492. JMLR Workshop and Conference Proceedings, 2010.
- Lugosi, G. and Mendelson, S. Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021.
- Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.
- Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pp. 3242–3245. PMLR, 2021.
- Ma, Y. *Adversarial Attacks in Sequential Decision Making and Control*. PhD thesis, The University of Wisconsin-Madison, 2021.
- Maillard, O.-A. and Mannor, S. Latent bandits. In *International Conference on Machine Learning*, pp. 136–144, 2014.
- Mobasher, B., Burke, R., Bhaumik, R., and Williams, C. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Transactions on Internet Technology (TOIT)*, 7(4):23–es, 2007.
- Olsen, S. Amazon blushes over sex link gaffe. *CNET News*, (December 6), 2002.
- O’Mahony, M., Hurley, N., Kushmerick, N., and Silvestre, G. Collaborative recommendation: A robustness analysis. *ACM Transactions on Internet Technology (TOIT)*, 4(4): 344–377, 2004.
- Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pp. 285–295, 2001.
- Sen, R., Shanmugam, K., Kocaoglu, M., Dimakis, A., and Shakkottai, S. Contextual bandits with latent confounders: An nmf approach. In *Artificial Intelligence and Statistics*, pp. 518–527. PMLR, 2017.
- Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.
- Tropp, J. A. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.
- Van Roy, B. and Yan, X. Manipulation robustness of collaborative filtering. *Management Science*, 56(11):1911–1929, 2010.
- Yang, J., Hu, W., Lee, J. D., and Du, S. S. Impact of representation learning in linear bandits. In *International Conference on Learning Representations*, 2020.
- Zhou, L. A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326*, 2015.
- Zhou, X., Xiong, Y., Chen, N., and Gao, X. Regime switching bandits. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

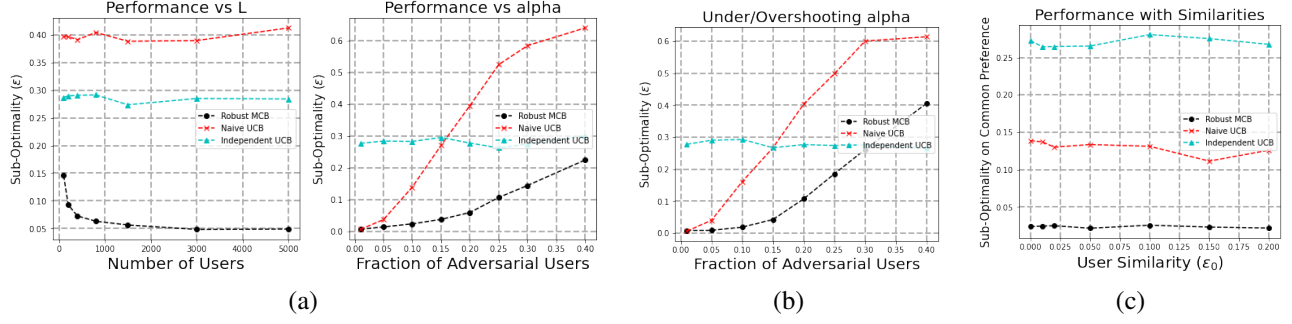


Figure 3: (a) Increase in the number of users only helps until $L = \tilde{O}(SA/\alpha)$. Afterwards, the corruption rules the required number of per-user interactions. (b) Robust MCB is tolerant to small differences in user preferences.

A. Additional Experiments

Effective Corruption-Rate. We recall that we require the number of users to be $\tilde{\Omega}(SA/\alpha)$ for robust estimation rewards. When L is small compared to inverse of α , we may consider $\alpha' = \max(\frac{SA}{L}, \alpha)$ to be an effective fraction of adversarial users. This effect can be seen in Figure 3 (a): we fix $S = 20$, $A = 10$, $T/L = 50$, and see how the policy is improved as L increases. The effective corruption rate decreases when L is below a threshold, and we see the improvement in the returned policy. When L becomes larger the threshold, increase in the number of users no longer improves the quality of the policy. In such case, we can only obtain a better policy by collecting more data from individual user.

Under/Over-Shooting Corruption Rate. We also perform an ablation study when we misspecify the hyper-parameter for the corruption rate α . We set our algorithm to be run with $\alpha = 0.1$, when the real fraction of adversaries varies from 0 to 0.3. When the actual corruption rate is less than the hyper-parameter $\alpha = 0.1$, the algorithm is robust and still outperforms other methods. When the actual corruption rate is larger than 0.1, our algorithm starts to lose the robustness against adversaries. We can conclude that over-shooting is always safer than under-shooting the rate of corruptions.

Similar Preferences. We make users' preferences slightly different from each other by adding small random perturbations of magnitude $O(\epsilon_0)$ to mean-rewards of all contexts and actions for each user. As we gradually increase the amount of perturbations, Robust MCB can tolerate small differences and find the best policy for the common preferences (Figure 3 (c)).

B. Deferred Proofs for Section 3

B.1. Conditions for Robust Estimators

As described in recent work for high-dimensional robust estimations (Diakonikolas et al., 2019; Cheng et al., 2019; Lugosi & Mendelson, 2021), Theorem 3.2 and 3.4 only require a few deterministic conditions to succeed on the concentration of the first and second-order moments. For these robust estimators for d -dimensional samples, we need the following deterministic conditions:

Condition B.1. Suppose N good sample points $X_1, X_2, \dots, X_N \in \mathbb{R}^d$ (not necessarily identically or independently distributed). Let $\mathcal{W}_\epsilon = \left\{ (w_1, w_2, \dots, w_N) \mid \sum_{i=1}^N w_i = 1, 0 \leq w_i \leq \frac{1}{(1-\epsilon)N} \right\}$ for a fixed $\epsilon < 1/3$. Then for all $w \in \mathcal{W}_{3\epsilon}$, there exists μ^* and some absolute constants $\delta_1 = O(\sqrt{\epsilon})$, $\delta_2 = O(1)$, $\delta_3 = O(\sqrt{d/\epsilon})$ such that

$$\left\| \sum_{i=1}^N w_i (X_i - \mu^*) \right\|_2 \leq \delta_1, \quad (7)$$

$$\left\| \sum_{i=1}^N w_i (X_i - \mu^*) (X_i - \mu^*)^\top \right\|_2 \leq \delta_2, \quad (8)$$

$$\forall i \in [N], \quad \|X_i - \mu^*\|_2 \leq \delta_3. \quad (9)$$

Note that i.i.d. assumption on samples is only a sufficient condition for Condition B.1 to hold. When $d = 1$, the above

condition also subsumes the precondition for a uni-variate robust estimator in (Lugosi & Mendelson, 2021). To ensure the correctness of robust estimators we use henceforth, we only need to check these deterministic conditions. When each X_i is an independent random variable with mean μ^* and covariance $Cov(X_i) \preceq I$, following the argument in Appendix A in (Diakonikolas et al., 2019) and Lemma A.18 in (Diakonikolas et al., 2017), we can consider $\{X_i\}_{i=1}^N$ as (0.1ϵ) -corrupted samples with $\{X_i | i \in [N], \|X_i - \mu^*\|_2 \leq \delta_3\}$ as good sample points that satisfy Condition B.1, with probability at least $1 - \delta$ as long as $N = \Omega(d \log(d/\delta)/\epsilon)$ (by Bernstein's inequality for the mean, and the matrix Chernoff bound (Tropp, 2015) for covariance). Henceforth, we only need to check whether the distribution of each sample point X_i has the common mean and bounded second-order moments.

B.2. Proof of Propositions 3.1 and 3.3

For each $a \in \mathcal{A}$, we only need to check Condition B.1 for mean-reward estimates \hat{r}_i from good users $i \in \mathcal{I}^* \cap \mathcal{I}_a$. Note that by construction, $|\mathcal{I}_a| = \Theta(\log(A)/\alpha)$ and $|\mathcal{I}^* \cap \mathcal{I}_a| \geq (1 - 1.1\alpha)|\mathcal{I}_a|$ with probability more than 0.99 for a sufficiently large $L = \Omega(A \log(A)/\alpha)$. Conditioned on the event that each $n_i \geq O(T/L)$, which holds with probability 1 under Assumption 2.1, we have

$$\mathbb{E}[\hat{r}_i] = \mu(a), \quad Var(\hat{r}_i) \leq O(L/T).$$

We can use either median-of-means for $\alpha \geq 1/16$, or Theorem 3.2 in (Lugosi & Mendelson, 2021) for $\alpha < 1/16$. Then robust estimator outputs $\hat{\mu}(a)$ such that $|\hat{\mu}(a) - \mu(a)| \leq O(\sqrt{\alpha L/T})$ with probability at least $1 - 0.1/A$. Taking union bound over all $a \in \mathcal{A}$, we get the lemma.

The proof for Lemma 3.3 follows similarly from equation (2) and Theorem 3.4.

Proof of equation (3): A simple algebra shows that

$$\begin{aligned} V_B^* - V_B^{\hat{\pi}} &= \frac{1}{S} \sum_{s \in \mathcal{S}} \mu(s, \pi^*(s)) - \mu(s, \hat{\pi}(s)) \\ &\leq \frac{1}{S} \sum_{s \in \mathcal{S}} \mu(s, \pi^*(s)) - \hat{\mu}(s, \pi^*(s)) + \hat{\mu}(s, \hat{\pi}(s)) - \mu(s, \hat{\pi}(s)) \\ &\leq \frac{1}{\sqrt{S}} \sqrt{\sum_{s \in \mathcal{S}} (\mu(s, \pi^*(s)) - \hat{\mu}(s, \pi^*(s)))^2} + \frac{1}{\sqrt{S}} \sqrt{\sum_{s \in \mathcal{S}} (\mu(s, \hat{\pi}(s)) - \hat{\mu}(s, \hat{\pi}(s)))^2} \\ &\leq \frac{2}{\sqrt{S}} \|\mu - \hat{\mu}\|_2 \leq O(\sqrt{\alpha L A/T}). \end{aligned}$$

C. Deferred Proofs in Section 4

In this appendix, we provide a full proof of our key negative results.

C.1. Proof of Lemma 4.2

We first note that a simple algebra shows

$$d_{TV}(B(1/2)^{\otimes n}, B(1/2 + \epsilon)^{\otimes n}) \leq \sqrt{\frac{n}{2} d_{KL}(B(1/2) || B(1/2 + \epsilon))} \leq 0.1\alpha,$$

where d_{KL} is a Kullback-Leibler divergence, and we used Pinsker's inequality with $n \leq 0.01\alpha^2/\epsilon^2$. For any sequence $e^{(n)} \in \{0, 1\}^n$, let $\mathbb{P}_{B(1/2)^{\otimes n}}(e^{(n)}) = 2^{-n}$ when sampled from $B(1/2)^{\otimes n}$. Let $e_j^{(n)}$ be a value at the j^{th} position in sequence $e^{(n)}$, and let $\Delta(e^{(n)})$ be defined as $\sum_{j=1}^n \mathbb{1}_{e_j^{(n)}=1} - \mathbb{1}_{e_j^{(n)}=0}$, i.e., the differences in the number of 1's and 0's in $e^{(n)}$. For any $e^{(n)} \in E_{good}^{(n)} := \{e^{(n)} | \Delta(e^{(n)}) \leq 4\sqrt{n \log L}\}$, we can show that

$$\mathbb{P}_{B(1/2)^{\otimes n}}(e^{(n)}) \geq (1 - \alpha) \mathbb{P}_{B(1/2+\epsilon)^{\otimes n}}(e^{(n)}).$$

Let $t = \Delta(e^{(n)}) < 4\sqrt{n \log L}$. The inequality is obvious for $t \leq 0$, and for $t > 0$,

$$\begin{aligned} \frac{\mathbb{P}_{B(1/2+\epsilon) \otimes n}(e^{(n)})}{\mathbb{P}_{B(1/2) \otimes n}(e^{(n)})} &= (1 - 4\epsilon^2)^{(n-t)/2} \cdot (1 + 2\epsilon)^t \\ &\leq (1 - n \cdot \epsilon^2) \cdot (1 + 4t\epsilon) \leq (1 - c\alpha^2) \cdot (1 + 4\sqrt{c}\alpha) \leq \frac{1}{1 - \alpha}. \end{aligned}$$

Let the distribution $E^{(n)}$ over $\{0, 1\}^n$ such that for $e^{(n)} \in E_{good}^{(n)}$,

$$\mathbb{P}_{E^{(n)}}(e^{(n)}) = \frac{1}{Z} \frac{1}{\alpha} \left(\mathbb{P}_{B(1/2) \otimes n}(e^{(n)}) - (1 - \alpha) \mathbb{P}_{B(1/2+\epsilon) \otimes n}(e^{(n)}) \right),$$

where Z is a normalizer to make $E^{(n)}$ a valid distribution, and $\mathbb{P}_{E^{(n)}}(e^{(n)}) = 0$ for $e^{(n)} \notin E_{good}^{(n)}$. Now to check the equation (4), we see that Z is less than:

$$\begin{aligned} Z &= \frac{1}{\alpha} \left(\mathbb{P}_{B(1/2) \otimes n}(e^{(n)} \in E_{good}^{(n)}) - (1 - \alpha) \mathbb{P}_{B(1/2+\epsilon) \otimes n}(e^{(n)} \in E_{good}^{(n)}) \right) \\ &= \frac{1}{\alpha} \left(\mathbb{P}_{B(1/2) \otimes n}(e^{(n)} \notin E_{good}^{(n)}) - \mathbb{P}_{B(1/2+\epsilon) \otimes n}(e^{(n)} \notin E_{good}^{(n)}) + \alpha \cdot \mathbb{P}_{B(1/2+\epsilon) \otimes n}(e^{(n)} \in E_{good}^{(n)}) \right) \\ &\leq \frac{1}{\alpha} \mathbb{P}_{B(1/2) \otimes n}(e^{(n)} \notin E_{good}^{(n)}) + \mathbb{P}_{B(1/2) \otimes n}(e^{(n)} \in E_{good}^{(n)}) \\ &\leq 1 + \frac{1 - \alpha}{\alpha} \cdot \mathbb{P}_{B(1/2) \otimes n}(e^{(n)} \notin E_{good}^{(n)}). \end{aligned}$$

The first inequality comes from the fact that $\mathbb{P}_{B(1/2+\epsilon) \otimes n}(e^{(n)} \in E_{good}^{(n)}) \leq \mathbb{P}_{B(1/2) \otimes n}(e^{(n)} \in E_{good}^{(n)})$. Standard Hoeffding's inequality shows that $\mathbb{P}_{B(1/2) \otimes n}(e^{(n)} \notin E_{good}^{(n)}) \leq \exp(-(4\sqrt{n \log L})^2 / (2n)) = 1/L^8$. For the lower bound on Z , we can simply check that

$$\begin{aligned} Z &= \frac{1}{\alpha} \left(\mathbb{P}_{B(1/2) \otimes n}(e^{(n)} \in E_{good}^{(n)}) - (1 - \alpha) \mathbb{P}_{B(1/2+\epsilon) \otimes n}(e^{(n)} \in E_{good}^{(n)}) \right) \\ &\geq \frac{1}{\alpha} \left(\mathbb{P}_{B(1/2) \otimes n}(e^{(n)} \in E_{good}^{(n)}) - (1 - \alpha) \mathbb{P}_{B(1/2) \otimes n}(e^{(n)} \in E_{good}^{(n)}) \right) \\ &= 1 - \mathbb{P}_{B(1/2) \otimes n}(e^{(n)} \notin E_{good}^{(n)}). \end{aligned}$$

Similarly, we can also show that $Z \geq 1 - \frac{1}{L^8}$. A simple algebra on total-variance distance shows us that (4) $\leq \frac{1}{L^4}$.

C.2. Full Proof of Theorem 4.1

For the construction of hard instances, we assume that $\alpha < 1/6$, $\epsilon \leq c \cdot \alpha / \log^{3/2} L$ for a sufficiently small constant $c \leq 0.01$ and $L \leq \min(S, A)^{O(1)}$. We also assume that a context s is always sampled independently from $Unif(S)$ regardless of incoming users and actions played.

We first show that in system A, it is not possible to play right actions more than $n \leq 0.01\alpha^2 / (\epsilon^2 \log L)$ times for any user i if we interact with user i less than $(\min(S, A) \cdot \alpha^2 / \epsilon^2)$ times. Let \mathcal{E} be the event that there exists at least one user $i \in [L]$ such that we played right actions more than n times with i after N per-user interactions, where N is specified in Lemma C.1. In system A, we can show that the chance of event \mathcal{E} is very small for any algorithm:

Lemma C.1. *In system A, let $\beta > 0$ be any small constant. Let the number of contexts S be sufficiently large so that $\frac{\log S}{\log(\log L)} > 2/\beta$. Suppose that we interact with any user no more than $N = n \cdot \min(S, A)^{1-\beta}$ times. Then, no algorithm can trigger the event \mathcal{E} with probability more than $1/L^2$.*

Proof. Note that in system A, observed reward sequences are independent of the choice of actions at every step. Consequently, any choice of actions by the agent is independent of correct actions $a^*(s)$. Suppose that an algorithm interacted with all users N times, and let $\{s_{i,t}, a_{i,t}\}_{t=1}^N$ be a length N sequence of contexts and actions when interacting with user i . Whatever action choices $\{a_{i,t}\}_{t=1}^N$ made by the agent is statistically independent of $a^*(s)$. Hence we can equivalently think that $a^*(s)$ is a random guess of chosen actions for a context s .

Now we can change the game to guessing more than n chosen actions $a_{i,t}$ by an user i using the completely random guess $a^*(s)$ for all $s \in \mathcal{S}$. Since $a^*(s)$ is completely independent of interaction histories, without loss of generality, we can assume that $s_{i,t}, a_{i,t}$ are fixed after algorithm interacts N times with every user. If there is at least one user i such that $\sum_{t=1}^N \mathbb{1}_{a_{i,t}=a^*(s_{i,t})} \geq n$, then we win the game, *i.e.*, the event \mathcal{E} is triggered.

Let us fix i for now and define a few variables:

$$X_s := \sum_{t:s_{i,t}=s} \mathbb{1}_{a_{i,t}=a^*(s)},$$

$$N_s := \sum_{t:s_{i,t}=s} 1, \quad N_{s,a} := \sum_{t:s_{i,t}=s} \mathbb{1}_{a_{i,t}=a},$$

for all $s \in \mathcal{S}, a \in \mathcal{A}$. Note that X_s is a random variable decided by $a^*(s)$, with $\mathbb{E}[X_s] = N_s/A$ and $\text{Var}(X_s) \leq N_s^2/A$, and $X_s \leq N_s$ almost surely. Furthermore, X_s and $X_{s'}$ are independent if $s \neq s'$. Thus, now we only need to check the probability of event $\sum_{s \in \mathcal{S}} X_s \geq n$. Let $t = n - N/A$. Since $\{X_s\}_{s \in \mathcal{S}}$ are independent random variables, we can apply Bernstein's inequality to obtain:

$$\begin{aligned} \mathbb{P}\left(\sum_{s \in \mathcal{S}} X_s \geq n\right) &\leq \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{s \in \mathcal{S}} \text{Var}(X_s) + \frac{1}{3}(\max_s N_s)t}\right) \\ &\leq \exp\left(-\frac{\frac{1}{2}t^2}{\frac{1}{A} \sum_{s \in \mathcal{S}} N_s^2 + \frac{1}{3}(\max_s N_s)t}\right). \end{aligned} \quad (10)$$

One thing we note here is that since we assumed uniformly sampled context every step, from Bernstein's inequality, we can bound the maximum of N_s as the following:

Lemma C.2. $N_{\max} := \max_s N_s \leq 2N/S + O(\log(L/\delta))$ with probability at least $1 - \delta/L^4$.

This is an application of basic Bernstein's inequality for sums of Bernoulli random variable with parameter $1/S$. Then, using this and $\sum_{s \in \mathcal{S}} N_s = N$, a simple algebra can show that

$$\sum_{s \in \mathcal{S}} N_s^2 \leq N \cdot N_{\max}.$$

Under such event, we can bound (10) further such that

$$\mathbb{P}\left(\sum_{s \in \mathcal{S}} X_s \geq n\right) \leq \exp\left(-\frac{\frac{1}{2}t^2}{N_{\max}(\frac{N}{A} + \frac{1}{3}t)}\right) \leq \exp\left(-\frac{n}{N_{\max}}\right),$$

where we used $n > 10N/A$ for sufficiently large A . Now if $N/S > \log(L/\delta)$, then it is less than $\exp(-nS/N) \leq \exp(-c_1 \cdot S^\beta)$ for some constant $c_1 > 0$. Otherwise, it is less than $\exp(-c_1 \cdot n/\log L)$. In either case, this probability is small enough so that for all users $i \in [L]$, we can take a union bound and show that with no user we have played right actions more than n times if:

$$c \cdot S^\beta \gg \log L, \text{ and } c \cdot n \gg \log^2 L,$$

which holds with our setup $\frac{\log S}{\log(\log L)} > 2/\beta$ and $n \geq 4 \log^2 L$. This concludes the proof for Lemma C.1. \square

Now suppose that there exists an algorithm that can trigger \mathcal{E} in system B with less than N times of per-user interactions with probability more than $2/3$. However, note that system A and B cannot be distinguished before \mathcal{E} is triggered due to Lemma 4.2 and Le Cam (LeCam, 1973) as argued before. However, if it is possible to trigger \mathcal{E} , *i.e.*, to play correct actions at least n times for any user, only using N interactions with probability more than $2/3$, then it is possible to distinguish A and B with probability better than $1/2 \cdot (1 - 1/L^2) + 1/2 \cdot 2/3 \approx 5/6$. Note that until \mathcal{E} is triggered, we can only observe reward sequences of length at most n for correct actions. Therefore, this means that we have a hypothesis testing mechanism by detecting the event \mathcal{E} only with length n reward sequences for correct actions. This contradicts the fundamental limit of two hypothesis testing.

Equivalently, if there exists an $(\epsilon/3, 1/3)$ -PAC algorithm using at most $N/2$ per-user interactions, then we can run this algorithm to obtain an $(\epsilon/3, 1/3)$ -PAC policy, and run this policy for the rest of $N/2$ per-user interactions to trigger \mathcal{E} in system B. This again contradicts the fundamental limit, and thus we conclude Theorem 4.1.

D. Deferred Details in Section 5

D.1. Recovery Procedure for $\hat{\mu}$ in Algorithm 2

We describe the procedure we deferred in Algorithm 2 after T time steps. For every context, let $n_i = |\{t | i_t = i\}|$ be the number of interactions with i . For top- A frequent contexts $s \in \mathcal{S}_+$, let $\hat{b}_{i,s} = \frac{1}{n_i} \sum_{t:i_t=i} r_t \sigma_s^{-1} \mathbb{1}_{s_t=s}$ be the empirical mean of reward times the context probability from the user i and the context s . Then, for every $a \in \mathcal{A}$, call the univariate robust estimator in Theorem 3.2 with input $\{\hat{b}_{i,s}\}_{i \in \mathcal{I}_{s,a}}$, and receive an estimate $\hat{b}(s, a)$. Set $\hat{\mu}(s, a) = \sigma_s^{-1} \cdot \hat{b}(s, a)$ for every $s \in \mathcal{S}_+$ and $a \in \mathcal{A}$.

For those contexts not in \mathcal{S}_+ , let $\mathcal{S}_- = \mathcal{S}/\mathcal{S}_+$ and $\hat{b}_i := \frac{A}{n_i} \sum_{t:i_t=i} r_t \sigma_{s_t}^{-1} \mathbb{1}_{s_t \in \mathcal{S}_-} \cdot e_{(s_t, a_t)}$ where we use σ_s found in Algorithm 1. We call the robust estimator in Theorem 3.4 with input $\{\hat{b}_i\}_{i=1}^L$, and get a returned \hat{b} . Note that we do not use this estimator for frequent contexts, and thus $\hat{b}(s, a) = 0$ for $s \in \mathcal{S}_+$. Set $\hat{\mu}(s, a) = \sigma_s^{-1} \cdot \hat{b}(s, a)$ for every $s \in \mathcal{S}_-$ and $a \in \mathcal{A}$.

D.2. Proof of Theorem 5.1

We first show that the context probability estimated in Algorithm 1 is approximately correct with the following guarantee:

Lemma D.1. *Let $\alpha < 1/3$ and $L = \Omega(S \log(S)/\alpha)$. Then Algorithm 1 with $T_0 = O(L\alpha/\epsilon^2)$ returns an estimator $\hat{\nu}$ that satisfies*

$$\|\hat{\nu} - \nu\|_1 \leq \epsilon,$$

with probability at least 9/10.

Proof. We first note that $\hat{\nu}_i := \frac{1}{n_i(T_0)} \sum_{t:i_t=i} e_{s_t}$ for each $i \in \mathcal{I}^*$ satisfies that

$$\mathbb{E}[\hat{\nu}_i] = \nu, \text{Cov}(\hat{\nu}_i) = \frac{1}{n_i(T_0)} \text{diag}(\nu^{-1}),$$

conditioned on the number of interactions $n_i(T_0)$, where $\text{diag}(\nu^{-1})$ is a diagonalized matrix of vector ν^{-1} such that $\nu^{-1}(s) = \nu(s)^{-1}$ for all $s \in \mathcal{S}$. Note that $n_i(T_0) = O(T_0/L)$. Let a vector $\sqrt{v} \in \mathbb{R}^S$ such that $\sqrt{v}(s) = \sqrt{\nu(s)}$. Then define $\tilde{b}_i := \text{diag}(\sqrt{v})^{-1} \hat{\nu}_i$ which satisfies

$$\mathbb{E}[\tilde{b}_i] = \sqrt{v}, \text{Cov}(\tilde{b}_i) = \frac{1}{n_i(T_0)} I \preceq O\left(\frac{L}{T_0} I\right).$$

Since all \tilde{b}_i are independent from each other (conditioned on the order of user interactions decided by external process Γ), we can find a set of N samples from $\{\tilde{b}_i\}_{i=1}^L$ that satisfies Condition B.1 with $N = (1 - 1.1\alpha)L$, $\mu^* = \sqrt{v}$, $\delta_1 = O(\sqrt{\alpha L/T_0})$, $\delta_2 = O(L/T_0)$, and $\delta_3 = O(\sqrt{SL/(T_0\alpha)})$.

Now we observe that $\hat{b}_i = M \tilde{b}_i$ where $M := \text{diag}(\{\sigma_s^{-1}\}_{s \in \mathcal{S}}) \text{diag}(\sqrt{v})$. Once we show that $\|M\|_2 \leq O(1)$, then the deterministic condition for robust estimation (Condition B.1) holds with μ^* : $\mu^*(s) = \sigma_s^{-1} \nu(s)$. Once we receive a robust estimate of samples $\{\hat{b}_i\}_{i=1}^L$, by the guarantee given by Theorem 3.4, we ensure that

$$\sum_s (\hat{b}(s) - \sigma_s^{-1} \nu(s))^2 \leq O(\sqrt{\alpha L/T_0}),$$

with probability at least 9/10. Therefore using $\hat{\nu}(s) := \sigma_s \hat{b}(s)$ is guaranteed as the following:

$$\sum_s |\sigma_s \hat{b}(s) - \nu(s)| = \sum_s \sigma_s \cdot |\hat{b}(s) - \sigma_s^{-1} \nu(s)|$$

$$\begin{aligned}
 &\leq \sqrt{\sum_s \sigma_s^2} \sqrt{\sum_s (\hat{b}(s) - \sigma_s^{-1} \nu(s))^2} \\
 &\leq \sqrt{\frac{\sum_s n_s(T_0) + O(\log S)}{T_0}} \cdot O\left(\sqrt{\alpha T_0/L}\right) \\
 &= \sqrt{1 + O(S \log S)/T_0} \cdot O\left(\sqrt{\alpha L/T_0}\right).
 \end{aligned}$$

Since $T_0 = O(L\alpha/\epsilon^2) \geq S \log(S)/\epsilon^2$, the right-hand side is less than ϵ .

Finally, we show that for all $s \in \mathcal{S}$,

$$\max(20 \log(S)/T_0, 2v(s)) \geq \sigma_s^2 \geq \max(\log(S)/T_0, v(s)/4). \quad (11)$$

If $v(s) < 4 \log(10S)/T_0$, then this is true by the definition of $\sigma_s^2 = \max(n_s(T_0), 20 \log S)/T_0$. Otherwise, we can show it by a straight-forward application of Bernstein's inequality: let $T'_0 = |\{t | i_t \in \mathcal{I}^*\}|$ be the number of times the system interacts with good users. With probability at least 0.99, $T'_0 \geq (1 - 1.1\alpha)T_0$. Then,

$$\begin{aligned}
 \mathbb{P}\left(\left|\sum_{t: i_t \in \mathcal{I}^*} \mathbb{1}_{s_t=s} - T'_0 \cdot v(s)\right| > t\right) &\leq \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{t: i_t \in \mathcal{I}^*} \mathbb{E}[\mathbb{1}_{s_t=s}^2] + \frac{1}{3}t}\right) \\
 &\leq \exp\left(-\frac{\frac{1}{2}t^2}{T'_0 \cdot v(s) + \frac{1}{3}t}\right).
 \end{aligned}$$

Plugging $t = T'_0 \cdot v(s)/2$ and $v(s) > 4 \log(10S)/T_0$, we can conclude that

$$n_s(T_0) \geq \sum_{t: i_t \in \mathcal{I}^*} \mathbb{1}_{s_t=s} \geq T'_0 \cdot v(s)/2 > T_0 \cdot v(s)/4,$$

with probability at least $1 - 1/(100S^3)$. Taking union bound over all s , we have $\sigma_s^2 \geq v(s)/4$ with probability at least 0.99, yielding $\|M\|_2 \leq 2 = O(1)$. This concludes Lemma D.1. \square

Guarantees for Univariate Estimators. Rest of the proof follows the similar logic. We first show that for each estimator $\hat{\mu}(s, a)$ for $s \in \mathcal{S}_+$ and $a \in \mathcal{A}$, it holds that

$$|\sigma_s^2 \hat{\mu}(s, a) - \nu(s) \mu(s, a)| \leq O\left(\sqrt{\nu(s) \alpha T/L}\right).$$

To show this, we only need to see that $\hat{b}_{i,s}$ satisfies

$$\mathbb{E}[\hat{b}_{i,s}] = \sigma_s^{-1} \nu(s) \mu(s, a_{i,s}), \quad \text{Cov}(\hat{b}_{i,s}) \preceq \frac{\sigma_s^{-2} \nu(s)}{n_i} I.$$

Therefore, after running an univariate robust estimator (Theorem 3.2) with input $\{\hat{b}_{i,s}\}_{i \in \mathcal{I}_{s,a}}$ for each $s \in \mathcal{S}_+, a \in \mathcal{A}$, and with Assumption 2.1 so that $n_i = O(L/T)$, we get

$$|\hat{b}_{s,a} - \sigma_s^{-1} \nu(s) \mu(s, a_{i,s})| \leq O\left(\sigma_s^{-1} \sqrt{\nu(s) \alpha L/T}\right) = O\left(\sqrt{\alpha L/T}\right),$$

with probability at least $1 - 0.1/(SA)$ given $|\mathcal{I}_{s,a}| = \Omega(\log(SA)/\alpha)$. We used the fact from (11).

To compute the error contributed from \mathcal{S}_+ part, we can observe that

$$\begin{aligned}
 \sum_{s \in \mathcal{S}_+} \nu(s) \cdot (\mu(s, \pi^*(s)) - \mu(s, \hat{\pi}(s))) &\leq \sum_{s \in \mathcal{S}_+} \nu(s) \mu(s, \pi^*(s)) - \nu(s) \mu(s, \hat{\pi}(s)) + \sigma_s^2 \hat{\mu}(s, \hat{\pi}(s)) - \sigma_s^2 \hat{\mu}(s, \pi^*(s)) \\
 &\leq \sum_{s \in \mathcal{S}_+} |\nu(s) \mu(s, \pi^*(s)) - \sigma_s^2 \hat{\mu}(s, \pi^*(s))| + |\nu(s) \mu(s, \hat{\pi}(s)) - \sigma_s^2 \hat{\mu}(s, \hat{\pi}(s))|
 \end{aligned}$$

$$\begin{aligned}
 &\leq \left(\sum_{s \in \mathcal{S}_+} \sigma_s \right) \cdot \max_{s \in \mathcal{S}_+} |\hat{b}_{s, \pi^*(s)} - \sigma_s^{-1} \nu(s) \mu(s, \pi^*(s))| \\
 &\quad + \left(\sum_{s \in \mathcal{S}_+} \sigma_s \right) \cdot \max_{s \in \mathcal{S}_+} |\hat{b}_{s, \hat{\pi}(s)} - \sigma_s^{-1} \nu(s) \mu(s, \hat{\pi}(s))| \\
 &\leq \left(\sum_{s \in \mathcal{S}_+} \sigma_s \right) \cdot O\left(\sqrt{\alpha L/T}\right).
 \end{aligned}$$

From (11), we have $\sigma_s^2 \leq O(\log S/T_0) + 2\nu(s)$, where $T_0 = O(\alpha L/\epsilon^2) = O(SA \log(SA)/\epsilon^2)$. Hence, $\sigma_s^2 \leq O(\epsilon^2/SA) + 2\nu(s)$, which gives

$$\sum_{s \in \mathcal{S}_+} \sigma_s \leq \sum_{s \in \mathcal{S}_+} O\left(\epsilon/\sqrt{SA} + \sqrt{\nu(s)}\right) \leq O\left(\epsilon + \sum_{s \in \mathcal{S}_+} \sqrt{\nu(s)}\right),$$

where we used $|\mathcal{S}_+| \leq \min(S, A) \leq \sqrt{SA}$. Note that by definition of $s_{[j]}$ for $j = 1, 2, \dots, S$,

$$\sum_{s \in \mathcal{S}_+} \sqrt{\nu(s)} \leq \sum_{i=1}^{\min(S, A)} \sqrt{\nu(s_{[i]})}.$$

Guarantees for High-Dimensional Estimators. Recall that $\hat{b}_i := \frac{A}{n_i} \sum_{t: i_t=i} r_t \sigma_{s_t}^{-1} \mathbb{1}_{s_t \in \mathcal{S}_-} \cdot e_{(s_t, a_t)}$. Since we only care about $s \in \mathcal{S}_-$, we restrict ourselves to coordinates in \mathcal{S}_- . For a vector v and a index-set \mathcal{I} , we denote $v_{\mathcal{I}}$ as a restriction of a vector to coordinates only in \mathcal{I} . b_{i, \mathcal{S}_-} . Similarly to the uni-variate case, we first see the expectation and covariance of \hat{b}_i :

$$\mathbb{E}[\hat{b}_{i, \mathcal{S}_-}] = \text{diag}(\{\sigma_s^{-1}\}_{s \in \mathcal{S}_-}) \text{diag}(\nu_{\mathcal{S}_-}) \mu_{\mathcal{S}_-}, \quad \text{Cov}(\hat{b}_{i, \mathcal{S}_-}) \preceq \text{diag}(\{\sigma_s^{-2}\}_{s \in \mathcal{S}_-}) \text{diag}(\nu_{\mathcal{S}_-}) \frac{LA}{T} I.$$

From this, the high-dimensional robust estimator in Theorem 3.4 is guaranteed to return $\hat{b}_{\mathcal{S}_-}$ such that

$$\|\hat{b}_{\mathcal{S}_-} - \text{diag}(\{\sigma_s^{-1}\}_{s \in \mathcal{S}_-}) \text{diag}(\nu_{\mathcal{S}_-}) \mu_{\mathcal{S}_-}\|_2 \leq O\left(\sqrt{\alpha AL/T}\right).$$

We used (11) to bound $\|\text{diag}(\{\sigma_s^{-2}\}_{s \in \mathcal{S}_-}) \text{diag}(\nu_{\mathcal{S}_-})\|_2 = O(1)$. From this, we can bound the errors from less frequent contexts \mathcal{S}_- . We first note that

$$\sum_{s \in \mathcal{S}_-} \hat{\nu}(s) \leq \sum_{j=A+1}^S \hat{\nu}(s_{[j]}) \leq \sum_{j=A+1}^S \nu(s_{[j]}) + \epsilon,$$

where the first inequality comes from the fact that \mathcal{S}_- is a collection of contexts that does not belong to top- A highest probabilities in $\hat{\nu}$. Also, by Lemma D.1, we have

$$\sum_{s \in \mathcal{S}_-} \nu(s) \leq \sum_{s \in \mathcal{S}_-} \hat{\nu}(s) + \epsilon \leq \sum_{j=A+1}^S \nu(s_{[j]}) + 2\epsilon. \quad (12)$$

Having this, we can show that

$$\begin{aligned}
 \sum_{s \in \mathcal{S}_-} \nu(s) \cdot (\mu(s, \pi^*(s)) - \mu(s, \hat{\pi}(s))) &\leq \sum_{s \in \mathcal{S}_-} |\nu(s) \mu(s, \pi^*(s)) - \sigma_s^2 \hat{\mu}(s, \pi^*(s))| + |\nu(s) \mu(s, \hat{\pi}(s)) - \sigma_s^2 \hat{\mu}(s, \hat{\pi}(s))| \\
 &\leq \sqrt{\sum_{s \in \mathcal{S}_-} \sigma_s^2} \cdot \sqrt{\sum_{s \in \mathcal{S}_-} (\hat{b}_{s, \pi^*(s)} - \sigma_s^{-1} \nu(s) \mu(s, \pi^*(s)))^2}
 \end{aligned}$$

$$\begin{aligned}
 & + \sqrt{\sum_{s \in \mathcal{S}_-} \sigma_s^2} \cdot \sqrt{\sum_{s \in \mathcal{S}_-} \left(\hat{b}_{s, \hat{\pi}(s)} - \sigma_s^{-1} \nu(s) \mu(s, \hat{\pi}(s)) \right)^2} \\
 & \leq \sqrt{\sum_{s \in \mathcal{S}_-} \sigma_s^2} \cdot O\left(\sqrt{\alpha L A / T}\right).
 \end{aligned}$$

In order to bound $\sum_{s \in \mathcal{S}_-} \sigma_s^2$, we use (11) and see that

$$\sum_{s \in \mathcal{S}_-} \sigma_s^2 \leq 2 \sum_{s \in \mathcal{S}_-} \nu(s) + 20 \frac{S \log(S)}{T_0} = O\left(\sum_{s \in \mathcal{S}_-} \nu(s) + \epsilon^2 / A\right),$$

where we use $T_0 = O(L\alpha/\epsilon^2) \geq SA \log(SA)/\epsilon^2$. Finally, we plug (12), and we have

$$\sum_{s \in \mathcal{S}_-} \nu(s) \cdot (\mu(s, \pi^*(s)) - \mu(s, \hat{\pi}(s))) \leq O\left(\epsilon + \sqrt{\sum_{j=A+1}^S \nu(s_{[j]}) \cdot A}\right) \sqrt{\frac{\alpha L}{T}}.$$

Combining this result with the bound for contexts in \mathcal{S}_+ , we get Theorem 5.1.