Truthful Mechanisms for Linear Bandit Games with Private Contexts

Yiting Hu
Singapore University of Technology and Design
Singapore
yiting_hu@mymail.sutd.edu.sg

Lingjie Duan
Singapore University of Technology and Design
Singapore
lingjie_duan@sutd.edu.sg

ABSTRACT

The contextual bandit problem, where agents arrive sequentially with personal contexts and the system adapts its arm allocation decisions accordingly, has recently garnered increasing attention for enabling more personalized outcomes. However, in many healthcare and recommendation applications, agents have private profiles and may misreport their contexts to gain from the system. For example, in adaptive clinical trials, where hospitals sequentially recruit volunteers to test multiple new treatments and adjust plans based on volunteers' reported profiles such as symptoms and interim data, participants may misreport severe side effects like allergy and nausea to avoid perceived suboptimal treatments. We are the first to study this issue of private context misreporting in a stochastic contextual bandit game between the system and non-repeated agents. We show that traditional low-regret algorithms, such as UCB family algorithms and Thompson sampling, fail to ensure truthful reporting and can result in linear regret in the worst case, while traditional truthful algorithms like explore-then-commit (ETC) and ϵ-greedy algorithm incur sublinear but high regret. We propose a mechanism that uses a linear program to ensure truthfulness while minimizing deviation from Thompson sampling, yielding an $O(\ln T)$ frequentist regret. Our numerical experiments further demonstrate strong performance in multiple contexts and across other distribution families.

KEYWORDS

Contextual linear bandit; private context; truthful mechanism; regret bound

ACM Reference Format:
Yiting Hu and Lingjie Duan. 2025. Truthful Mechanisms for Linear Bandit Games with Private Contexts. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

1 INTRODUCTION

The contextual bandit problems have received increasing attention over the past decade, beginning with Auer's introduction of the concept [1, 2, 5, 7, 10]. In the contextual bandit model, an arbitrary set of observable actions is available at each time step, and the reward for each action is determined by an unknown parameter shared across all actions. The contextual bandit excels in making personalized decisions by using contextual information to select the best possible actions. This allows for more efficient learning and better adaptation to dynamic environments compared to traditional bandit models.

However, these models do not align with scenarios involving private contexts and fail to capture the challenges posed by private information. In the new stochastic bandit problem involving private contexts [8, 13, 20, 25], at each time step, a new agent arrives, reports her private context, and the system selects one of the available K arms, where each arm is associated with a different unknown parameter. The agent then receives a stochastic reward based on the system's chosen action, after which she leaves. This scenario is common in applications like clinical trials and online recommendations, where agents may strategically misreport private contexts to maximize single-round personal rewards. For example, in adaptive clinical trials of phase 2 INSIGHT trial where hospitals test treatments for glioblastoma based on patients' symptoms and medical history, some patients may misreport side effect histories like allergies or anemia to avoid the less-established abemaciclib treatment [12, 24]. On online platforms like Netflix or Amazon, many users prefer recommendations based on popular choices or expert curation [14].

In this new stochastic contextual linear bandit problem with private contexts, previous works assume observable and public contexts and do not consider the agents' strategic behavior to game the system. The conflict between the system's long-term reward and the individual's immediate reward in the multi-armed bandit problem has been studied for the past decade [15, 17, 18, 21, 22, 28, 29]. Kremer et al. [18] initiate the research within a Bayesian exploration framework, introducing a recommendation mechanism for incentivizing exploration with deterministic rewards. Mansour et al. [21] further develop the problem to the stochastic rewards. Sellke and Slivkins [29] first prove that Thompson sampling algorithm can be naturally incentive-compatible (IC) if provided with sufficient initial samples. Then Hu et al. [15] and Simchowitz and Slivkins [28] extend this result to the combinatorial and linear bandit problems. Beyond recommendation mechanisms, Immorlica et al. [17] apply selective disclosure of historical information to encourage exploration. Simchowitz and Slivkins [30] also study this problem in reinforcement learning. These works assume that the system has the full information about agents for the recommendation, and then the problem is to design the IC mechanisms that ensure agents follow the recommendation. In contrast, our system needs agents to report their private contexts, where the system lacks context information, and agents may strategically misreport their private contexts, rendering these methods ineffective for the problem addressed in this paper.

Our main contributions are summarized as follows:

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

- We are the first to study agents' strategic context misreporting to maximize their one-time individual expected rewards in the new contextual bandit problem. We demonstrate that existing algorithms perform poorly under misreporting. Specifically, existing truthful algorithms, such as the greedy and explore-then-commit (ETC) methods, suffer from relatively high regret, while low-regret algorithms like UCB family algorithms and Thompson sampling exhibit a regret of $O(T)$ under strategic misreporting.
- We propose a truthful mechanism based on Thompson sampling algorithm which guarantees that agents have no incentive to misreport their contexts. We prove that our algorithm achieves a frequentist regret upper bound of $O(\ln T)$ in the Bayesian contextual linear bandit setting. Additionally, our experiments show that the mechanism has sublinear regret when applied to multiple contexts and across some other sub-Gaussian distributions.

All the missing proofs are available in the full version of the paper (see https://arxiv.org/pdf/2501.03865.pdf, [16])

1.1 Related work

Stochastic contextual linear bandit algorithms can be categorized into deterministic algorithms, which make deterministic choices, and stochastic algorithms, which maintain a probability distribution among arms for selection. When facing agents' context misreporting, deterministic algorithms in which the choice depends on context cannot ensure truthful reporting in exploration because the resulting arm choices are predictable. Therefore, deterministic low-regret algorithms like the UCB family [1, 10] suffer from linear regret in the worst-case scenario as shown in Section 3 in this paper. Another deterministic algorithm, the Explore-Then-Commit (ETC) algorithm, does not rely on agents' context but is inefficient, incurring a relatively high regret of order $O(T^{2/3})$ [19]. For stochastic algorithms, the ϵ^t-greedy algorithm is truthful as its exploration probability is independent of the context, but it also incurs a regret order of $O(T^{2/3})$ [31].

Another stochastic low-regret algorithm is Thompson sampling, which was first adapted by Agrawal and Goyal in [5] for the contextual linear bandits problem. Abeille and Lazaric in [2] further improve the frequentist regret of linear Thompson sampling. Thompson sampling is also widely applied to Bayesian bandit problems, as it naturally leverages posterior distributions. Russo and Van Roy [26, 27] provide the Bayesian regret upper bound for Thompson sampling, with frequentist regret serving as an upper bound on Bayesian regret. However, we will show in Section 3 that Thompson sampling is not truthful and still suffers from linear regret under misreporting behavior.

Regarding context misreporting behavior, Buening et al. examine this phenomenon in a different problem in [9]. Their work considers arms as repeated strategic entities that manipulate rewards to increase their chances of being chosen. While they also address context misreporting behavior, their focus fundamentally differs from ours, and their approach is inapplicable to our problem, as our agents are non-repeated and myopic.

2 PROBLEM FORMULATION

We consider the Bayesian contextual linear bandit model in [8]. There are K arms in the set $[K] = \{1, \dots, K\}$, each associated with an unknown, fixed d-dimensional hidden parameter $\theta_k \in \mathbb{R}^d$. These parameters $\{\theta_k\}_{k\in[K]}$ are unknown to both the system and the agents but are drawn from a known prior distribution $\mathcal{P}_k : \mathbb{R}^d \to \mathbb{R}$. The prior distributions $\mathcal{P}_k$ for any $k \in [K]$ are common knowledge for both the system and the agents. We define the prior mean of each $\mathcal{P}_k$ as $\mu_k \in \mathbb{R}^d$ and the covariance matrix as $V_k \in \mathbb{S}^d_+$.

At each time $t \in \{1, \dots, T\}$, a new agent arrives with a private context $x_t \in \mathcal{X}$, where $\mathcal{X} = \{\chi_1, \dots, \chi_N\} \subsetneq \mathbb{R}^d$ is a set of finite N contexts. We assume that each x_t takes the value $\chi_n \in \mathcal{X}$ with a known probability $\beta_n > 0$, where $\sum_{n=1}^N \beta_n = 1$ [1]. The system first provides its history $\mathcal{F}_t$, which includes previously observed contexts, actions, and rewards, as well as the arm selection policy $\pi(x, \mathcal{F}_t)$ for all contexts $x \in \mathcal{X}$ which defines a probability distribution over the K arms, to the agent. After receiving this information, the agent reports the context $x'_t \in \mathcal{X}$ to the system. Based on the reported context x'_t, the system selects an arm $a_t \in [K]$ and only observes the corresponding reward $r_t = x_t^\top \theta_{a_t} + \eta_t$, where η_t is a zero-mean random variable. The system then updates the posterior estimation of the chosen arm a_t to $\hat{\theta}^{t+1}_{a_t}$, and the posterior distribution to $\mathcal{P}^{t+1}_{a_t} = \mathcal{P}_{a_t}(\cdot|\mathcal{F}_{t+1})$.

We begin by considering Gaussian priors and Gaussian noise to provide a clearer illustration of the problem and to facilitate the analysis of frequentist regret, as in [5, 6]. Specifically, we assume $\mathcal{P}_k \sim \mathcal{N}(\mu_k, V_k)$ and $\eta_t \sim \mathcal{N}(0, 1)$. This assumption allows for closed-form updates of $\hat{\theta}^t_k$ and $\mathcal{P}^t_{a_t}$, making the analysis more tractable. Still, our mechanism is applicable to any family of prior and noise distributions. In Section 7, we use simulations to demonstrate that the mechanism design in Section 4 achieves good regret performance under other sub-Gaussian distributions. Under Gaussian priors and Gaussian noise, the posterior distribution is $\mathcal{P}^t_k(\cdot) \sim \mathcal{N}(\hat{\theta}^t_k, \hat{V}^t_k)$, where $\hat{\theta}^t_k$ and $\hat{V}^t_k$ are updated as follows:

$$\hat{V}^t_k = \left(V_k^{-1} + \sum_{\tau\in\mathcal{T}^t_k} x_\tau x_\tau^\top\right)^{-1}, \hat{\theta}^t_k = \hat{V}^t_k\left(V_k^{-1}\mu_k + \sum_{\tau\in\mathcal{T}^t_k} x_\tau r_\tau\right), \quad (1)$$

where $\mathcal{T}^t_k$ denotes the set of time steps when the system chooses arm k before time t.

We now formulate the objectives of both the agents and the system. Given the arm choice policy $\pi(x, \mathcal{F}_t)$ for any $x \in \mathcal{X}$ provided by the system, the agent arriving with context x_t chooses to report the context x'_t that maximizes her expected reward, expressed as:

$$x'_t = \arg\max_{x\in\mathcal{X}} {x_t}^\top \Theta^t \pi(x, \mathcal{F}_t), \quad (2)$$

where $\Theta^t = [\hat{\theta}^t_1, \dots, \hat{\theta}^t_K]$ represents the matrix of posterior estimates at time t. Based on Eq. (2), we then present the definition of the truthful mechanism as follows:

Definition 2.1. A mechanism is considered truthful in our Bayesian contextual linear bandit problem if no agent can increase her expected reward by misreporting her true context at any time step.

[1] Our mechanism also works when the arrival probability β_n is unknown. We can initialize the mechanism with a uniform discrete context distribution and adjust β_n as we learn and update it throughout the process.

Formally, for any history $\mathcal{F}_t$ and any pair of distinct contexts x_t and x'_t in $\mathcal{X}$, the following condition holds at each time step $t \in [T]$:

$$x_t^\top \Theta^t \pi(x_t, \mathcal{F}_t) \geq x_t{}^\top \Theta^t \pi(x'_t, \mathcal{F}_t). \tag{3}$$

When a context $x \in \mathcal{X}$ has an incentive to misreport as another context $x' \in \mathcal{X}$, we say that contexts x and x' have conflict.

Conversely, the system's objective is to maximize the cumulative reward, or equivalently, to minimize the expected total regret by choosing a_t for each time step t in the truthful mechanism $\pi(\cdot)$. Let a_t^* denote the optimal arm for the agent arriving at time t. The expected total regret is:

$$\mathbb{E}[\mathcal{R}(T)] = \mathbb{E}\left[\sum_{t=1}^{T}\left(x_t^\top \theta_{a_t^*} - x_t^\top \theta_{a_t}\right)\right]. \tag{4}$$

Building on the truthful mechanism defined above, we will next analyze the behavior of existing bandit algorithms under misreporting.

3 PERFORMANCE ANALYSIS OF EXISTING ALGORITHMS UNDER MISREPORTING

In this section, we present comprehensive studies on the performances of existing deterministic and stochastic algorithms under agents' possible misreporting.

3.1 Deterministic Algorithms

In deterministic algorithms, the algorithm selects one of the K arms based on the history with probability 1 at each time step. A well-known class of deterministic algorithms for the contextual linear bandit problem is the UCB family, including LinUCB [10] and OFUL [1], which select arms at each time step based on upper confidence bounds. However, UCB family algorithms are vulnerable to misreporting because their allocation is predictable, allowing agents to easily manipulate their reported context to favor the currently optimal arm, which can lead to linear regret in the worst-case scenario.

In contrast, the deterministic Explore-Then-Commit (ETC) algorithm, which first operates in a round-robin exploration phase before switching to a purely greedy strategy, is truthful because its decisions are independent of personal contexts, making agents' context reporting irrelevant to the algorithm's choice. However, the ETC algorithm incurs a relatively high regret of $O(T^{2/3})$ [19].

3.2 Stochastic Algorithms

In stochastic algorithms, the algorithm maintains a probability distribution over the arms at each time step and selects an arm according to this distribution. The ϵ^t-greedy algorithm, which selects the greedy arm with probability ϵ^t and chooses an arm uniformly at random with probability $1-\epsilon^t$, is truthful since ϵ^t can be set so that the selection probability is independent of the contexts. However, it also suffers from a relatively high regret of $O(T^{2/3})$ [31]. Next, we consider Thompson sampling algorithm [5].

In Thompson sampling, given the reported context x'_t at each time step, the system draws a sample $\tilde{\theta}_k^t$ from the posterior distribution of $x'^\top_t \theta_k$, denoted as $\mathcal{P}_k^t(\cdot|x'_t) : \mathbb{R} \to \mathbb{R}$, and then selects arm $\bar{a}_t = \arg\max_k \tilde{\theta}_k^t$. Note that $\mathcal{P}_k^t(\cdot)$ represents the posterior distribution of θ_k at time t, whereas $\mathcal{P}_k^t(\cdot|x'_t)$ is the posterior distribution of $x'^\top_t \theta_k$. The process of first sampling θ_k from $\mathcal{P}_k^t$ and then multiplying it by x'_t yields the same result as directly sampling from $\mathcal{P}_k^t(\cdot|x'_t)$ when assuming Gaussian prior and noise. Therefore, the distribution of arm selection $\mathbf{p}^t(x'_t) = (p_1^t(x'_t), \ldots, p_K^t(x'_t))$ in Thompson sampling, which is also the policy $\pi(x'_t, \mathcal{F}_t)$ at time t is:

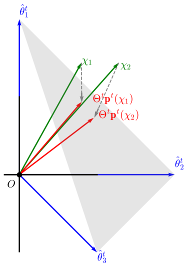

Figure 1: Geometric illustration of context χ_2's incentive to misreport as χ_1.

$$p_k^t(x'_t) = \int \cdots \int_{\tilde{\theta}_k^t \geq \tilde{\theta}_j^t,\, j \in [K],\, j \neq k} d\mathcal{P}_1^t(\tilde{\theta}_1^t|x'_t) \cdots d\mathcal{P}_K(\tilde{\theta}_K^t|x'_t). \tag{5}$$

We can demonstrate that Thompson sampling is not truthful through a simple example in Fig. 1. At time t, given the posterior estimation $\hat{\theta}_k^t$ for all $k \in [3]$ and the expected arm choice policy $\pi(x'_t, \mathcal{F}_t) = (p_1^t(x'_t), p_2^t(x'_t), p_3^t(x'_t))$ in (5), the resulting expected parameter $\Theta^t \mathbf{p}^t(x'_t)$ lies within the convex hull $\mathrm{conv}(\hat{\theta}_1^t, \hat{\theta}_2^t, \hat{\theta}_3^t)$. Therefore, Thompson sampling can be seen as a mapping from any context $x'_t \in \{\chi_1, \chi_2\}$ to $\Theta^t \mathbf{p}^t(x'_t)$ within the convex hull. In the Thompson sampling mapping of $\Theta^t \mathbf{p}^t(\chi_1)$ and $\Theta^t \mathbf{p}^t(\chi_2)$ in Fig. 1, context χ_2 yields a higher inner product with $\Theta^t \mathbf{p}^t(\chi_1)$ than with $\Theta^t \mathbf{p}^t(\chi_2)$. Consequently, if χ_2 arrives at time t, it will misreport as χ_1.

We further prove the regret of Thompson sampling under the context misreporting.

Lemma 3.1. *Thompson sampling algorithm cannot ensure truthful context reporting and results in linear regret $O(T)$ in the worst case.*

Sketch of Proof. We prove the lemma by constructing an example in which, given a specific prior $\mathcal{P}_1 \times \cdots \times \mathcal{P}_K$, one of the contexts has an incentive to misreport. Then, under a certain context arrival distribution $\{\beta_n\}_{n \in [N]}$, we find a positive probability that the algorithm fails to identify the true optimal arm for this context throughout the process and ultimately converge to a suboptimal arm for this context. □

Given that existing deterministic and stochastic algorithms fail due to either untruthfulness or inefficiency, there is a need to design more effective, truthful mechanisms.

4 TRUTHFUL THOMPSON SAMPLING MECHANISM

In this section, we introduce our truthful-Thompson sampling (truthful-TS) mechanism for contextual linear bandits to ensure truthful reporting under Thompson sampling. Our objective is to determine a new probability distribution $\mathbf{q}^t(\chi_n)$ across the K arms at each time t for any $n \in [N]$ that guarantees truthfulness. We derive $\{\mathbf{q}^t(\chi_n)\}_{n\in[N]}$ from the Thompson sampling probabilities $\{\mathbf{p}^t(\chi_n)\}_{n\in[N]}$ in Eq. (5), aiming to keep $\{\mathbf{q}^t(\chi_n)\}_{n\in[N]}$ as close as possible to $\{\mathbf{p}^t(\chi_n)\}_{n\in[N]}$. To achieve this, we formulate a linear optimization problem (LP) at each time t, given by:

$$\begin{aligned}
\text{minimize } & \max_{n\in[N]} (||\mathbf{p}^t(\chi_n) - \mathbf{q}^t(\chi_n)||_\infty) \\
\text{s.t. } & \chi_i^\top \Theta^t (\mathbf{q}^t(\chi_i) - \mathbf{q}^t(\chi_j)) \geq 0, \quad \forall i \neq j, i, j \in [N] \\
& \sum_{n\in[N]} \beta_n \mathbf{q}^t(\chi_n) = \sum_{n\in[N]} \beta_n \mathbf{p}^t(\chi_n), \\
& q_1^t(\chi_n) + \cdots + q_K^t(\chi_n) = 1, \quad n \in [N] \\
& q_k^t(\chi_n) \geq 0 \quad k \in [K], \quad n \in [N].
\end{aligned} \tag{6}$$

In (6), the objective is to minimize the maximum difference between any $p_k^t(\chi_n)$ and $q_k^t(\chi_n)$ across all possible contexts, keeping the new distribution as aligned as possible with the Thompson sampling probabilities. The first constraint ensures that the agent with private context χ_i cannot obtain a higher expected reward $\chi_i^\top \Theta^t \mathbf{q}^t(\chi_j)$ by misreporting as χ_j than the reward $\chi_i^\top \Theta^t \mathbf{q}^t(\chi_i)$ by truthfully reporting. The second constraint ensures that the weighted average of choosing arm k across all contexts, based on the arrival probability β_n of context χ_n, remains the same as the weighted average of Thompson sampling probability $p_k^t(\cdot)$. The third and fourth constraints ensure that the solution $\mathbf{q}^t(x)$ for each $x \in \mathcal{X}$ forms a valid probability distribution.

Based on (6), we present our truthful-TS mechanism in Mechanism 1. To demonstrate the feasibility of our mechanism, we first need to show that the LP in problem (6) has a feasible and convergent solution, where the probability of selecting the optimal arm converges to 1. We can easily construct such a feasible solution. Define the conflict clustering $\mathbf{C} = \{C_1, \ldots, C_j\}$, where each $C \in \mathbf{C}$ is a subset of contexts and each context $x \in \mathcal{X}$ belongs to exactly one cluster $C \in \mathbf{C}$. Within each cluster, every context has a conflict with at least one other context in the same cluster and has no conflicts with contexts in any other clusters. Let C be a mapping from any context $x \in \mathcal{X}$ to its respective cluster, such that $C : \mathcal{X} \to \mathbf{C}$. Then, we can construct a feasible solution for (6) as follows:

$$\mathbf{q}^t(x) = \sum_{\chi_i \in C(x)} \frac{\beta_i \mathbf{p}^t(\chi_i)}{\sum_{\chi_i \in C(x)} \beta_i}, \forall x \in \mathcal{X}. \tag{7}$$

However, a better solution for $\mathbf{q}^t(x)$ can be obtained by solving (6), under the condition in Lemma 4.1.

Lemma 4.1. *Problem (6) must have a feasible and convergent solution as in (7). Additionally, as long as $\chi_n^\top(\Theta^t(\cdot, 1 : K-1) - \hat{\theta}_K^t \mathbf{1}^\top)$ are linearly independent across all $n \in [N]$, i.e.,*

$$\sum_{n\in[N]} \lambda_n \chi_n^\top (\Theta^t(\cdot, 1 : K-1) - \hat{\theta}_K^t \mathbf{1}^\top) \neq 0, \ \forall (\lambda_1, \ldots, \lambda_N) \neq \mathbf{0}, \tag{8}$$

the feasible solution space of problem (6) has a non-zero measure around $\{\mathbf{q}^t(x)\}_{x\in\mathcal{X}}$ in (7) with infinitely many possible solutions.

Mechanism 1: Truthful-Thompson sampling mechanism

```
for $t = 1$ to $T$ do
    Calculate the arm choice distribution $\mathbf{p}^t(x)$ for each $x \in \mathcal{X}$ in Thompson sampling algorithm by Eq. (5).
    Solve $\{\mathbf{q}^t(x)\}_{x\in\mathcal{X}}$ in (6), using $\mathcal{X}$, $\{\mathbf{p}^t(x)\}_{x\in\mathcal{X}}$, $\{\hat{\theta}_k^t\}_{k\in[K]}$ and $\{\hat{V}_k^t\}_{k\in[K]}$ as inputs.
    Provide the solution $\{\mathbf{q}^t(x)\}_{x\in\mathcal{X}}$, the posterior estimates $\{\hat{\theta}_k^t\}_{k\in[K]}$ and $\{\hat{V}_k^t\}_{k\in[K]}$ to the agent arriving at time $t$.
    Observe the context $x_t'$ reported by the agent.
    Choose arm $a_t$ according to the probability distribution $\mathbf{q}^t(x_t')$.
    Observe the reward $r_t$, then update $\hat{V}_{a_t}^t$ and $\hat{\theta}_{a_t}^t$ based on Eq. (1).
```

Sketch of proof. It is straightforward to observe that setting $\mathbf{q}^t(x)$ as the weighted average of $\mathbf{p}^t(x)$ within each subset of conflicting contexts, with weights proportional to β_n, yields a feasible solution. To prove the second part of the lemma, we start by noting that the feasible solution of (6) must take the form $\mathbf{q}^t(\chi_n) = \sum_{i\in[N]} \beta_i \mathbf{p}^t(\chi_i) + \xi_n$, where $\sum_{n\in[N]} \beta_n \xi_n = 0$. By substituting $\mathbf{q}^t(\chi_n) = \sum_{i\in[N]} \beta_i \mathbf{p}^t(\chi_i) + \xi_n$ into (6) and redefining the variables in terms of ξ_n for all $n \in [N]$, we then reformulate problem (6) into an equivalent form as follows:

$$\begin{aligned}
\min \max_{n\in[N]} & \left\| \sum_{i\in[N]} \beta_i \mathbf{p}^t(\chi_i) - \mathbf{p}^t(\chi_n) + \xi_n \right\|_\infty \\
\text{s.t. } & \chi_i^\top \Theta^t (\xi_i - \xi_j) \geq 0, \quad \forall i \neq j, \ i, j \in [N] \\
& \xi_{n,1} + \cdots + \xi_{n,K} = 0, \quad \forall n \in [N] \\
& \beta_1 \xi_{1,k} + \cdots + \beta_N \xi_{N,k} = 0, \quad \forall k \in [K] \\
& 0 \leq \sum_i \beta_i p_k^t(\chi_i) + \xi_{n,k} \leq 1, \quad \forall n \in [N], k \in [K].
\end{aligned} \tag{9}$$

We can reformulate the first three constraints of (9) as a convex cone given by

$$\begin{pmatrix} \chi_1^\top (\Theta^t(\cdot, 1 : K-1) - \hat{\theta}_K^t \mathbf{1}^\top) \otimes A_1^\top \\ \vdots \\ \chi_N^\top (\Theta^t(\cdot, 1 : K-1) - \hat{\theta}_K^t \mathbf{1}^\top) \otimes A_N^\top \end{pmatrix} \mathrm{vec}(\mathcal{E}^\top) \geq 0,$$

where $\mathcal{E}$ represents the first $K-1$ rows and first $N-1$ columns of matrix $(\xi_1, \ldots, \xi_N)$, and A_n is a constant matrix constructed for each $n \in [N]$. We show that this convex cone has a non-zero measure when all vectors $\chi_n^\top(\Theta^t(\cdot, 1 : K-1) - \hat{\theta}_K^t \mathbf{1}^\top)$ for $n \in [N]$ are linearly independent, based on the properties of the constructed matrix A_n. Furthermore, since the distribution $\mathbf{p}^t(x)$ for any $x \in \mathcal{X}$ lies within the interior of the simplex Δ^K, representing all probability distributions over K arms, we can construct a feasible solution space with a non-zero measure. □

When Eq. (8) is satisfied, we can modify the first constraint of problem (6) to $\chi_i^\top \Theta^t (\mathbf{q}^t(\chi_i) - \mathbf{q}^t(\chi_j)) \geq \epsilon$, where ϵ is a sufficiently small positive value to ensure that truthful reporting becomes a strictly dominant strategy for agents. Furthermore, when the feasible space has a non-zero measure, it can yield an improved optimal value for problem (6) compared to $\{\mathbf{q}^t(x)\}_{x\in\mathcal{X}}$ in (7).

However, solving the linear program in (6) incurs a complexity higher than $O((KN)^{2+1/6})$ [11], whereas identifying (7) has a lower complexity of $O(KN^2)$, which arises from the process of identifying conflict clusters. For larger K and N, we can improve efficiency of Mechanism 1 by using $\{\mathbf{q}^t(x)\}_{x\in\mathcal{X}}$ from (7) as a substitute for solving (6).

5 REGRET ANALYSIS UNDER THE SAME OPTIMAL ARM FOR TWO CONTEXTS

In this section, we analyze the regret performance of our truthful mechanism in the simple but fundamental case of two contexts. The two-context scenario is common in real-world settings. For example, online platforms often categorize users as either Mac or Windows users to tailor sales strategies [4]. Similarly, in clinical trials, hospitals categorize patients as either treatment-naive or treatment-experienced when conducting studies [23]. For agents with two possible contexts, the scenarios are limited to either having the same or different optimal arms to misreport the other context. As N increases, the misreporting patterns become exponentially more intricate among agents, significantly complicating the regret analysis. In Section 7, we still run simulations for multiple contexts to show similar results as presented in this section.

We focus on problem-dependent frequentist regret because misreporting behavior is influenced by the specific realization of the prior, requiring separate analysis for different cases. Specifically, we divide the realizations into two cases: (1) when the two contexts share the same optimal arm, and (2) when the two contexts have different optimal arms. The differing misreporting incentives for each case lead to major differences in regret analysis. Still, Bayesian regret can be obtained by taking the expectation over our frequentist regret, yielding the same order. We begin by analyzing the regret in the scenario where the two contexts, χ_1 and χ_2, share the same optimal arm in this section. Addressing the other scenario requires new extra techniques (see Section 6).

When two contexts share the same optimal arm, the probability of selecting the optimal arm will eventually converge to 1 for both contexts. However, one context must converge faster than the other. Consequently, the context with a slower convergence rate may have an incentive to misreport for most of the time steps during the process. Inspired by this, we consider the two contexts collectively and derive an upper bound on the total number of suboptimal pulls for both contexts.

Theorem 5.1. *For the realization of prior, $\{\theta_k\}_{k\in[K]}$, such that the two contexts χ_1 and χ_2 share the same optimal arm α, the frequentist regret of our truthful-TS mechanism in Mechanism 1 is $O(\ln T)$ to be upper bounded by*

$$\sum_{j\neq\alpha}\left(\sum_{n=1}^{2}\frac{18}{\Delta_{n,j}^2}\ln\frac{T\Delta_{n,j}^2}{36}+C_{n,j}\right)\max_{n=1,2}\Delta_{n,j},$$

where $\Delta_{n,j}=\chi_n^\top\theta_\alpha-\chi_n^\top\theta_j$ is the reward gap between the optimal arm α and arm $j\neq\alpha$ for context χ_n. $C_{n,j}$ is a constant for $n\in\{1,2\}$ and $j\in[K]$.

Proof. Let α denote the optimal arm for both contexts. Recall that a_t represents the arm chosen under our truthful-TS mechanism in Mechanism 1, and $\bar{a}_t$ represents the arm chosen under the standard TS algorithm. Since the two contexts share the same optimal arm, we can decompose the regret in equation (4) as follows:

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)] =&\mathbb{E}\left[\sum_{t=1}^{T}\left(x_t^\top\theta_{a_t^*}-x_t^\top\theta_{a_t}\right)\right]\\ &=\sum_{t=1}^{T}\sum_{n=1}^{2}\sum_{j\neq\alpha}\mathbb{E}[\mathbf{1}(x_t=\chi_n,a_t=j)]\Delta_{n,j}\\ &\le\sum_{t=1}^{T}\sum_{n=1}^{2}\sum_{j\neq\alpha}\mathbb{E}[\mathbf{1}(x_t=\chi_n,a_t=j)]\max_{n=1,2}\Delta_{n,j}\\ &=\sum_{t=1}^{T}\sum_{n=1}^{2}\sum_{j\neq\alpha}\mathbb{E}[\beta_n\mathbb{E}[\mathbf{1}(a_t=j)|x_t=\chi_n,\mathcal{F}_t]]\max_{n=1,2}\Delta_{n,j}.\end{aligned}\tag{10}$$

The expectation in the first equality is taken over the arrivals of contexts, the arm selections, and the observed rewards. Then, using the second constraint of our LP in (6), the last expression of (10) equals the following:

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{n=1}^{2}\sum_{j\neq\alpha}\mathbb{E}[\beta_n q_j^t(\chi_n)]\max_{n=1,2}\Delta_{n,j}\\ =&\sum_{t=1}^{T}\sum_{n=1}^{2}\sum_{j\neq\alpha}\mathbb{E}[\beta_n p_j^t(\chi_n)]\max_{n=1,2}\Delta_{n,j}\\ =&\sum_{t=1}^{T}\sum_{n=1}^{2}\sum_{j\neq\alpha}\mathbb{E}[\mathbb{E}[\mathbf{1}(x_t=\chi_n)]\cdot\mathbb{P}(\bar{a}_t=j\mid x_t=\chi_n,\mathcal{F}_t)]\max_{n=1,2}\Delta_{n,j}.\end{aligned}$$

In this way, we convert the regret of our truthful-TS mechanism into the regret under the Thompson sampling algorithm. We then rewrite each term involving arm j in the last expression as follows:

$$\sum_{t=1}^{T}\sum_{n=1}^{2}\mathbb{E}[\mathbb{P}(\bar{a}_t=j\mid x_t=\chi_n,\mathcal{F}_t)\mathbf{1}(x_t=\chi_n)],\tag{11}$$

which is obtained by moving $\mathbb{P}(\bar{a}_t=j\mid x_t=\chi_n,\mathcal{F}_t)$ inside the inner expectation and applying the tower rule. Then (11) is upper bounded by

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{n=1}^{2}\mathbb{E}[\mathbb{P}(\bar{a}_t=j\mid x_t=\chi_n,\mathcal{F}_t,M_n(t)=t-1)]\\ =&\sum_{t=1}^{T}\sum_{n=1}^{2}\mathbb{E}[\mathbf{1}(\bar{a}_t=j)|x_t=\chi_n,M_n(t)=t-1],\end{aligned}\tag{12}$$

where $M_n(t)$ represent the number of agent arrivals with context χ_n before time t. From the derivation above, we upper bound the $\mathbb{E}[\beta_n\mathbb{E}[\mathbf{1}(a_t=j)|x_t=\chi_n,\mathcal{F}_t]]$ in (10) by the expected number of pulls of suboptimal arm j in Thompson sampling, conditioned on the history where χ_n consistently arrives.

Define $\tilde{\theta}_{n,k}^t$ as the value sampled from the posterior distribution for arm k by context χ_n at time t in the Thompson sampling algorithm. Inspired by the method from [6], which bounds the number of suboptimal arm pulls in Thompson sampling, we decompose

Eq. (12) by considering the following two events: one event $E^{\mu}_{n,j}(t)$ where the posterior mean estimate $\chi_n^\top \hat{\theta}^t_k$ does not deviate significantly from the true value $\chi_n^\top \theta_k$, and the other event $E^{\theta}_{n,j}(t)$ where $\tilde{\theta}^t_{n,k}$ remains close to the posterior mean $\chi_n^\top \hat{\theta}^t_k$ at time t. These events are formally defined as follows:

$$E^{\mu}_{n,j}(t) : \chi_n^\top \hat{\theta}^t_j \le \chi_n^\top \theta_j + \frac{\Delta_{n,j}}{3},$$
$$E^{\theta}_{n,j}(t) : \tilde{\theta}^t_{n,j} \le \chi_n^\top \hat{\theta}^t_j + \frac{\Delta_{n,j}}{3}, \quad n \in \{1,2\},\ j \in [K].$$

Considering the realization of these events above, we decompose (12) as follows:

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{n=1}^{2}\mathbb{E}[\mathbf{1}(\bar{a}_t = j) | x_t = \chi_n, M_n(t) = t-1]\\ =&\sum_{t=1}^{T}\sum_{n=1}^{2}\mathbb{E}\Big[\mathbf{1}(\bar{a}_t = j, E^{\mu}_{n,j}(t), E^{\theta}_{n,j}(t)) + \mathbf{1}(\bar{a}_t = j, E^{\mu}_{n,j}(t), \overline{E^{\theta}_{n,j}(t)})\\ &+\mathbf{1}(\bar{a}_t = j, \overline{E^{\mu}_{n,j}(t)}) \mid x_t = \chi_n, M_n(t) = t-1\Big]. \qquad (13)\end{aligned}$$

Inspired by the method in [6], we upper bound the three terms using the following Lemmas 5.2, 5.3, and 5.4, respectively. The upper bound of the first term directly follows from Lemma 5.2. The upper bounds of the second and third terms are obtained from Lemma 5.3 and Lemma 5.4 by setting $\delta = \Delta_{n,j}/3$. Let $C_{n,j}$ summarize all the constant parts in (12), then we obtain and complete the proof of Theorem 5.1.

Lemma 5.2. *In Thompson sampling with arm action $\bar{a}_t$ at time t, the expected total number of pulls of a suboptimal arm $j \ne \alpha$ by context $\chi_n \in \mathcal{X}$, together with the occurrence of events $E^{\mu}_{n,j}(t)$ and $E^{\theta}_{n,j}(t)$, can be upper bounded by a constant $C^1_{n,j}$ for $n \in \{1,2\}$ and $j \in [K]$:*

$$\sum_{t=1}^{T}\mathbb{E}\Big[\mathbf{1}(\bar{a}_t = j, E^{\mu}_{n,j}(t), E^{\theta}_{n,j}(t)) \mid x_t = \chi_n, M_n(t) = t-1\Big] < C^1_{n,j}.$$

Lemma 5.3. *In Thompson sampling with arm action $\bar{a}_t$ at time t, the expected total number of pulls of a suboptimal arm $j \ne \alpha$ by context $\chi_n \in \mathcal{X}$, together with the occurrence of event $\tilde{\theta}^t_{n,j} - \chi_n^\top \hat{\theta}^t_j > \delta$ for $\delta > 0$, can be upper bounded as follows:*

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}\Bigg[\mathbf{1}(\bar{a}_t = j, \tilde{\theta}^t_{n,j} - \chi_n^T \hat{\theta}^t_j > \delta) \mid x_t = \chi_n, M_n(t) = t-1\Bigg]\\ \le&\frac{2}{\delta^2}\ln\frac{T\delta^2}{4} - \frac{1}{\chi_n^\top V_j \chi_n} + \frac{2}{\delta^2}.\end{aligned}$$

Lemma 5.4. *In Thompson sampling with arm action $\bar{a}_t$ at time t, the expected total number of pulls of suboptimal arm $j \ne \alpha$ under any context $\chi_n \in \mathcal{X}$, together with the occurrence of event $\chi_n^\top \hat{\theta}^t_j - \chi_n^\top \theta_j > \delta$ for $\delta > 0$, can be upper bounded as follows:*

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}\Bigg[\mathbf{1}(\bar{a}_t = j, \chi_n^T \hat{\theta}^t_j - \chi_n^\top \theta_j > \delta) \mid x_t = \chi_n, M_n(t) = t-1\Bigg]\\ \le&\max\left(\left\lceil \frac{\chi_n^\top(\mu_j - \theta_j)}{\delta\chi_n^\top V_j \chi_n} - \frac{1}{\chi_n^\top V_j \chi_n}\right\rceil, 0\right) + 1 + \frac{\exp\left(\frac{\delta\chi_n^T(\mu_j - \theta_j)}{2\chi_n^T V_j \chi_n}\right)}{\delta^2}.\end{aligned}$$

□

Theorem 5.1 and our proof demonstrate that our truthful-TS mechanism shares the same regret order as the Thompson sampling algorithm, indicating that our mechanism can achieve an optimal regret order while ensuring truthfulness.

6 REGRET ANALYSIS UNDER DIFFERENT OPTIMAL ARMS FOR TWO CONTEXTS

When two contexts have different optimal arms, define the true optimal arms for χ_1 and χ_2 as α_1 and α_2, respectively, with $\alpha_1 \ne \alpha_2$. The regret-bounding method used in the previous section, where contexts share the same optimal arm, cannot be applied here. This is because arm α_{3-n} is suboptimal for context χ_n but optimal for context χ_{3-n}, which prevents us from jointly bounding the total number of pulls of arm α_{3-n} across both contexts. As a result, we must also consider the number of pulls of the optimal arm α_{3-n} for context χ_{3-n}, which is not considered when bounding the regret in Thompson sampling.

First, we illustrate the misreporting incentives over time in Thompson sampling when agents have different optimal arms based on their beliefs, which provides insight into our proof approach. For illustration, we assume an ideal Thompson sampling scenario in which agents report truthfully in Fig. 2 and 3. Initially, agents may have an incentive to misreport even if they have different prior optimal arms, as illustrated in Fig. 2. This is because, at first, the variance of the prior distribution is relatively large to encourage exploration, causing the range under the Thompson sampling mapping from $\mathcal{X}$ to $\mathrm{conv}(\hat{\theta}^t_1, \hat{\theta}^t_2, \hat{\theta}^t_3)$ to be concentrated near the center of $\mathrm{conv}(\hat{\theta}^t_1, \dots, \hat{\theta}^t_K)$. As a result, even though the probability of selecting χ_2's prior optimal arm 2 is higher than that of χ_1 ($p^t_2(\chi_2) > p^t_2(\chi_1)$), χ_2 can still obtain a higher expected reward $\chi_2^\top \Theta^t \mathbf{p}^t(\chi_1)$ by misreporting as χ_1. As the algorithm progresses and posterior variance decreases, the range of the Thompson sampling mapping shifts towards the extreme points of the simplex Δ^3, as shown in Fig. 3, where each context ultimately achieves a higher expected reward under its own Thompson sampling distribution. Thus, when contexts have different optimal arms, the number of pulls of arm α_{3-n} by χ_n can be analyzed by separately considering the time steps when the contexts are in conflict and when they are not.

Theorem 6.1. *For the realization of prior, $\{\theta_k\}_{k\in[K]}$ such that the two contexts χ_1 and χ_2 have the different optimal arms α_1 and α_2, the frequentist regret of our truthful-TS mechanism in Mechanism 1 is $O(\ln T)$ to be upper bounded by*

$$\begin{aligned}&\sum_{n=1}^{2}\Bigg(\sum_{j\ne\alpha_1,\alpha_2}\frac{18}{\Delta^2_{n,j}}\ln\frac{T\Delta^2_{n,j}}{36} + C_{n,j}\Bigg)\Bigg(\frac{\beta_n}{\beta_{3-n}}\Delta_{n,\alpha_{3-n}} + \max_{n=1,2}\Delta_{n,j}\Bigg)\\ &+\sum_{n=1}^{2}\Bigg(\Bigg(2 + \frac{\beta_n}{\beta_{3-n}}\Bigg)\Bigg(\frac{18}{\Delta^2_{n,\alpha_{3-n}}}\ln\frac{T\Delta^2_{n,\alpha_{3-n}}}{36} + C_{n,\alpha_{3-n}}\Bigg)\\ &+\frac{2048}{\Delta^2_{3-n}}\ln\frac{T\Delta^2_{3-n}}{2048} + \frac{64}{\Delta^2_{3-n}}\ln\sqrt{\frac{2}{\pi}}\frac{T\Delta^2_{3-n}}{64} + D_{3-n}\\ &+\frac{\beta_n}{\beta_{3-n}}\Bigg(\frac{2048}{\Delta^2_n}\ln\frac{T\Delta^2_n}{2048} + \frac{64}{\Delta^2_n}\ln\sqrt{\frac{2}{\pi}}\frac{T\Delta^2_n}{64} + D_n\Bigg)\Bigg)\Delta_{n,\alpha_{3-n}},\end{aligned}$$

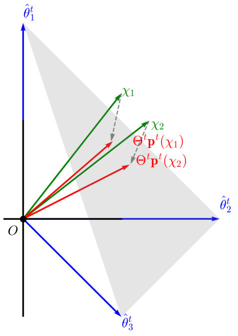

Figure 2: Initial stage for contexts with different optimal arms

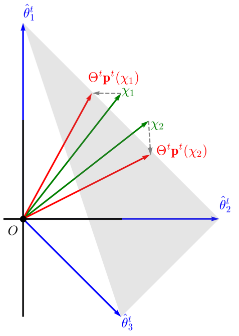

Figure 3: Converging stage for contexts with different optimal arms

where $\Delta_{n,j} = \chi_n^\top\theta_{\alpha_n} - \chi_n^\top\theta_j$ is the reward gap between the optimal arm α_n and arm $j \neq \alpha_n$ for context χ_n. $\Delta_n = \min_{j\neq n}\Delta_{n,j}$ is the minimum reward gap for context χ_n. $C_{n,j}$ is a constant for $n \in \{1,2\}$ and $j \in [K]$, and D_n is the constant part for $n \in \{1,2\}$.

Proof. To upper bound the regret when the contexts have different optimal arms, we need to consider the expected number of times that either context has an incentive to misreport, which introduces an additional part on regret analysis compared to Theorem 5.1. Define $I(t)$ as the event that the two contexts have conflict at time t, meaning:

$$I(t) : \chi_1^\top\Theta^t(\mathbf{p}^t(\chi_2) - \mathbf{p}^t(\chi_1)) > 0 \text{ or } \chi_2^\top\Theta^t(\mathbf{p}^t(\chi_1) - \mathbf{p}^t(\chi_2)) > 0.$$

Similar to the proof of Theorem 5.1, we first decompose the regret in (4) as follows:

$$
\begin{aligned}
&\mathbb{E}[\mathcal{R}(T)]\\
=&\sum_{t=1}^{T}\sum_{n=1}^{2}\sum_{j\neq\alpha_n}\mathbb{E}[\mathbf{1}(x_t=\chi_n, a_t=j)]\Delta_{n,j}\\
\leq&\sum_{t=1}^{T}\sum_{n=1}^{2}\left(\left(\sum_{j\neq\alpha_1,\alpha_2}\mathbb{E}[\mathbf{1}(x_t=\chi_n,\bar{a}_t=j)]\max_{n=1,2}\Delta_{n,j}\right)+\Delta_{n,\alpha_{3-n}}\right.\\
&\left.\cdot\,\mathbb{E}[\mathbf{1}(x_t=\chi_n, a_t=\alpha_{3-n},\overline{I(t)})+\mathbf{1}(x_t=\chi_n,a_t=\alpha_{3-n},I(t))]\right)\\
\leq&\sum_{t=1}^{T}\sum_{n=1}^{2}\left(\left(\sum_{j\neq\alpha_1,\alpha_2}\mathbb{E}[\mathbf{1}(x_t=\chi_n,\bar{a}_t=j)]\max_{n=1,2}\Delta_{n,j}\right)+\right.\\
&\left.\mathbb{E}[\mathbf{1}(x_t=\chi_n,\bar{a}_t=\alpha_{3-n})+\mathbf{1}(a_t=\alpha_{3-n},I(t))]\Delta_{n,\alpha_{3-n}}\right),
\end{aligned}
\tag{14}
$$

where the last inequality is because when $\overline{I(t)}$ happens, the mechanism will operate the same as Thompson sampling. We further derive an upper bound for the last term in the last expression of (14) as follows:

$$
\begin{aligned}
&\mathbb{E}[\mathbf{1}(a_t=\alpha_{3-n},I(t))]\\
=&\mathbb{E}[\mathbf{1}(x_t=\chi_n,\bar{a}_t=\alpha_{3-n},I(t))+\mathbf{1}(x_t=\chi_{3-n},\bar{a}_t=\alpha_{3-n},I(t))]\\
\leq&\mathbb{E}[\mathbf{1}(x_t=\chi_n,\bar{a}_t=\alpha_{3-n})+\mathbf{1}(x_t=\chi_{3-n},\bar{a}_t=\alpha_{3-n},I(t))],
\end{aligned}
$$

where the equality holds because a_t is chosen according to $\mathbf{q}^t(x_t)$ from (6), whose second constraint enables us to convert it to $\bar{a}_t$ for Thompson sampling. Substituting the last expression into the last expression of (14), we obtain an upper bound for (14) given by

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{n=1}^{2}((\sum_{j\neq\alpha_n,\alpha_2}\mathbb{E}[\mathbf{1}(x_t=\chi_n,\bar{a}_t=j)]\max_{n=1,2}\Delta_{n,j})\\
&+(2\mathbb{E}[\mathbf{1}(x_t=\chi_n,\bar{a}_t=\alpha_{3-n})]\\
&+\mathbb{E}[\mathbf{1}(x_t=\chi_{3-n},\bar{a}_t=\alpha_{3-n},I(t))])\Delta_{n,\alpha_{3-n}}).
\end{aligned}
\tag{15}
$$

The first two lines of Eq. (15) can be upper bounded using the procedure for bounding the number of suboptimal pulls described in (11) to (13) from the last section. For the last term, since it only relates to actions under Thompson sampling, we proceed to upper bound it by analyzing a virtual process with an ideal Thompson sampling algorithm that does not involve misreporting. We first present Lemma 6.2 that transforms the event including $I(t)$ to another event involving $p_{\alpha_1}^t(\chi_1) + p_{\alpha_2}^t(\chi_2)$.

Lemma 6.2. *Let $\underline{k_n}$ denote the empirically worst arm for context χ_n in any bandit process at any given time. Let ϵ^t denote*

$$\min_n \frac{\chi_n^\top\hat{\theta}_{\alpha_n}^t - \chi_n^\top\hat{\theta}_{\alpha_{3-n}}^t}{2(\chi_n^\top\hat{\theta}_{\alpha_n}^t - \chi_n^\top\hat{\theta}_{\underline{k_n}}^t + \max_{k\neq\alpha_1,\alpha_2}\chi_n^T(\hat{\theta}_k^t - \hat{\theta}_{\underline{k_n}}^t))}. \tag{16}$$

The probability of event $I(t)$, where either χ_1 or χ_2 has an incentive to misreport at time t, can be upper bounded by the probability of the following event:

$$E[\mathbf{1}(I(t))] \leq E\left[\mathbf{1}(p_{\alpha_n}^t(\chi_n) + p_{\alpha_{3-n}}^t(\chi_{3-n}) \leq 2 - 2\epsilon^t\right].$$

Using Lemma 6.2, we then decompose the last term in (15) below:

$$
\begin{aligned}
&\mathbb{E}[\mathbf{1}(x_t=\chi_{3-n},\bar{a}_t=\alpha_{3-n},I(t))]\\
\leq&\mathbb{E}[\mathbf{1}(x_t=\chi_{3-n},\bar{a}_t=\alpha_{3-n},p_{\alpha_1}^t(\chi_1)+p_{\alpha_2}^t(\chi_2)<2-2\epsilon^t)]\\
\leq&\mathbb{E}[\mathbf{1}(x_t=\chi_{3-n},\bar{a}_t=\alpha_{3-n},p_{\alpha_{3-n}}^t(\chi_{3-n})<1-\epsilon^t)]\\
&+\mathbb{E}[\mathbf{1}(x_t=\chi_{3-n},\bar{a}_t=\alpha_{3-n},p_{\alpha_n}^t(\chi_n)<1-\epsilon^t)]\\
\leq&\mathbb{E}[\mathbf{1}(x_t=\chi_{3-n},\bar{a}_t=\alpha_{3-n},p_{\alpha_{3-n}}^t(\chi_{3-n})<1-\epsilon^t)]\\
&+\mathbb{E}[\mathbf{1}(x_t=\chi_{3-n},p_{\alpha_n}^t(\chi_n)<1-\epsilon^t)]\\
=&\mathbb{E}[\mathbf{1}(x_t=\chi_{3-n},\bar{a}_t=\alpha_{3-n},p_{\alpha_{3-n}}^t(\chi_{3-n})<1-\epsilon^t)]\\
&+\mathbb{E}[\mathbb{E}[\mathbf{1}(x_t=\chi_{3-n})|\mathcal{F}_t]\mathbb{E}[\mathbf{1}(p_{\alpha_n}^t(\chi_n)<1-\epsilon^t)|\mathcal{F}_t]]\\
=&\mathbb{E}[\mathbf{1}(x_t=\chi_{3-n},\bar{a}_t=\alpha_{3-n},p_{\alpha_{3-n}}^t(\chi_{3-n})<1-\epsilon^t)]\\
&+\mathbb{E}[\frac{\beta_n}{\beta_{3-n}}\mathbb{E}[\mathbf{1}(x_t=\chi_n)|\mathcal{F}_t]\mathbb{E}[\mathbf{1}(p_{\alpha_n}^t(\chi_n)<1-\epsilon^t)|\mathcal{F}_t]]\\
=&\mathbb{E}[\mathbf{1}(x_t=\chi_{3-n},\bar{a}_t=\alpha_{3-n},p_{\alpha_{3-n}}^t(\chi_{3-n})<1-\epsilon^t)]\\
&+\frac{\beta_n}{\beta_{3-n}}\mathbb{E}[\mathbf{1}(x_t=\chi_n,p_{\alpha_n}^t(\chi_n)<1-\epsilon^t)].
\end{aligned}
$$

The equations above hold because, given the history $\mathcal{F}_t$, the context arrival is independent of $p_{\alpha_{3-n}}^t(\chi_{3-n})$. The final expression can be further decomposed and upper bounded as follows:

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}(x_t = \chi_{3-n}, \bar{a}_t = \alpha_{3-n}, p^t_{\alpha_{3-n}}(\chi_{3-n}) < 1 - \epsilon^t)] \\
&+ \frac{\beta_n}{\beta_{3-n}} \mathbb{E}[\mathbf{1}(x_t = \chi_n, \bar{a}_t = \alpha_n, p^t_{\alpha_n}(\chi_n) < 1 - \epsilon^t)] \\
&+ \sum_{i \neq \alpha_n} \frac{\beta_n}{\beta_{3-n}} \mathbb{E}[\mathbf{1}(x_t = \chi_n, \bar{a}_t = i)]. \qquad (17)
\end{aligned}$$

When summing from $t = 1$ to T, the term in the last line above can be upper bounded using the same method as in (11) through (13) from the last section, then applying Lemmas 5.2, 5.3, and 5.4. The upper bound for the summation from $t = 1$ to T of the first two terms in (17) is provided in Lemma 6.3 below.

Lemma 6.3. *Let ϵ^t denote the expression in (16). Then the expected number of pulls of α_n by χ_n together with the occurrence of event $p^t_{\alpha_n}(\chi_n) < 1 - \epsilon^t$ is upper bounded by:*

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}[\mathbf{1}(x_t = \chi_n, \bar{a}_t = \alpha_n, p^t_{\alpha_n}(\chi_n) < 1 - \epsilon^t)] \\
\leq &\frac{2048}{\Delta_n^2} \ln \frac{T\Delta_n^2}{2048} + \frac{64}{\Delta_n^2} \ln \sqrt{\frac{2}{\pi}} \frac{T\Delta_n^2}{64} + D_n,
\end{aligned}$$

where D_n is a constant for $n \in \{1, 2\}$.

By substituting (17) into (15) and then substituting Lemmas 5.2, 5.3, 5.4, and 6.3, we derive the final result of Theorem 6.1. □

7 SIMULATION EXPERIMENTS

In this section, we use simulation experiments to evaluate the performance of our truthful-TS mechanism for more than two contexts and non-Gaussian noise distribution.

We first extend Gaussian noise $\eta_t \sim \mathcal{N}(0, 1)$ to Laplace noise as a typical sub-Gaussian distribution in Fig. 4. Since the Laplace noise yields a non-closed form and non-standard posterior distribution, we turn to numerical methods to update the posterior and compute the probabilities $\mathbf{p}^t(x)$ in (5). To ensure a small error of inherent approximation in numerical methods, we conduct this experiment for the small scale case with $N = d = K = 2$ and $T = 50$ time steps. For a fair regret comparison, we set the same Gaussian prior for both noises where prior mean $\mu_1 = (0, 1)$ and $\mu_2 = (1, 0)$ and prior covariance matrix $V_1 = V_2 = I$. The variances of both Laplace and Gaussian noises are set to 1 in Fig. 4. To compare with our mechanism in Mechanism 1, we use the ideal Thompson sampling (always assuming agents' truthful reporting) as the performance upper bound. Fig. 4 implies that our truthful-TS mechanism yields a similar regret order to Thompson sampling. Since Laplace noise has a heavier tail than Gaussian noise, the regret order under Laplace noise remains sublinear but is slightly higher than that under Gaussian noise. As observed, the regrets for both the truthful-TS mechanism and the Thompson sampling algorithm under Laplace noise will exceed those under Gaussian noise after $t = 50$. We can have the same conclusion when extending to other sub-Gaussian noises such as uniform and Cauchy distributions.

Similar to the experiment setting in [8], we then extend the setting to $N = 3, 5, 9$ contexts and $d = 5, 9$ dimensions under Gaussian prior and noise. We consider $K = 6$ arms for recommendations among these contexts. For each arm k, we set its prior distribution $\mathcal{P}_k \sim \mathcal{N}(\mu_k, V_k)$ such that μ_k is a vector with a single 1 in the k-th entry and V_k is the identity matrix. Different contexts are sampled from a multivariate uniform distribution over $[0, 1]^d$. For each parameter setting, we run 100 simulations, generating new realizations of $\{\theta_k\}_{k \in [K]}$ from the prior distribution of $\mathcal{N}(\mu_k, V_k)$ each time, and calculate the average regret. The results are displayed in Fig. 5. According to Fig. 5, like the ideal Thompson sampling, our truthful-TS mechanism still exhibits sublinear order for different N and d, which is consistent with our Theorems 5.1 and 6.1. Though ideal Thompson sampling algorithm's regret grows with N and d, our mechanism grows faster. The reason is that as there are more contexts, a context may envy more other contexts with higher convergence rates and our Mechanism 1 will reduce the exploitation of those contexts, thereby slowing down the overall convergence.

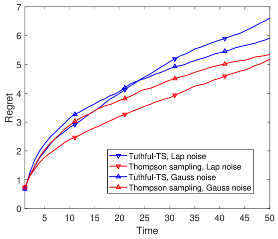

Figure 4: Cumulative regret at each time step under our truthful-TS mechanism in Mechanism 1 and Thompson sampling algorithm, and Gaussian and Laplace noises with 2 contexts, 2 arms, and 2 dimensions.

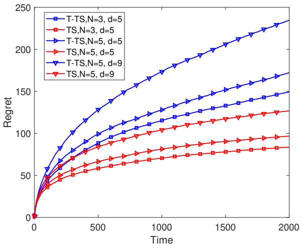

Figure 5: Cumulative regret at each time step under our Mechanism 1 (T-TS) and Thompson sampling algorithm (TS), where the number of contexts varies between $N \in \{3, 5, 9\}$ and $d \in \{5, 9\}$.

8 CONCLUSION

In this paper, we investigate the problem of strategic misreporting of private contexts by agents within the Bayesian contextual linear bandit framework. We are the first to analyze this issue and demonstrate that existing algorithms fail to perform effectively under such misreporting behavior. To address this, we propose a novel truthful mechanism based on the Thompson sampling algorithm, which solves an LP at each time step to ensure incentive compatibility. We prove that our mechanism achieves a problem-dependent regret bound of $O(\ln T)$ in the two-context case with Gaussian priors and noise. Furthermore, our numerical results suggest that the proposed mechanism retains a comparable regret order across multiple contexts and under heavier tails of noise.

ACKNOWLEDGMENTS

This work is also supported in part by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 Grant with Award no. MOE-T2EP20121-0001; in part by SUTD Kickstarter Initiative (SKI) Grant with no. SKI 2021_04_07; and in part by the Joint SMU-SUTD Grant with no. 22-LKCSB-SMU-053.

REFERENCES

[1] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, Vol. 24.

[2] Marc Abeille and Alessandro Lazaric. 2017. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*. PMLR, 176–184.

[3] Milton Abramowitz and Irene A Stegun. 1948. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Vol. 55. US Government printing office.

[4] Akinlolu Adekotujo, Adedoyin Odumabo, Ademola Adedokun, and Olukayode Aiyeniko. 2020. A comparative study of operating systems: Case of windows, unix, linux, mac, android and ios. *International Journal of Computer Applications* 176, 39 (2020), 16–23.

[5] Shipra Agrawal and Navin Goyal. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*. PMLR, 127–135.

[6] Shipra Agrawal and Navin Goyal. 2017. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)* 64, 5 (2017), 1–24.

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

[8] Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. 2021. Mostly exploration-free algorithms for contextual bandits. *Management Science* 67, 3 (2021), 1329–1349.

[9] Thomas Kleine Buening, Aadirupa Saha, Christos Dimitrakakis, and Haifeng Xu. 2024. Strategic Linear Contextual Bandits. *arXiv preprint arXiv:2406.00551* (2024).

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

[11] Michael B Cohen, Yin Tat Lee, and Zhao Song. 2021. Solving linear programs in the current matrix multiplication time. *Journal of the ACM (JACM)* 68, 1 (2021), 1–39.

[12] U.S. Food and Drug Administration. 2017. *Drug Approval Package: Verzenio (abemaciclib)*. Technical Report. U.S. Food and Drug Administration. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2017/208716Orig1s000TOC.cfm Accessed: 2024-09-04.

[13] Alexander Goldenshluger and Assaf Zeevi. 2013. A linear response bandit problem. *Stochastic Systems* 3, 1 (2013), 230–261.

[14] GrowthSetting. n.d.. How Netflix Enhances User Experience with AI Recommendations. https://growthsetting.com Accessed: 2024-10-11.

[15] Xinyan Hu, Dung Ngo, Aleksandrs Slivkins, and Steven Z Wu. 2022. Incentivizing combinatorial bandit exploration. *Advances in Neural Information Processing Systems* 35 (2022), 37173–37183.

[16] Yiting Hu and Lingjie Duan. 2025. Truthful mechanisms for linear bandit games with private contexts. *arXiv preprint arXiv:2501.03865* (2025).

[17] Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. 2020. Incentivizing Exploration with Selective Data Disclosure. In *Proceedings of the 21st ACM Conference on Economics and Computation*. 647–648.

[18] Ilan Kremer, Yishay Mansour, and Motty Perry. 2014. Implementing the "wisdom of the crowd". *Journal of Political Economy* 122, 5 (2014), 988–1012.

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

[20] Alessandro Lazaric and Rémi Munos. 2009. Hybrid Stochastic-Adversarial On-line Learning.. In *COLT*. Citeseer.

[21] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. 2020. Bayesian incentive-compatible bandit exploration. *Operations Research* 68, 4 (2020), 1132–1161.

[22] Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. 2022. Bayesian exploration: Incentivizing exploration in Bayesian games. *Operations Research* 70, 2 (2022), 1105–1127.

[23] Joanne Neale, Michele Robertson, and Michael Bloor. 2007. 'Treatment experienced'and 'treatment naïve'drug agency clients compared. *International Journal of Drug Policy* 18, 6 (2007), 486–493.

[24] Rifaquat Rahman, Lorenzo Trippa, Eudocia Q Lee, Isabel Arrillaga-Romany, Geoffrey Fell, Mehdi Touat, Christine McCluskey, Jennifer Wiley, Sarah Gaffey, Jan Drappatz, et al. 2023. Inaugural results of the individualized screening trial of innovative glioblastoma therapy: a phase II platform trial for newly diagnosed glioblastoma using Bayesian adaptive randomization. *Journal of Clinical Oncology* 41, 36 (2023), 5524–5535.

[25] Alexander Rakhlin and Karthik Sridharan. 2016. Bistro: An efficient relaxation-based method for contextual bandits. In *International Conference on Machine Learning*. PMLR, 1977–1985.

[26] Daniel Russo and Benjamin Van Roy. 2014. Learning to optimize via posterior sampling. *Mathematics of Operations Research* 39, 4 (2014), 1221–1243.

[27] Daniel Russo and Benjamin Van Roy. 2016. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research* 17, 68 (2016), 1–30.

[28] Mark Sellke. 2023. Incentivizing exploration with linear contexts and combinatorial actions. In *International Conference on Machine Learning*. PMLR, 30570–30583.

[29] Mark Sellke and Aleksandrs Slivkins. 2021. The price of incentivizing exploration: A characterization via thompson sampling and sample complexity. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 795–796.

[30] Max Simchowitz and Aleksandrs Slivkins. 2024. Exploration and incentives in reinforcement learning. *Operations Research* 72, 3 (2024), 983–998.

[31] Aleksandrs Slivkins et al. 2019. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning* 12, 1-2 (2019), 1–286.