STRATEGYPROOF ALGORITHMS IN LINEAR CONTEXTUAL BANDITS WITH PRIVATE CONTEXTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contextual bandits have broad applications, including clinical trials and LLM routing. In a clinical trial, a doctor uses patients' private information, such as symptoms and family history, to assign treatments and learn from observed outcomes. To improve overall patient welfare, the doctor may assign seemingly suboptimal treatments to some patients to reduce uncertainty, whereas individual patients seek to maximize their immediate outcomes. This misalignment may incentivize patients' misreporting to avoid exploration, substantially biasing treatment-effect estimates and reducing long-term patient welfare. It is therefore important for bandit algorithms to be strategyproof, ensuring patients cannot benefit from misreporting their contexts. In this paper, we conduct a comprehensive study of strategyproof algorithms for linear bandits with private contexts. We first establish complementary gaps of classical strategyproof bandit algorithms. Explore-then-commit and truthful Thompson sampling either lack $O(\mathrm{polylog}\, T)$ problem-dependent regret guarantees or can incur linear problem-independent regret, while non-strategyproof algorithms such as LinUCB generally suffer linear regret in strategic settings. Secondly, we establish a novel minimax regret lower bound of $\Omega(T^{2/3})$ for strategyproof bandit algorithms, revealing the fundamental challenge posed by strategic misreporting. Motivated by this, we propose Confidence-Gated Truthful UCB (CGT-UCB), which first employs a confidence gate to construct a proposal policy that adaptively switches between greedy and optimistic policies for agents with different contexts. This mechanism avoids the excessive exploration induced by the standard optimism-in-the-face-of-uncertainty principle. CGT-UCB then feeds the proposal policy into a specially designed linear programming problem that enforces strategyproofness while balancing exploration and exploitation under strategyproofness. CGT-UCB simultaneously achieves $\widetilde{O}(T^{2/3})$ problem-independent regret and $O(\mathrm{polylog}\, T)$ problem-dependent regret. Experiments on synthetic and real-world datasets support our theoretical findings and demonstrate the practical effectiveness of CGT-UCB.

1 INTRODUCTION

Human-in-the-loop online learning systems repeatedly make decisions for agents with heterogeneous contextual information. For example, in a clinical trial, a doctor assigns a treatment based on a patient's contextual information, such as symptoms and medical history, and observes the corresponding treatment outcome to improve future treatment decisions (Food & Administration, 2017; Rahman et al., 2023). In LLM routing, a platform assigns different LLMs to queries of different types and observes the response quality of the chosen LLM to improve future model assignments (Ong et al., 2025; Poon et al., 2026). Contextual bandits provide a principled framework for making such decisions under uncertainty and partial feedback (Auer, 2002; Chu et al., 2011; Abbasi-yadkori et al., 2011; Agrawal & Goyal, 2013; Abeille & Lazaric, 2017). In each round, the system collects an agent's context, selects an action, and observes the reward generated by the selected action, which it uses to improve future decisions. The system's objective is to minimize cumulative regret, that is, the expected cumulative reward loss relative to an oracle that knows the expected rewards and selects an optimal action under each context.

However, because the system's learning process relies on the participation of many agents with diverse personal objectives, the system's long-term learning objective may conflict with agents' indi-

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

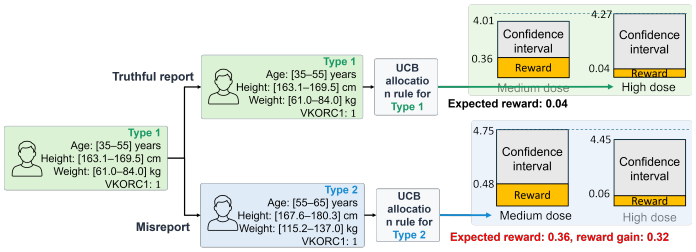

Figure 1: Misreporting in warfarin dose allocation on the Warfarin dataset (Consortium, 2009).

vidual interests. For example, learning the relative effectiveness of treatments can require assigning patients to treatments that currently appear less promising, while each patient seeks the highest expected benefit for their own health. This conflict leads agents to withhold cooperation from the system's learning process. A related line of work studies settings in which self-interested agents can select their own actions and disregard the system's recommendations (Mansour et al., 2020; Sellke & Slivkins, 2021; Hu et al., 2022; Sellke, 2023; Immorlica et al., 2020; Simchowitz & Slivkins, 2024). A more serious challenge arises when agents strategically misreport their private contexts to manipulate the system's allocations (Harris et al., 2022; Kleine Buening et al., 2024; Hu & Duan, 2025). Such agents' behavior corrupts the system's learning data and biases reward estimates, causing persistent allocation errors that harm future agents and undermine long-term social welfare. Existing studies explore several distinct settings of context misreporting. Kleine Buening et al. (2024) model arms as strategic agents that manipulate their contexts to increase their chances of selection, while Harris et al. (2022) studies a bounded-manipulation model in the strategic classification problem.

More closely aligned with the applications discussed at the beginning of this section, Hu & Duan (2025) assumes that agents misreport private contexts to avoid exploration and improve their immediate expected rewards. This work also studies this model, which can be illustrated by the example of warfarin dose allocation in Figure 1, where a doctor assigns different warfarin doses to patients based on their contextual information. In a particular round, according to the standard OFUL algorithm (Abbasi-yadkori et al., 2011), the UCB estimates of the high and medium doses are 4.27 and 4.01, respectively, for type 1 patients, and 4.45 and 4.75, respectively, for type 2 patients. OFUL accordingly assigns the high dose to type 1 patients and the medium dose to type 2 patients. However, the estimated reward for a type 1 patient is 0.04 under the high dose but 0.36 under the medium dose. The type 1 patient can therefore misreport her type as type 2 to receive the medium dose and increase her immediate reward from 0.04 to 0.36.

Despite its practical importance, strategic context misreporting to avoid exploration remains largely underexplored. Standard bandit algorithms can incur linear regret under strategic misreporting, motivating strategyproof algorithms that make misreporting unprofitable for agents (Hu & Duan, 2025). However, few existing algorithms achieve both strategyproofness and sublinear regret. Regret guarantees are studied in two important complementary regimes. Problem-independent bounds provide worst-case guarantees uniformly over instances (Abbasi-yadkori et al., 2011), while problem-dependent bounds account for instance-specific hardness (Hao et al., 2020). Among strategyproof bandit algorithms, explore-then-commit (ETC) (Langford & Zhang, 2007; Komiyama & Imaizumi, 2023) achieves strategyproofness through context-independent exploration but incurs $\widetilde{O}(T^{2/3})$ regret in both problem-dependent and problem-independent regimes. Greedy allocation is also strategyproof but suffers linear problem-independent regret. Truthful Thompson sampling algorithm (Hu & Duan, 2025) achieves an $O(\log T)$ problem-dependent regret bound for a special case of only two possible contexts, but its problem-independent regret is actually $\Omega(T)$ as analyzed in this work. Given the currently limited understanding of this problem, we aim to answer the following questions in this work:

What is the fundamental regret lower bound under strategyproofness, and how can strategyproof contextual bandit algorithms simultaneously achieve low problem-dependent and problem-independent regret?

Our main contributions are summarized as follows.

Table 1: Comparison of regret bounds. Here, M denotes the number of agent types, d the model dimension, and K the number of actions.

Method/Lower bound	Problem-independent bound	Problem-dependent bound
Lower bound	$\Omega(\min\{M,d\}^{2/3}T^{2/3})$ (Thm. 3)	$\Omega(\log T)^{\dagger}$
Greedy, OFUL (Abbasi-yadkori et al., 2011), Thompson samp. (Abeille & Lazaric, 2017)	$O(T)$	$O(T)$
ETC (Komiyama & Imaizumi, 2023)	$\widetilde{O}((KMd)^{1/3}T^{2/3})$ (Thm. 1)	$\widetilde{O}((KMd)^{1/3}T^{2/3})$ (Thm. 1)
Rev. truthful Thompson samp. (Alg. 2 in our work)	$O(T)$ (Thm. 2)	$O(d^3\log^3(dT))$ (Thm. 2)
Conf.-gated truthful UCB (Alg. 1 in our work)	$\widetilde{O}(M^{1/3}d^{2/3}T^{2/3})$ (Thm. 4)	$O(Md^2\log^2 T)$ (Thm. 5)

†Asymptotic lower bound. The problem-dependent bound assumes a unique optimal action for each type.

- **Baseline regret bounds.** We evaluate existing algorithms under context misreporting using both problem-independent and problem-dependent regret as complementary measures to evaluate algorithms, as summarized in Table 1. After formulating both frequentist and Bayesian models of linear contextual bandits with private contexts and defining strategyproofness and the regret objective in Section 2, we establish regret bounds for two strategyproof baseline algorithms in Section 3. In particular, ETC achieves $\widetilde{O}(T^{2/3})$ regret in both problem-dependent and problem-independent regimes, whereas truthful Thompson sampling can incur $\Omega(T)$ problem-independent regret despite its $O(\mathrm{polylog}\, T)$ problem-dependent regret. Thus, neither baseline simultaneously guarantees sublinear problem-independent regret and $O(\mathrm{polylog}\, T)$ problem-dependent regret.
- **Worst-case regret lower bound.** To characterize the fundamental performance limit of strategyproof algorithms, we establish a novel minimax regret lower bound of $\Omega(\min\{M,d\}^{2/3}T^{2/3})$ in Section 3. Our lower-bound analysis shows that incentives for strategic misreporting interact across agent types, thereby coupling the system's exploration across types and precluding the optimal learning trajectory.
- **Strategyproof algorithm with low regret in both regimes.** We propose Confidence-Gated Truthful UCB (CGT-UCB) in Section 4. CGT-UCB first uses a confidence gate to construct a proposal policy that adaptively switches between greedy and optimistic actions for agents with different contexts, avoiding the excessive exploration that can arise from the standard optimism-in-the-face-of-uncertainty (OFU) principle. It then incorporates the proposal policy into a specially designed linear programming problem that enforces strategyproofness while balancing exploration and exploitation under strategyproofness. CGT-UCB simultaneously achieves $\widetilde{O}(T^{2/3})$ problem-independent regret and $O(\mathrm{polylog}\, T)$ problem-dependent regret. In Section 5, experiments on synthetic and Warfarin datasets support our theoretical findings and demonstrate CGT-UCB's superior performance compared with the baselines.

2 Problem Formulation

We study a stochastic contextual linear bandit problem with private agent contexts. Let $\mathcal{K} = \{1,\dots,K\}$ denote the set of K actions, and let $\mathcal{M} = \{1,\dots,M\}$ denote the set of M agent types. Each type-action pair $(m,a) \in \mathcal{M}\times\mathcal{K}$ is associated with a known context vector $\phi(m,a) \in \mathbb{R}^d$, where $\|\phi(m,a)\|_2 \le 1$.

The system and agents interact over T rounds. At each round t, a new agent arrives with a private type $I_t \in \mathcal{M}$. Types are drawn independently from a distribution $\boldsymbol{\beta} = (\beta_1,\dots,\beta_M)$ with full support, so that $\Pr(I_t = m) = \beta_m$ for all t and $m \in \mathcal{M}$, with $\beta_{\min} := \min_{m\in\mathcal{M}}\beta_m > 0$. The parameter $\boldsymbol{\beta}$ can be known or unknown. The agent reports $I'_t \in \mathcal{M}$, which may differ from her actual type I_t. The system selects A_t according to $\pi(I'_t \mid \mathcal{F}_t)$, a distribution over the K actions conditional on the reported type I'_t and the observation history $\mathcal{F}_t$. Specifically, $\pi_a(m \mid \mathcal{F}_t) = \Pr(A_t = a \mid I'_t = m, \mathcal{F}_t)$. The agent then receives reward $r_t = \langle\phi(I_t,A_t),\theta^*\rangle + \eta_t$, where $\theta^* \in \mathbb{R}^d$ is the ground truth parameter unknown to both the system and the agents, and η_t is the noise variable at time t. The observation history before round t is $\mathcal{F}_t := (I'_s, A_s, r_s : 1 \le s < t)$. Define the optimality gap

as follows

$$\Delta_{\min} := \min_{m,a:\Delta_{m,a}>0} \Delta_{m,a}, \text{ where } \Delta_{m,a} := \max_{a' \in \mathcal{K}} \langle \phi(m,a'), \theta^* \rangle - \langle \phi(m,a), \theta^* \rangle.$$

We comprehensively study both **frequentist** and **Bayesian** models. In the frequentist model, we assume θ^* is fixed but unknown, satisfying $\|\theta^*\| \leq L/2$ with known L. η_t is conditionally σ-sub-Gaussian, that is, $\mathbb{E}[\exp(u\eta_t) \,|\, \mathcal{F}_t, I_t, I'_t, A_t] \leq \exp\big(u^2\sigma^2/2\big)$ with $u \in \mathbb{R}$. The environment computes the ridge regression estimate $\hat{\theta}_t$ with $\lambda > 0$, which can be observed by the system and all agents:

$$\hat{\theta}_1 = 0,\ \hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t-1} \phi(I'_s, A_s) r_s,\ t > 1, \text{ where } V_t = \lambda I_d + \sum_{s=1}^{t-1} \phi(I'_s, A_s)\phi(I'_s, A_s)^\top. \quad (1)$$

In the Bayesian model, we take Gaussian priors and noise as a representative model. At the very beginning, the parameter θ^* is drawn from a commonly known Gaussian prior, $\theta^* \sim \mathcal{N}(\mu_0, \Sigma_0^{-1})$, where $\Sigma_0 \succ 0$. The $\{\eta_t\}_{t=1}^T$ are i.i.d $\mathcal{N}(0, \sigma^2)$ random variables, independent of θ^* and the type sequence. At each time t, the posterior distribution of θ^* is $\theta^* \mid \mathcal{F}_t \sim \mathcal{N}(\hat{\theta}_t, \Sigma_t^{-1})$, whose mean $\hat{\theta}_t$ serves as the environment's public estimator and is computed as follows:

$$\hat{\theta}_t = \Sigma_t^{-1} \left(\Sigma_0 \mu_0 + \frac{1}{\sigma^2} \sum_{s=1}^{t-1} \phi(I'_s, A_s) r_s \right), \text{where } \Sigma_t = \Sigma_0 + \frac{1}{\sigma^2} \sum_{s=1}^{t-1} \phi(I'_s, A_s)\phi(I'_s, A_s)^\top. \quad (2)$$

Strategic reporting and strategyproofness. As the system can not access the true context, agents with true type m can report a type $i \in \mathcal{M}$ with non-negative cost $c_{m,i}$, where $c_{m,m} = 0$. Since agents are rational and participate only once (e.g., in a clinical trail), they seek to maximize their expected single-round reward based on the public estimate $\hat{\theta}_t$. Under the system's policy π, the agent arriving at time t receives an expected single-round reward of $\mathbb{E}_{A \sim \pi(I'_t | \mathcal{F}_t)} \left[\langle \phi(I_t, A), \hat{\theta}_t \rangle \right]$. Thus, the optimal context reporting strategy satisfies

$$I'_t \in \arg\max_{m \in \mathcal{M}} \mathbb{E}_{A \sim \pi(m | \mathcal{F}_t)} \left[\langle \phi(I_t, A), \hat{\theta}_t \rangle - c_{I_t, m} \right]. \quad (3)$$

We focus on strategyproof algorithms defined as follows.

Definition 1 (Strategyproofness)**.** *An algorithm is strategyproof with respect to the prescribed public estimator $\hat{\theta}_t$, regardless of the misreporting cost c, if its induced policy π satisfies the following condition for every $t \in [T]$, every history $\mathcal{F}_t$, and all $m, j \in \mathcal{M}$:*

$$\mathbb{E}_{A \sim \pi(m | \mathcal{F}_t)} \left[\langle \phi(m, A), \hat{\theta}_t \rangle \right] \geq \mathbb{E}_{A \sim \pi(j | \mathcal{F}_t)} \left[\langle \phi(m, A), \hat{\theta}_t \rangle \right].$$

Strategyproofness ensures that no agent can obtain a higher reward by misreporting and thus eliminates agents' incentive to misreport their features, regardless of the value of the misreporting cost. A strategyproof system aims to minimize the cumulative regret defined below, in contrast to each agent's objective of maximizing their single-round expected reward.

Definition 2 (Regret)**.** *In the frequentist setting, let $\nu(\theta^*) := (\theta^*, \mathcal{K}, \mathcal{M}, \phi, \beta)$ denote a problem instance. For any policy π, we define the following notions of regret.*

1. ***Problem-dependent frequentist regret.*** *For a fixed problem instance ν, the expected cumulative regret is*

$$R_T(\pi; \nu(\theta^*)) := \mathbb{E}\left[\sum_{t=1}^{T} \left(\max_{a \in \mathcal{K}} \langle \phi(I_t, a), \theta^* \rangle - \langle \phi(I_t, A_t), \theta^* \rangle \right) \right]. \quad (4)$$

2. ***Problem-independent frequentist regret.*** *Let $\mathfrak{V}$ denote a class of admissible problem instances. The problem-independent expected regret is*

$$R_T(\pi) := \sup_{\nu(\theta^*) \in \mathfrak{V}} R_T(\pi; \nu(\theta^*)). \quad (5)$$

In the Bayesian setting, we consider Bayesian regret, which additionally takes an expectation over the prior distribution of θ^* for the problem-dependent regret: $\mathrm{BR}_T(\pi) := \mathbb{E}_{\theta^* \sim \mathcal{N}(\mu_0, \Sigma_0^{-1})} \left[R_T\big(\pi; \nu(\theta^*)\big) \right]$.

3 Analysis of Baseline Strategyproof Algorithms and the Generic Lower Bound

In this section, we establish regret upper bounds for existing strategyproof bandit algorithms to highlight a research gap: none of them simultaneously achieves sublinear problem-independent regret and polylogarithmic problem-dependent regret. Next, we derive a generic minimax regret lower bound, characterizing the fundamental cost of strategyproofness under strategic misreporting.

3.1 Regret Bounds for Baseline strategyproof algorithms

We first adapt ETC algorithm to the linear bandit model with private contexts (Komiyama & Imaizumi, 2023), which we continue to refer to as ETC. The algorithm (Langford & Zhang, 2007; Li et al., 2022; Komiyama & Imaizumi, 2023) divides the T rounds into a uniform exploration phase over the first T_0 rounds and a greedy exploitation phase thereafter based on the public estimate $\hat{\theta}_t$. ETC balances exploration and exploitation through the transition time T_0. When the minimum positive optimality gap $\Delta_{\min}$ is known, appropriately tuning T_0 would yield logarithmic problem-dependent regret. However, when $\Delta_{\min}$ is unknown, such problem-dependent tuning is generally unavailable, causing a higher regret. The description of ETC is postponed to Appendix A.1. We establish problem-dependent and problem-independent regret upper bounds for ETC. The proof is deferred to Appendix A.1.

Theorem 1. *In the frequentist model, the ETC policy π^{ETC} satisfies*

$$R_T(\pi^{\mathrm{ETC}}) \in \widetilde{O}\Big((MKd)^{1/3}T^{2/3} + (MK)^{2/3}(T/d)^{1/3}\Big).$$

The Bayesian regret $\mathrm{BR}_T(\pi^{\mathrm{ETC}})$ follows the same order as above. Moreover, assuming that each agent type has a unique true optimal action with respect to θ^*, we have

$$R_T(\pi^{\mathrm{ETC}}; \nu(\theta^*)) \in O\left((MKd)^{1/3}T^{2/3}(\log T)^{1/3} + (MKd)^{2/3}\left(\frac{1}{\Delta_{\min}}+1\right)T^{1/3}(\log T)^{2/3}\right).$$

Although Theorem 1 suggests sublinear problem-independent regret, the bound explicitly depends on the number of actions K, which can greatly exceed the feature dimension d and is undesirable in linear bandits. This dependence stems from ETC's uniform exploration, which assigns probability $1/K$ to every action regardless of the informativeness of each type–action pair.

In contrast, Hu & Duan (2025) propose the Truthful Thompson Sampling (TTS) algorithm when there are two agent types and prove TTS is with an $O(\log T)$ problem-dependent regret bound. TTS adjusts the traditional Thompson sampling (TS) policy at each round through a specially designed LP that guarantees its strategyproofness while maintaining the efficiency of the original TS policy. In this work, we extend the results in Hu & Duan (2025). Specifically, we adapt TTS to the model with $M > 2$ agent types, yielding revised truthful Thompson sampling (R-TTS), which is detailed in Algorithm 2. However, R-TTS incurs linear worst-case regret when an infrequent type m with small arrival probability β_m is with a context direction that cannot be learned from frequent types' observations. We formalize the bounds on R-TTS's regret as follows.

Theorem 2. *The revised truthful Thompson sampling policy $\pi^{\text{R-TTS}}$ satisfies*

$$R_T(\pi^{\text{R-TTS}}) \in \Omega(T),\ \mathrm{BR}_T(\pi^{\text{R-TTS}}) \in \Omega(T).$$

Moreover, assuming that each type has a unique true optimal action with respect to θ^*, and that $\boldsymbol{\beta}$ is known, we have

$$R_T(\pi^{\text{R-TTS}}; \nu) \in O\left(\frac{d^3\log^3(dT)}{\beta_{\min}^2\Delta_{\min}^3}\right). \tag{6}$$

R-TTS's linear problem-independent regret suggests that the standard OFU principle can fail in strategyproof settings. This conclusion is further supported by the fact that the results continue to hold when the TS policy in R-TTS is replaced by a UCB-based policy, which also follows the OFU principle. In summary, ETC achieves sublinear problem-independent regret but cannot efficiently handle easy instances to attain polylogarithmic problem-dependent regret, whereas R-TTS achieves the latter but can incur linear regret on difficult instances.

3.2 REGRET LOWER BOUND FOR STRATEGYPROOF ALGORITHMS

We first quantify the minimum achievable regret for strategyproof algorithms, characterizing the fundamental cost of learning under the strategic environment.

Theorem 3. *Let Π denote the class of all bandit policies satisfying the strategyproofness condition in Definition 1 and let $\mathfrak{V}$ denote the set of all admissible problem instances with $\|\theta^*\| \leq L/2$. Let $M, d > 1$ and $L > 2$. Then, in the frequentist model, if*

$$T > \max\left\{8\sigma^2(K-1)^2,\ \sigma^2(K-1)^2\left(\frac{d}{L^2/4-1}\right)^{3/2},\ 16(K-1)\left(8+\frac{1}{\sigma^2}+\log 4\right)\right\},$$

the minimax regret satisfies

$$\inf_{\pi\in\Pi}\sup_{\nu(\theta^*)\in\mathfrak{V}} R_T(\pi;\nu(\theta^*)) \in \Omega\Big(\min\{M,d\}^{2/3}T^{2/3}\Big).$$

The minimax lower bound in Theorem 3 scales as $T^{2/3}$, in contrast to the order of $\sqrt{T}$ for standard linear bandits (Slivkins et al., 2019), revealing the unavoidable cost induced by strategic misreporting. We say an agent type m is infrequent and hard if it has a small arrival probability β_m and a small minimum positive optimality gap $\min_{a:\Delta_{m,a}>0}\Delta_{m,a}$, and say type m is common and comparatively easy if β_m and $\min_{a:\Delta_{m,a}>0}\Delta_{m,a}$ are comparatively large. To prove Theorem 3, we construct an instance with infrequent hard types and common easy types. The context vectors of different types lie in mutually orthogonal subspaces, so a good understanding of the common types cannot help to identify the optimal actions for the hard types. Since strategyproofness forces the system to sample actions for hard and common types with similar distributions, the sampling rule which can handle the hardness for the hard type will incur additional regret for the common type. The complete proof is provided in Appendix A.3.

This lower bound along with the complementary limitations of existing strategyproof algorithms which has been presented in Section 3.1 motivates us to develop a strategyproof bandit algorithm in the next section that simultaneously achieves both $\widetilde{O}(T^{2/3})$ problem-independent regret and $O(\mathrm{polylog}\,T)$ problem-dependent regret.

4 CONFIDENCE-GATED TRUTHFUL UCB

The results in Section 3 reveal two fundamental challenges of strategyproof learning. First, while strategyproofness couples exploration across types, efficient learning for each agent type must adapt to the type's needs. Second, due to heterogeneous agent arrival frequencies, resolving uncertainty for an infrequent type may force undesirable action choices of common types. As a result, an effective algorithm must balance the favorable action choices across agents under the strategyproofness constraint.

We address these challenges with *Confidence-Gated Truthful UCB* (CGT-UCB), as presented in Algorithm 1. At each round, we construct policy $\pi^{\text{CGT-UCB}}$ through two phases. Phase 1 constructs a proposal policy that selects an exploration or exploitation action for each type, avoiding excessive exploration induced by the standard OFU principle under strategic environment. Phase 2 converts this proposal policy into a strategyproof policy, $\pi^{\text{CGT-UCB}}$ by solving an LP. We further elaborate the algorithms below.

In Phase 1, at each time step t, CGT-UCB first computes the estimated-optimal action $\hat{a}_m^t$ and the exploration action $b_{m,t}$ for each agent $m \in \mathcal{M}$. It then computes $U_{m,t}$ as defined in (7), which serves as an upper bound on the instantaneous regret of $\hat{a}_m^t$. If $U_{m,t}$ passes the confidence gate tailored to the agent's type, that is, $U_{m,t} < \varepsilon_t/\sqrt{\beta_m}$, CGT-UCB selects $\hat{a}_m^t$ as the proposal action for agent m; otherwise, it selects $b_{m,t}$.

In Phase 2, CGT-UCB converts each proposal action from Phase 1 into a one-hot distribution p_m^t for all $m \in \mathcal{M}$ and passes them to the LP in (8)–(11), whose solution defines the policy $\pi^{\text{CGT-UCB}}$. The LP minimizes the $\boldsymbol{\beta}$-weighted distance between $\pi^{\text{CGT-UCB}}$ and p^t subject to the following constraints: the constraint in (9) ensures the strategyproofness of $\pi^{\text{CGT-UCB}}$ in every round, while the retention constraint in (10) ensures that $\pi^{\text{CGT-UCB}}$ retains enough exploration compared to the proposal distribution. Moreover, CGT-UCB operates in both frequentist and Bayesian settings by

Algorithm 1 Confidence-Gated Truthful UCB (CGT-UCB)

Require: Type probabilities $(\beta_m)_{m\in\mathcal{M}}$; feature vectors $\{\phi(m,a) : m \in \mathcal{M}, a \in \mathcal{K}\}$, where $\Phi_i \in \mathbb{R}^{K\times d}$ has a-th row $\phi(i,a)^\top$ for each $i \in \mathcal{M}$; regularization parameter $\lambda > 0$; confidence radii $(\rho_t)_{t\geq 1}$; resolution schedule $(\varepsilon_t)_{t\geq 1}$.

1: Initialize $(\hat{\theta}_1, W_1) \leftarrow (0_d, \lambda I_d)$ (frequentist) or $(\hat{\theta}_1, W_1) \leftarrow (\mu_0, \Sigma_0)$ (Bayesian).
2: **for** $t = 1, \ldots, T$ **do**
Phase 1: Construct proposal policy
3: **for** each type $m \in \mathcal{M}$ **do**
4: Compute the estimated-optimal action $\hat{a}^t_m \in \arg\max_{a\in\mathcal{K}} \left\langle \hat{\theta}_t, \phi(m,a) \right\rangle$.
5: Compute the exploration action

$$b_{m,t} \in \arg\max_{a\in\mathcal{K}} \left\{ \left\langle \hat{\theta}_t, \phi(m,a) \right\rangle + \rho_t \|\phi(m,a)\|_{W_t^{-1}} \right\}.$$

6: Compute the greedy simple-regret upper bound

$$U_{m,t} \leftarrow \left\langle \hat{\theta}_t, \phi(m,b_{m,t}) - \phi(m,\hat{a}^t_m) \right\rangle + \rho_t \left(\|\phi(m,b_{m,t})\|_{W_t^{-1}} + \|\phi(m,\hat{a}^t_m)\|_{W_t^{-1}} \right). \quad (7)$$

7: **if** $U_{m,t} \geq \varepsilon_t/\sqrt{\beta_m}$ **then**
8: Set the proposal action to the exploration action: $a^{\text{prop}}_{m,t} \leftarrow b_{m,t}$.
9: **else**
10: Set the proposal action to the estimated-optimal action: $a^{\text{prop}}_{m,t} \leftarrow \hat{a}^t_m$.
11: **end if**
12: **end for**
Phase 2: Construct CGT-UCB policy
13: Define the deterministic proposal policy $p^t_m \leftarrow e_{a^{\text{prop}}_{m,t}} \in \Delta_K$. Define $\mathcal{M}_t := \{m \in \mathcal{M} : U_{m,t} \geq \varepsilon_t/\sqrt{\beta_m}\}$. Compute $q^t = (q^t_m)_{m\in\mathcal{M}}$ by solving

$$\underset{(q^t_m)_{m\in\mathcal{M}}}{\text{minimize}} \quad \sum_{m\in\mathcal{M}} \beta_m \|q^t_m - p^t_m\|_1 \quad (8)$$

$$\text{subject to} \quad (q^t_i - q^t_j)^\top \Phi_i \hat{\theta}_t \geq 0, \qquad \forall i,j \in \mathcal{M}, \quad (9)$$

$$q^t_m(b_{m,t}) \geq \frac{1}{|\mathcal{M}_t|} p^t_m(b_{m,t}), \qquad \forall m \in \mathcal{M}_t, \quad (10)$$

$$q^t_m \in \Delta_K, \qquad \forall m \in \mathcal{M}. \quad (11)$$

Omit (10) if $\mathcal{M}_t = \varnothing$.
14: $\pi^{\text{CGT-UCB}}(m|\mathcal{F}_t) \leftarrow q^t_m$, for all $m \in \mathcal{M}$.
15: Observe the reported type $I'_t \in \mathcal{M}$, and sample $A_t \sim \pi^{\text{CGT-UCB}}(I'_t|\mathcal{F}_t)$.
16: Observe the reward r_t, and update $\hat{\theta}_{t+1}$ and W_{t+1} using (1) (frequentist) or (2) (Bayesian).
17: **end for**

executing different parameter initialization and parameter update procedures in lines 1 and 17 and setting different confidence radii ρ_t input for the two settings. In summary, the proposal policy in Phase 1 controls when exploration is proposed, while the constraint in Phase 2 ensures sufficient exploration survives the LP adjustment.

To ensure the effectiveness of Algorithm 1, it is essential to show that the LP in (8)–(11) is feasible, as established in the following lemma. The proof is deferred to Appendix B.

Lemma 1. *The LP in* (8)–(11) *is feasible for every time* t *and every history* $\mathcal{F}_t$.

Moreover, we establish the problem-independent frequentist and Bayesian regret bounds of our Algorithm 1, with proof deferred to Appendix B.1.

Theorem 4. *In Algorithm 1, let*

$$\rho_t = \alpha_t := \sigma\sqrt{2\log T + d\log\left(1 + \frac{t}{\lambda d}\right)} + \sqrt{\lambda}L, \ \varepsilon_t = \left(\frac{8dL\alpha_T^2 \max\{1,\lambda^{-1}\}\log\left(1+\frac{T}{\lambda d}\right)}{t\sum_{m=1}^M \sqrt{\beta_m}} \right)^{1/3} \quad (12)$$

for the frequentist setting, and let

$$\rho_t = \alpha'_T := \sqrt{d + 2\sqrt{2d\log T} + 4\log T}, \tag{13}$$

$$\varepsilon_t = \left(\frac{8(\alpha'_T)^2 d \max\{\sigma^2, \lambda_{\max}(\Sigma_0^{-1})\} \log\left(\frac{\mathrm{tr}(\Sigma_0)+T/\sigma^2}{d\ \det(\Sigma_0)^{1/d}}\right) \mathbb{E}_{\theta^*}[\Delta_{\max}]}{t \sum_{m=1}^{M} \sqrt{\beta_m}} \right)^{1/3} \tag{14}$$

for the Bayesian setting. Then the policy $\pi^{\text{CGT-UCB}}$ *is strategyproof and satisfies*

$$R_T(\pi^{\text{CGT-UCB}}) \leq 5 \left(dL\alpha_T^2 \max\{1, \lambda^{-1}\} \log\left(1 + \frac{T}{\lambda d}\right) \right)^{1/3} T^{2/3} M^{1/3} + L, \tag{15}$$

$$\begin{aligned} \mathrm{BR}_T(\pi^{\text{CGT-UCB}}) \leq & 5 \left((\alpha'_T)^2 d \max\{\sigma^2, \lambda_{\max}(\Sigma_0^{-1})\} \log\left(\frac{\mathrm{tr}(\Sigma_0)+T/\sigma^2}{d\ \det(\Sigma_0)^{1/d}}\right) \mathbb{E}_{\theta^*}[\Delta_{\max}] \right)^{1/3} \\ & \times T^{2/3} M^{1/3} + \sqrt{\mathbb{E}[\Delta_{\max}^2]}. \end{aligned} \tag{16}$$

In particular,

$$R_T(\pi^{\text{CGT-UCB}}) \in \widetilde{O}(d^{2/3} M^{1/3} T^{2/3}),\ \mathrm{BR}_T(\pi^{\text{CGT-UCB}}) \in \widetilde{O}(d^{2/3} M^{1/3} T^{2/3}). \tag{17}$$

In Theorem 4, CGT-UCB leverages $\boldsymbol{\beta}$ to establish the threshold for the confidence gate. Together with the retention constraint (10), this mitigates the regret's dependence on M and d. When $\boldsymbol{\beta}$ is unknown, the algorithm employs an alternative threshold, which incurs a higher regret bound, as detailed in Appendix B.3.

Compared with the ETC bound in Theorem 1, the bound for CGT-UCB in Theorem 4 does not explicitly depend on K, suggesting that CGT-UCB outperforms ETC in the regime where $K \gg d$. Furthermore, the CGT-UCB upper bound matches the general lower bound in Theorem 3 in terms of its dependence on T, up to logarithmic factors. The gap in M arises from its influence through the type distribution $\boldsymbol{\beta}$, whereas the gap in d is a well-known artifact in linear bandits (Lattimore & Szepesvári, 2020). Closing these gaps remains challenging, and resolving this tension presents an intriguing direction for future work.

We also establish the problem-dependent regret upper bound for CGT-UCB as follows. The proof is provided in Appendix B.

Theorem 5. *In the frequentist setting, we again set* ρ_t, α_t, *and* ε_t *as in* (12). *Suppose each agent type has a unique optimal action and* $\Delta_{\min} > 0$. *Then* $\pi^{\text{CGT-UCB}}$ *satisfies*

$$R_T(\pi^{\text{CGT-UCB}}; \nu) \leq \frac{20L + 8LM}{\beta_{\min}\Delta_{\min}^2} \alpha_T^2 \max\{1, \lambda^{-1}\} d \log\left(1 + \frac{T}{\lambda d}\right) + L \in O\left(\frac{Md^2 \log^2 T}{\beta_{\min}\Delta_{\min}^2}\right). \tag{18}$$

Compared to the baseline R-TTS bound in Theorem 2, the CGT-UCB bound in Theorem 5 improves the dimension dependence from $O(d^3 \log^3(dT))$ to $O(Md^2 \log^2 T)$, and the instance parameter dependence from $O(\beta_{\min}^{-2}\Delta_{\min}^{-3})$ to $O(\beta_{\min}^{-1}\Delta_{\min}^{-2})$. Consequently, for fixed instance parameters, CGT-UCB yields a strictly superior dependence on d whenever $M < d$. The additional factor M in the CGT-UCB bound originates from the exploration retention constraint in (10), which is necessary for CGT-UCB to achieve sublinear problem-independent regret. This highlights a fundamental design tradeoff required to simultaneously secure sublinear problem-independent regret and logarithmic problem-dependent regret.

5 Experiments

In this section, we evaluate CGT-UCB on the synthetic instances with varying dimensions, the instance used in our lower-bound construction, and a real-world Warfarin dose-selection task with binary feedback. We compare CGT-UCB with several strategyproof baselines, including ETC Komiyama & Imaizumi (2023), truthful Thompson sampling (TTS) Hu & Duan (2025), and the revised truthful Thompson sampling algorithm (R-TTS) described in Algorithm 2. For all instances, we run each algorithm for $T = 6000$ over 1000 independent trials. We report the mean results together with 95% confidence intervals. Complete experimental details are provided in Appendix D.

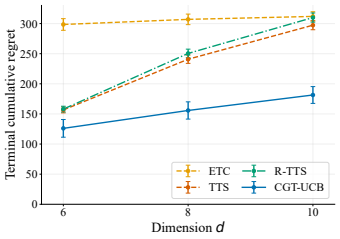

Figure 2: Terminal regret versus dimension d. Terminal values report averages over seeds, while the error bars indicate pointwise 95% confidence intervals.

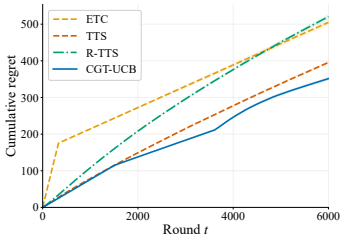

Figure 3: Empirical regret on the worst-case instance. Curves report averages over seeds, while the shaded regions indicate pointwise 95% confidence intervals.

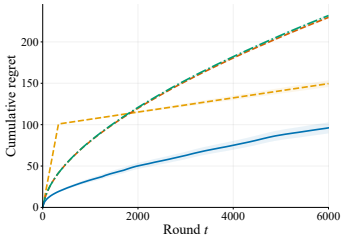

Figure 4: Empirical regret on Warfarin dose selection. Curves report averages over seeds, while the shaded regions indicate pointwise 95% confidence intervals.

Regret dependence on d. We vary the feature dimension over $d \in \{6, 8, 10\}$ while fixing $M = K = 5$. We use a challenging synthetic instance in which the type-action feature vectors occupy different dimensions. As shown in Figure 2, CGT-UCB achieves the lowest mean terminal regret at all three dimensions, with values of 31.90, 162.25, and 175.70, respectively. Regret increases with d for all algorithms, albeit at different rates, reflecting the greater difficulty of learning in higher dimensions. CGT-UCB is less sensitive to increases in d than the competing methods.

$T^{2/3}$ dependence of problem-independent regret. We instantiate the worst-case problem instance used to establish the problem-independent lower bound in Theorem 3 and Appendix A.3, setting $d = M = K = 10$. The results are shown in Figure 3. CGT-UCB achieves the lowest mean terminal regret of 354.28, compared with 403.22 for TTS, 486.09 for ETC, and 517.54 for R-TTS. Moreover, the confidence band for CGT-UCB lies uniformly below those of the competing methods. On this challenging instance, CGT-UCB accumulates less regret than the baselines in later rounds, demonstrating the effectiveness of confidence-gated exploration under strategyproofness constraints. Its regret curve appears to scale as $T^{2/3}$ over the observed horizon, consistent with Theorems 3 and 4.

Practical efficacy. We evaluate the algorithms on a real-world Warfarin dose-selection task using the Warfarin dataset (Consortium, 2009). Patients are clustered into eight types according to age, height, weight, and VKORC1 genotype, yielding $M = 8$, $K = 3$, and $d = 15$. The task is to select one of three weekly dose classes: high (above 49 mg/week), medium (21–49 mg/week), or low (below 21 mg/week). The reward is one if the selected dose class matches the patient's recorded class and zero otherwise. As shown in Figure 4, CGT-UCB achieves a substantially lower mean terminal regret of 85.58, compared with 145.36 for ETC, 227.95 for TTS, and 230.95 for R-TTS. These results demonstrate the practical effectiveness of CGT-UCB in a real-world setting.

6 Conclusion

We studied strategyproof algorithms for linear contextual bandits with private contexts. We established complementary limitations of existing strategyproof algorithms and a generic regret lower bound of $\Omega(T^{2/3})$. We proposed CGT-UCB, which combines confidence-gated exploration with an LP that enforces strategyproofness. CGT-UCB simultaneously achieves $\widetilde{O}(T^{2/3})$ problem-independent regret and $O(\mathrm{polylog}\, T)$ problem-dependent regret, matching the minimax dependence on T up to logarithmic factors. Experiments on synthetic and real-world datasets demonstrate its advantages over competing algorithms and are consistent with the theoretically predicted dependence on T and d.

Several questions remain open. In particular, closing the gap between the lower and upper bounds in their dependence on M and d remains a challenge. Another open direction is to study strategyproof algorithms when the type-action context vectors lie in a compact set with infinitely many agent types, rather than in the discrete setting modeled in this work. We leave these questions for future work.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript, search for related work, and assist with running experiments. We did not use generative AI tools to develop the ideas. The remaining disclosure categories are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide implementation details in Appendix D. All assumptions and complete proofs of our theoretical results are included in the corresponding appendices. We will provide the complete experimental code upon acceptance.

REFERENCES

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

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

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

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

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. 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.

International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.

U.S. Food and Drug Administration. U.S. Food and Drug Administration. Technical report, U.S. Food and Drug Administration, 2017. URL https://www.accessdata.fda.gov/drugsatfda_docs/nda/2017/208716Orig1s000TOC.cfm. Accessed: 2024-09-04.

Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pp. 3536–3545. PMLR, 2020.

Keegan Harris, Chara Podimata, and Steven Wu. Strategy-aware contextual bandits. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

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

Yiting Hu and Lingjie Duan. Truthful mechanisms for linear bandit games with private contexts. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1006–1014. International Foundation for Autonomous Agents and Multiagent Systems, 2025. URL https://dl.acm.org/doi/10.5555/3709347.3743620.

Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Incentivizing exploration with selective data disclosure. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 647–648, 2020.

Thomas Kleine Buening, Aadirupa Saha, Christos Dimitrakakis, and Haifeng Xu. Strategic linear contextual bandits. *Advances in Neural Information Processing Systems*, 37:116638–116675, 2024.

Junpei Komiyama and Masaaki Imaizumi. High-dimensional contextual bandit problem without sparsity. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49416–49427. Curran Associates, Inc., 2023. doi: 10.52202/075280-2150. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9b35a0a20d617dc68ae98a7a57df2f51-Paper-Conference.pdf.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.

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

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

Wenjie Li, Adarsh Barik, and Jean Honorio. A simple unified framework for high dimensional bandit problems. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12619–12655. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22a.html.

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

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

Manhin Poon, Xiangxiang Dai, Xutong Liu, Fang Kong, John C. S. Lui, and Jinhang Zuo. Online multi-llm selection via contextual bandits under unstructured context evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24855–24863, 2026. doi: 10.1609/aaai.v40i29.39672.

Rifaquat Rahman, Lorenzo Trippa, Eudocia Q Lee, Isabel Arrillaga-Romany, Geoffrey Fell, Mehdi Touat, Christine McCluskey, Jennifer Wiley, Sarah Gaffey, Jan Drappatz, et al. 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):5524–5535, 2023.

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

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

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

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

A Missing Details for Section 3

A.1 Details of the ETC Algorithm

In ETC, the policy π^{ETC} operates as follows.

$$\pi_a^{\mathrm{ETC}}(m \mid \mathcal{F}_t) = \begin{cases} \frac{1}{K}, & \text{if } t \le T_0, \\ \mathbf{1}\{a = \hat{a}_m^t\}, & \text{if } t > T_0. \end{cases}, \text{ where } \hat{a}_m^t \in \arg\max_{a \in \mathcal{K}} \langle \hat{\theta}_t, \phi(m,a) \rangle. \tag{19}$$

We present the complete version of Theorem 1, providing problem-independent and problem-dependent regret bounds, respectively, in the following two theorems.

Theorem 6 (Complete version of problem-independent regret of Theorem 1)**.** *π^{ETC} is strategyproof. Define $\Delta_{\max} := \max_{m,a} \Delta_{m,a}$. The frequentist regret of π^{ETC} satisfies*

$$\begin{aligned} R_T(\pi^{\mathrm{ETC}}) &\le 4\alpha_T^{2/3}(MKL)^{1/3}T^{2/3} + 2L \\ &\quad + \frac{4(KM)^{2/3}L^{5/3}T^{1/3}}{\alpha_T^{2/3}} \left[1 + \log\left(1 + \frac{d}{8K}\left(\frac{(2\alpha_T T)^{2/3}(2MK)^{1/3}}{L^{2/3}} + 1\right)\right)\right] \end{aligned} \tag{20}$$

where

$$\alpha_t := \sigma\sqrt{2\log T + d\log\left(1 + \frac{t}{\lambda d}\right)} + \sqrt{\lambda}L,$$

for

$$T_0 = \min\left\{T, \max\left\{1, \left\lceil \frac{(2\alpha_T T)^{2/3}(2MK)^{1/3}}{L^{2/3}} \right\rceil\right\}\right\}.$$

Moreover, the Bayesian regret of π^{ETC} satisfies

$$\begin{aligned} &\mathrm{BR}_T(\pi^{\mathrm{ETC}}) \\ &\le 4(\sigma\alpha'_T)^{2/3}(MK)^{1/3}\mathbb{E}_{\theta^*}[\Delta_{\max}]^{1/3}T^{2/3} + \mathbb{E}_{\theta^*}[\Delta_{\max}] + \sqrt{\mathbb{E}_{\theta^*}[\Delta_{\max}^2]} \\ &+ \frac{4(KM)^{2/3}\mathbb{E}_{\theta^*}[\Delta_{\max}]^{5/3}T^{1/3}}{(\sigma\alpha'_T)^{2/3}} \left[1 + \log\left(1 + \frac{d}{8K}\left(\frac{(2\sigma\alpha'_T T)^{2/3}(2MK)^{1/3}}{\mathbb{E}_{\theta^*}[\Delta_{\max}]^{2/3}} + 1\right)\right)\right], \end{aligned} \tag{21}$$

where

$$\alpha'_T := \sqrt{d + 2\sqrt{2d\log T} + 4\log T},$$

for

$$T_0 = \min\left\{T, \max\left\{1, \left\lceil \frac{(2\sigma\alpha'_T T)^{2/3}(2MK)^{1/3}}{\mathbb{E}_{\theta^*}[\Delta_{\max}]^{2/3}} \right\rceil\right\}\right\}.$$

If $\mathbb{E}_{\theta^*}[\Delta_{\max}] = 0$, take $T_0 = 1$; then the Bayesian regret is zero. Moreover,

$$R_T(\pi^{\mathrm{ETC}}) \in \widetilde{O}((MKd)^{1/3}T^{2/3} + (MK)^{2/3}(T/d)^{1/3}) \tag{22}$$

and

$$\mathrm{BR}_T(\pi^{\mathrm{ETC}}) \in \widetilde{O}((MKd)^{1/3}T^{2/3} + (MK)^{2/3}(T/d)^{1/3}). \tag{23}$$

Theorem 7 (Complete version of problem-dependent regret of Theorem 1)**.** *For any fixed instance $\nu(\theta^*)$ with unique true optimal action for all agent types with respect to θ^*, and for*

$$T_0 = \min\left\{T, \max\left\{1, \left\lceil \frac{(2\alpha_T T)^{2/3}(2MK)^{1/3}}{L^{2/3}} \right\rceil\right\}\right\},$$

ETC gives the following problem-dependent regret bound

$$\begin{aligned} &R_T(\pi^{\mathrm{ETC}}; \nu(\theta^*)) \\ &\le \frac{2\Delta_{\max}\alpha_T^{2/3}(MK)^{1/3}}{L^{2/3}}T^{2/3} + \frac{4(KM)^{2/3}L^{2/3}\alpha_T^{4/3}}{\Delta_{\min}}T^{1/3} \\ &\quad + \frac{4(KM)^{2/3}L^{2/3}\Delta_{\max}}{\alpha_T^{2/3}}T^{1/3}\left[1 + \log\left(1 + \frac{d}{8K}\left(\frac{(2\alpha_T T)^{2/3}(2MK)^{1/3}}{L^{2/3}} + 1\right)\right)\right] \\ &\quad + 2\Delta_{\max}. \end{aligned} \tag{24}$$

Moreover,

$$R_T(\pi^{\mathrm{ETC}}; \nu(\theta^*)) \in O((MKd)^{1/3} T^{2/3} (\log T)^{1/3}).$$

Proof of Theorem 6. First, the policy π^{ETC} in (19) is independent of the reported type for $t \le T_0$ and is greedy with respect to the current public estimate $\hat{\theta}_t$ for $t > T_0$. Therefore, it satisfies Definition 1 and is strategyproof.

We bound the problem-independent frequentist and Bayesian regret separately, beginning with the frequentist regret.

Frequentist regret. For each $t \in [T]$, define

$$\mathcal{E}_t := \left\{ \left| \langle u, \hat{\theta}_t - \theta^* \rangle \right| \le \alpha_t \|u\|_{V_t^{-1}} \text{ for every } u \in \mathbb{R}^d \right\}.$$

Since $\left| \langle u, \hat{\theta}_t - \theta^* \rangle \right| \le \|\hat{\theta}_t - \theta^*\|_{V_t} \|u\|_{V_t^{-1}}$ always holds and Theorem 2 in Abbasi-yadkori et al. (2011) guarantees that $\|\hat{\theta}_t - \theta^*\|_{V_t} \le \alpha_t$ for all t with probability at least $1 - 1/T$, we have

$$\Pr(\mathcal{E}_t^c) \le \frac{1}{T}.$$

For each type $m \in \mathcal{M}$ and exploration round $s \le T_0$, define

$$G_m := \mathbb{E}\left[\mathbf{1}\{I_s = m\} \phi(m, A_s) \phi(m, A_s)^\top \right] = \frac{\beta_m}{K} \sum_{a \in \mathcal{K}} \phi(m, a) \phi(m, a)^\top,$$

where the second equality follows because the exploration policy selects each action uniformly.

Also define the empirical type-specific covariance matrix

$$V_{T_0+1}^m := \sum_{s=1}^{T_0} \mathbf{1}\{I_s = m\} \phi(m, A_s) \phi(m, A_s)^\top.$$

Then

$$\max_{a \in \mathcal{K}} \|\phi(m, a)\|_{G_m^+}^2 \le \frac{K}{\beta_m}.$$

For $G_m \neq 0$, applying Lemma 5 with $\delta = 1/2$ yields

$$\begin{aligned} \Pr\left(V_{T_0+1}^m \not\succeq \frac{T_0}{2} G_m \right) &\le d \exp\left(- \frac{T_0}{8 \max_{a \in \mathcal{K}} \|\phi(m, a)\|_{G_m^+}^2} \right) \\ &\le d \exp\left(- \frac{\beta_m T_0}{8K} \right). \end{aligned} \tag{25}$$

If $G_m = 0$, all features of type m are zero and the failure probability above is zero; the bound therefore also holds in this case.

For a threshold $b > 0$, partition the type set into

$$\mathcal{M}_b^+ := \{m \in \mathcal{M} : \beta_m \ge b\}, \qquad \mathcal{M}_b^- := \{m \in \mathcal{M} : \beta_m < b\}.$$

Define the event

$$\mathcal{E}_V := \bigcap_{m \in \mathcal{M}_b^+} \left\{ V_{T_0+1}^m \succeq \frac{T_0}{2} G_m \right\}.$$

By a union bound and (25),

$$\Pr(\mathcal{E}_V^c) \le \sum_{m \in \mathcal{M}_b^+} d \exp\left(- \frac{\beta_m T_0}{8K} \right) \le dM \exp\left(- \frac{b T_0}{8K} \right). \tag{26}$$

Regret on the good event. Consider any $m \in \mathcal{M}_b^+$. On $\mathcal{E}_V$, after time T_0,

$$V_t \succeq V_{T_0+1} \succeq V_{T_0+1}^m \succeq \frac{T_0}{2} G_m.$$

Since every $\phi(m,a)$ belongs to $\mathrm{range}(G_m)$, it follows that

$$\|\phi(m,a)\|^2_{V_t^{-1}} \le \frac{2}{T_0}\|\phi(m,a)\|^2_{G_m^+} \le \frac{2K}{T_0\beta_m}. \tag{27}$$

After T_0, because $\hat{a}_m^t$ is greedy with respect to $\hat{\theta}_t$,

$$\langle \phi(m,a_m^*) - \phi(m,\hat{a}_m^t), \hat{\theta}_t\rangle \le 0.$$

Thus, on $\mathcal{E}_t \cap \mathcal{E}_V$, using $\alpha_t \le \alpha_T$,

$$\begin{aligned}
\Delta_{m,\hat{a}_m^t} &= \langle \phi(m,a_m^*) - \phi(m,\hat{a}_m^t), \theta^*\rangle \\
&= \langle \phi(m,a_m^*) - \phi(m,\hat{a}_m^t), \theta^* - \hat{\theta}_t\rangle + \langle \phi(m,a_m^*) - \phi(m,\hat{a}_m^t), \hat{\theta}_t\rangle \\
&\le \alpha_T \|\phi(m,a_m^*) - \phi(m,\hat{a}_m^t)\|_{V_t^{-1}} \\
&\le \alpha_T \left(\|\phi(m,a_m^*)\|_{V_t^{-1}} + \|\phi(m,\hat{a}_m^t)\|_{V_t^{-1}}\right) \\
&\le 2\alpha_T \sqrt{\frac{2K}{T_0\beta_m}}.
\end{aligned} \tag{28}$$

It follows that the expected one-round regret contributed by the types in $\mathcal{M}_b^+$, on the good event, is at most

$$\sum_{m\in\mathcal{M}_b^+} \beta_m \Delta_{m,\hat{a}_m^t} \le 2\alpha_T \sqrt{\frac{2K}{T_0}} \sum_{m\in\mathcal{M}_b^+} \sqrt{\beta_m} \le 2\alpha_T \sqrt{\frac{2MK}{T_0}}. \tag{29}$$

For the types in $\mathcal{M}_b^-$, we use the worst-case bound $\Delta_{m,a} \le \Delta_{\max}$, and

$$\sum_{t=1}^{T} \sum_{m\in\mathcal{M}_b^-} \beta_m \Delta_{m,a} \le b|\mathcal{M}_b^-|\Delta_{\max} T \le bM\Delta_{\max} T.$$

For $t > T_0$, I_t is independent of the pre-round history and the events $\mathcal{E}_t, \mathcal{E}_V$. Combining the preceding results yields, for every fixed instance, the following regret upper bound:

$$\begin{aligned}
R_T(\pi^{\mathrm{ETC}};\nu) &= \mathbb{E}\left[\sum_{t=1}^{T} \left(\Delta_{I_t,A_t}\mathbf{1}\{I_t \in \mathcal{M}_b^+\} + \Delta_{I_t,A_t}\mathbf{1}\{I_t \in \mathcal{M}_b^-\}\right)\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} \left(\Delta_{I_t,A_t}\mathbf{1}\{\mathcal{E}_t,\, I_t \in \mathcal{M}_b^+\} + \Delta_{I_t,A_t}\mathbf{1}\{\mathcal{E}_t^c,\, I_t \in \mathcal{M}_b^+\}\right)\right] + bMT\Delta_{\max} \\
&\le \mathbb{E}\Bigg[\sum_{t=1}^{T} \left(\Delta_{I_t,A_t}\mathbf{1}\{\mathcal{E}_t,\, t\le T_0,\, I_t \in \mathcal{M}_b^+\} + \Delta_{I_t,A_t}\mathbf{1}\{\mathcal{E}_t,\, t > T_0,\, I_t \in \mathcal{M}_b^+\}\right)\Bigg] \\
&\quad + \Delta_{\max} + bMT\Delta_{\max} \\
&\le T_0\Delta_{\max} + \sum_{t=T_0+1}^{T} \mathbb{E}\left[\sum_{m\in\mathcal{M}_b^+} \beta_m \Delta_{m,\hat{a}_m^t}\mathbf{1}\{\mathcal{E}_t \cap \mathcal{E}_V\}\right] \\
&\quad + \sum_{t=T_0+1}^{T} \mathbb{E}\left[\sum_{m\in\mathcal{M}_b^+} \beta_m \Delta_{m,\hat{a}_m^t}\mathbf{1}\{\mathcal{E}_t \cap \mathcal{E}_V^c\}\right] + \Delta_{\max} + bMT\Delta_{\max} \\
&\le T_0\Delta_{\max} + 2\alpha_T T\sqrt{\frac{2MK}{T_0}} + dM\exp\left(-\frac{bT_0}{8K}\right)T\Delta_{\max} + \Delta_{\max} + bMT\Delta_{\max}.
\end{aligned} \tag{30}$$

Then, choose

$$b = \frac{8K}{T_0}\log\left(1 + \frac{dT_0}{8K}\right).$$

Then

$$d \exp\left(-\frac{bT_0}{8K}\right) \le \frac{8K}{T_0},$$

and the preceding regret bound gives

$$R_T(\pi^{\mathrm{ETC}}; \nu) \le T_0 \Delta_{\max} + 2\alpha_T T \sqrt{\frac{2MK}{T_0}} + \frac{8KMT\Delta_{\max}}{T_0}\left[1 + \log\left(1 + \frac{dT_0}{8K}\right)\right] + \Delta_{\max}. \tag{31}$$

In the frequentist setting, $\Delta_{\max} \le 2\|\theta^*\|_2 \le L$. Choose the integer-valued exploration length

$$T_0 = \min\left\{T, \max\left\{1, \left\lceil \frac{(2\alpha_T T)^{2/3}(2MK)^{1/3}}{L^{2/3}} \right\rceil\right\}\right\}.$$

Let $x = (2\alpha_T T)^{2/3}(2MK)^{1/3}/L^{2/3}$. If $x \ge T$, then $T_0 = T$ and $R_T \le LT \le Lx$, which is already bounded by the first term in (20). Otherwise, $x \le T_0 \le x + 1$. Using these inequalities, in particular $T_0 \le \frac{(2\alpha_T T)^{2/3}(2MK)^{1/3}}{L^{2/3}} + 1$, we obtain

$$R_T(\pi^{\mathrm{ETC}}; \nu) \le 4\alpha_T^{2/3}(MK)^{1/3}L^{1/3}T^{2/3} + 2L + \frac{4(KM)^{2/3}L^{5/3}T^{1/3}}{\alpha_T^{2/3}}\left[1 + \log\left(1 + \frac{d}{8K}\left(\frac{(2\alpha_T T)^{2/3}(2MK)^{1/3}}{L^{2/3}} + 1\right)\right)\right], \tag{32}$$

The bound is uniform over admissible instances, which proves (20).

Bayesian regret. We next consider the Bayesian setting. Under the Bayesian model, conditional on the observation history $\mathcal{F}_t$, we have for every fixed $u \in \mathbb{R}^d$,

$$\left\langle u, \theta^* - \hat{\theta}_t \right\rangle \mid \mathcal{F}_t \sim \mathcal{N}\left(0, u^\top \Sigma_t^{-1} u\right).$$

For each $t \in [T]$, define the good event

$$\mathcal{E}_t := \left\{ \left| \langle u, \hat{\theta}_t - \theta^* \rangle \right| \le \alpha'_T \|u\|_{\Sigma_t^{-1}} \text{ for every } u \in \mathbb{R}^d \right\}.$$

By Lemma 6, choosing

$$\alpha'_T = \sqrt{d + 2\sqrt{2d \log T} + 4 \log T},$$

we have

$$\Pr\left(\mathcal{E}_t^c \mid \mathcal{F}_t\right) \le \frac{1}{T^2}.$$

Taking expectations with respect to $\mathcal{F}_t$ yields $\Pr(\mathcal{E}_t^c) \le \frac{1}{T^2}$. Therefore, by the Cauchy–Schwarz inequality,

$$\mathbb{E}\left[\Delta_{\max} \mathbf{1}\{\mathcal{E}_t^c\}\right] \le \sqrt{\mathbb{E}_{\theta^*}[\Delta_{\max}^2]}\sqrt{\Pr(\mathcal{E}_t^c)} \le \frac{1}{T}\sqrt{\mathbb{E}_{\theta^*}[\Delta_{\max}^2]}. \tag{33}$$

Moreover, $\mathcal{E}_V$ depends only on the type arrivals and the uniformly sampled exploration actions and is therefore independent of θ^*. For $t > T_0$, on $\mathcal{E}_V$, we have

$$\Sigma_t \succeq \sigma^{-2} V_{T_0+1}^m \succeq \frac{T_0}{2\sigma^2} G_m, \qquad m \in \mathcal{M}_b^+.$$

Consequently, $\|\phi(m,a)\|_{\Sigma_t^{-1}}^2 \le 2\sigma^2 K/(T_0 \beta_m)$. The same greedy-action argument in the frequentist model bounds the common-type one-round regret on $\mathcal{E}_t \cap \mathcal{E}_V$ by $2\sigma\alpha'_T\sqrt{2MK/T_0}$. Repeating

the decomposition in (30), now averaging over the prior and applying (33) at each round, gives

$$\begin{aligned}
&\mathrm{BR}_T(\pi^{\mathrm{ETC}}) \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(\Delta_{I_t,A_t}\mathbf{1}\{I_t \in \mathcal{M}_b^+\} + \Delta_{I_t,A_t}\mathbf{1}\{I_t \in \mathcal{M}_b^-\}\right)\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}\left(\Delta_{I_t,A_t}\mathbf{1}\{\mathcal{E}_t,\, I_t \in \mathcal{M}_b^+\} + \Delta_{I_t,A_t}\mathbf{1}\{\mathcal{E}_t^c,\, I_t \in \mathcal{M}_b^+\}\right)\right] + bMT\mathbb{E}_{\theta^*}[\Delta_{\max}] \\
&\le \mathbb{E}\Bigg[\sum_{t=1}^{T}\left(\Delta_{I_t,A_t}\mathbf{1}\{\mathcal{E}_t,\, t \le T_0,\, I_t \in \mathcal{M}_b^+\} + \Delta_{I_t,A_t}\mathbf{1}\{\mathcal{E}_t,\, t > T_0,\, I_t \in \mathcal{M}_b^+\}\right)\Bigg] \\
&\quad + \sum_{t=1}^{T}\mathbb{E}[\Delta_{\max}\mathbf{1}\{\mathcal{E}_t^c\}] + bMT\mathbb{E}_{\theta^*}[\Delta_{\max}] \\
&\le T_0\mathbb{E}_{\theta^*}[\Delta_{\max}] + \sum_{t=T_0+1}^{T}\mathbb{E}\left[\sum_{m\in\mathcal{M}_b^+}\beta_m\Delta_{m,\hat{a}_m^t}\mathbf{1}\{\mathcal{E}_t \cap \mathcal{E}_V\}\right] \\
&\quad + \sum_{t=T_0+1}^{T}\mathbb{E}\left[\sum_{m\in\mathcal{M}_b^+}\beta_m\Delta_{m,\hat{a}_m^t}\mathbf{1}\{\mathcal{E}_t \cap \mathcal{E}_V^c\}\right] + \sqrt{\mathbb{E}_{\theta^*}[\Delta_{\max}^2]} + bMT\mathbb{E}_{\theta^*}[\Delta_{\max}] \\
&\le T_0\mathbb{E}_{\theta^*}[\Delta_{\max}] + 2\sigma\alpha_T' T\sqrt{\frac{2MK}{T_0}} + dM\exp\left(-\frac{bT_0}{8K}\right)T\mathbb{E}_{\theta^*}[\Delta_{\max}] \\
&\quad + \sqrt{\mathbb{E}_{\theta^*}[\Delta_{\max}^2]} + bMT\mathbb{E}_{\theta^*}[\Delta_{\max}]. \qquad (34)
\end{aligned}$$

Thus, choosing

$$b = \frac{8K}{T_0}\log\left(1 + \frac{dT_0}{8K}\right)$$

and

$$T_0 = \min\left\{T,\, \max\left\{1,\, \left\lceil\frac{(2\sigma\alpha_T' T)^{2/3}(2MK)^{1/3}}{\mathbb{E}_{\theta^*}[\Delta_{\max}]^{2/3}}\right\rceil\right\}\right\},$$

we obtain

$$\begin{aligned}
&\mathrm{BR}_T(\pi^{\mathrm{ETC}}) \\
&\le 4(\sigma\alpha_T')^{2/3}(MK)^{1/3}\mathbb{E}_{\theta^*}[\Delta_{\max}]^{1/3}T^{2/3} + \mathbb{E}_{\theta^*}[\Delta_{\max}] + \sqrt{\mathbb{E}_{\theta^*}[\Delta_{\max}^2]} \\
&+ \frac{4(KM)^{2/3}\mathbb{E}_{\theta^*}[\Delta_{\max}]^{5/3}T^{1/3}}{(\sigma\alpha_T')^{2/3}}\left[1 + \log\left(1 + \frac{d}{8K}\left(\frac{(2\sigma\alpha_T' T)^{2/3}(2MK)^{1/3}}{\mathbb{E}_{\theta^*}[\Delta_{\max}]^{2/3}} + 1\right)\right)\right], \qquad (35)
\end{aligned}$$

To justify the truncation, let $x = (2\sigma\alpha_T' T)^{2/3}(2MK)^{1/3}/\mathbb{E}_{\theta^*}[\Delta_{\max}]^{2/3}$. If $x \ge T$, then $\mathrm{BR}_T \le T\mathbb{E}_{\theta^*}[\Delta_{\max}] \le \mathbb{E}_{\theta^*}[\Delta_{\max}]x$, which is bounded by the first term above. Otherwise, $x \le T_0 \le x+1$ justifies the substitution. This proves (21).

□

Proof of Theorem 7. During the first T_0 rounds, actions are selected uniformly. Hence, the expected exploration regret is exactly

$$\mathbb{E}\left[\sum_{t=1}^{T_0}\Delta_{I_t,A_t}\right] = T_0\sum_{m=1}^{M}\sum_{a\in\mathcal{K}}\frac{\beta_m}{K}\Delta_{m,a}.$$

For each type m, define the covariance event

$$\mathcal{C}_m := \left\{V_{T_0+1}^m \succeq \frac{T_0}{2}G_m\right\}.$$

By (25),

$$\Pr(\mathcal{C}_m^c) \le d \exp\left(-\frac{\beta_m T_0}{8K}\right).$$

We consider types with $\beta_m > 0$, since types with zero arrival probability contribute no regret.

For any exploitation round $t > T_0$, on $\mathcal{E}_t \cap \mathcal{C}_m$, the same argument as in (28) gives

$$\Delta_{m,\hat{a}_m^t} \le 2\alpha_T \sqrt{\frac{2K}{T_0 \beta_m}}.$$

If $\hat{a}_m^t$ is suboptimal, then $\Delta_{m,\hat{a}_m^t} \ge \Delta_{\min}$, and therefore

$$\Delta_{m,\hat{a}_m^t} \le \frac{\Delta_{m,\hat{a}_m^t}^2}{\Delta_{\min}} \le \frac{8K\alpha_T^2}{T_0 \beta_m \Delta_{\min}}.$$

The same upper bound holds when $\hat{a}_m^t$ is optimal, since its regret is zero.

For $t > T_0$, the arriving type I_t is independent of the pre-round history. Decomposing according to $\mathcal{E}_t$ and $\mathcal{C}_m$, and using $\Pr(\mathcal{E}_t^c) \le 1/T$, yields

$$\begin{aligned}
\mathbb{E}[\Delta_{I_t,A_t}] &\le \sum_{m\in\mathcal{M}} \beta_m \mathbb{E}\left[\Delta_{m,\hat{a}_m^t} \mathbf{1}\{\mathcal{E}_t \cap \mathcal{C}_m\}\right] \\
&\quad + \Delta_{\max} \Pr(\mathcal{E}_t^c) + \Delta_{\max} \sum_{m\in\mathcal{M}} \beta_m \Pr(\mathcal{C}_m^c) \\
&\le \frac{8KM\alpha_T^2}{T_0 \Delta_{\min}} + \frac{\Delta_{\max}}{T} + d\Delta_{\max} \sum_{m\in\mathcal{M}} \beta_m \exp\left(-\frac{\beta_m T_0}{8K}\right).
\end{aligned}$$

Summing over the $T - T_0$ exploitation rounds and adding the exploration regret proves (24).

Finally, since $xe^{-x} \le e^{-1}$ for $x \ge 0$,

$$\beta_m \exp\left(-\frac{\beta_m T_0}{8K}\right) \le \frac{8K}{eT_0}.$$

Summing over types gives the results. □

A.2 Details of revised Truthful Thompson Sampling

Hu and Duan Hu & Duan (2025) propose a truthful Thompson sampling algorithm with a problem-dependent regret bound for two agent types, i.e., $M = 2$. Its performance for the general multi-type setting with $M > 2$ remains an open question. We next introduce a key modification of their algorithm.

Let π^{TTS} denote their truthful Thompson sampling policy and π^{TS} the Thompson sampling action distribution in Abeille & Lazaric (2017), with its frequentist sampling rule specified below. Set $\delta = 1/(4T^2)$ for $T \ge 2$ and

$$\alpha_t^1 := \sigma \sqrt{2\log\left(\frac{(1+t/\lambda)^{d/2}}{\delta}\right)} + \sqrt{\lambda} L, \qquad \alpha_t^2 := \alpha_t^1 \sqrt{2d\log(2d/\delta)}, \tag{36}$$

$$\widetilde{\theta}_t \mid \mathcal{F}_t \sim \mathcal{N}\left(\hat{\theta}_t, (\alpha_t^1)^2 V_t^{-1}\right),$$

$$\pi_a^{\mathrm{TS}}(m \mid \mathcal{F}_t) := \Pr\left(a = \arg\max_{b\in\mathcal{K}} \langle \phi(m,b), \widetilde{\theta}_t \rangle \mid \mathcal{F}_t\right). \tag{37}$$

Use a fixed measurable rule to break ties.

Write $\Phi_m \in \mathbb{R}^{K\times d}$ for the matrix whose a-th row is $\phi(m,a)^\top$. At each time t, define $\mathbf{q}_t(m) = (q_t(m,1), \ldots, q_t(m,K)) \in \mathbb{R}^K$ for each agent type $m \in \mathcal{M}$, and solve the following linear pro-

Algorithm 2 Revised truthful Thompson sampling algorithm

Require: Known type probabilities $(\beta_m)_{m\in\mathcal{M}}$; feature vectors $\{\phi(m,a) : m \in \mathcal{M}, a \in \mathcal{K}\}$; regularization parameter $\lambda > 0$.

1: Initialize $(\hat{\theta}_1, W_1) \leftarrow (0_d, \lambda I_d)$ (frequentist) or $(\hat{\theta}_1, W_1) \leftarrow (\mu_0, \Sigma_0)$ (Bayesian).
2: Set $\gamma = \beta_{\min}$.
3: **for** $t = 1, \ldots, T$ **do**
4: Compute, for all $m \in \mathcal{M}$ and $a \in \mathcal{K}$,

$$\pi_a^{\mathrm{TS}}(m \mid \mathcal{F}_t) \leftarrow \Pr\left(a = \arg\max_{b\in\mathcal{K}} \langle \phi(m,b), \widetilde{\theta}_t\rangle \,\middle|\, \mathcal{F}_t\right),$$

where

$$\widetilde{\theta}_t \mid \mathcal{F}_t \sim \begin{cases} \mathcal{N}\left(\hat{\theta}_t, (\alpha_t^1)^2 V_t^{-1}\right), & \text{frequentist,} \\ \mathcal{N}\left(\hat{\theta}_t, \Sigma_t^{-1}\right), & \text{Bayesian.} \end{cases}$$

5: Solve the LP in (39) and obtain the solution $\mathbf{q}_t$. $\pi^{\text{R-TTS}}(\cdot|\mathcal{F}_t) \leftarrow \mathbf{q}_t$.
6: Observe the reported type $I_t' \in \mathcal{M}$, and sample $A_t \sim \pi^{\text{R-TTS}}(I_t'|\mathcal{F}_t)$.
7: Observe the reward r_t, and update $\hat{\theta}_{t+1}$ and W_{t+1} using (1) (frequentist) or (2) (Bayesian).
8: **end for**

gramming problem (LP):

$$\begin{aligned} \text{minimize} \quad & \max_{m\in\mathcal{M}} \left\|\pi^{\mathrm{TS}}(m|\mathcal{F}_t) - \mathbf{q}_t(m)\right\|_\infty \\ \text{s.t.} \quad & \left(\mathbf{q}_t(i) - \mathbf{q}_t(j)\right)^\top \Phi_i \hat{\theta}_t \geq 0, \quad \forall i \neq j,\ i,j \in \mathcal{M}, \\ & \sum_{m\in\mathcal{M}} \beta_m \mathbf{q}_t(m) = \sum_{m\in\mathcal{M}} \beta_m \pi^{\mathrm{TS}}(m|\mathcal{F}_t), \\ & \mathbf{q}_t(i) \in \Delta_K, \quad i \in \mathcal{M}. \end{aligned} \tag{38}$$

The solution defines π^{TTS}, i.e., $\pi^{\mathrm{TTS}}(m \mid \mathcal{F}_t) \leftarrow \mathbf{q}_t(m)$ for every $m \in \mathcal{M}$.

To obtain a regret upper bound for general $M > 2$, we modify the LP (38) as follows:

$$\begin{aligned} \text{minimize} \quad & \max_{m\in\mathcal{M}} \left\|\pi^{\mathrm{TS}}(m|\mathcal{F}_t) - \mathbf{q}_t(m)\right\|_\infty \\ \text{s.t.} \quad & \left(\mathbf{q}_t(i) - \mathbf{q}_t(j)\right)^\top \Phi_i \hat{\theta}_t \geq 0, \quad \forall i \neq j,\ i,j \in \mathcal{M}, \\ & \sum_{m\in\mathcal{M}} \beta_m \mathbf{q}_t(m) = \sum_{m\in\mathcal{M}} \beta_m \pi^{\mathrm{TS}}(m|\mathcal{F}_t), \\ & q_t(m,a) \geq \gamma \pi_a^{\mathrm{TS}}(m|\mathcal{F}_t), \quad m \in \mathcal{M},\ a \in \mathcal{K}, \\ & \mathbf{q}_t(i) \in \Delta_K, \quad i \in \mathcal{M}, \end{aligned} \tag{39}$$

where γ is a constant specified below. We give the following result to show that the new LP (39) is valid.

Lemma 2. *The LP* (39) *admits a feasible solution at every time* t *for* $\gamma = \beta_{\min}$.

Proof. Assign the same distribution to every report:

$$\bar{q}_t(m,a) := \sum_{i\in\mathcal{M}} \beta_i \pi_a^{\mathrm{TS}}(i \mid \mathcal{F}_t). \tag{40}$$

The incentive constraints hold with equality, and the action marginals are preserved. Moreover, $\bar{q}_t(m,a) \geq \beta_m \pi_a^{\mathrm{TS}}(m \mid \mathcal{F}_t) \geq \beta_{\min} \pi_a^{\mathrm{TS}}(m \mid \mathcal{F}_t)$. □

The solution of (39) defines the revised truthful Thompson sampling policy via $\pi^{\text{R-TTS}}(m \mid \mathcal{F}_t) \leftarrow \mathbf{q}_t(m)$ for every $m \in \mathcal{M}$. The complete procedure of the revised truthful Thompson sampling algorithm is specified in Algorithm 2.

We first give the problem-dependent regret bound of revised truthful Thompson sampling algorithm in Theorem 8, and then the problem-independent regret bounds in Theorem 9.

Theorem 8 (Complete version of problem-dependent regret of Theorem 2)**.** *Consider the frequentist model with a unique true optimal action for each agent type. Use $\gamma = \beta_{\min}$. Let Φ be the standard normal cumulative distribution function and set*

$$p := \Phi(-1), \qquad \tau := \min\left\{\Phi(-2), \frac{\Delta_{\min}}{4L}\right\}.$$

Then $\pi^{\text{R-TTS}}$ is strategyproof and

$$R_T(\pi^{\text{R-TTS}}; \nu(\theta^*)) \le \frac{2dL\max\{1,\lambda^{-1}\}}{\beta_{\min}^2\tau\Delta_{\min}^2}\left(\frac{4\alpha_T^2}{p} + \alpha_T^1 + \alpha_T^2\right)^2 \log\left(1 + \frac{T}{\lambda d}\right) + \frac{1}{4T}\left(L + \frac{L^3}{\beta_{\min}\tau\Delta_{\min}^2}\right). \tag{41}$$

This implies

$$R_T(\pi^{\text{R-TTS}}; \nu(\theta^*)) \in O\left(\frac{d^3\log^3(dT)}{\beta_{\min}^2\Delta_{\min}^3}\right).$$

Theorem 9 (Complete version of problem-independent regret of Theorem 2)**.** *Use $\gamma = \beta_{\min}$. $\pi^{\text{R-TTS}}$ satisfies*

$$R_T(\pi^{\text{R-TTS}}) = \Omega(T),\ \text{BR}_T(\pi^{\text{R-TTS}}) = \Omega(T). \tag{42}$$

Proof of Theorem 8. **Virtual clock.** We use the following implementation rule to apply the revised truthful Thompson sampling policy. The third constraint in (39) means that each round can use TS with probability γ and the remaining distribution $\big(\mathbf{q}_t(m) - \gamma\pi^{\text{TS}}(m \mid \mathcal{F}_t)\big)/(1-\gamma)$ otherwise. Thus, on average, one TS round occurs every $1/\gamma$ rounds.

Feasibility of $\pi^{\text{R-TTS}}$ follows from Lemma 2, and the first constraint of (39) gives strategyproofness. We next evaluate regret. Define

$$\mathcal{E}_t := \{\|\hat{\theta}_t - \theta^*\|_{V_t} \le \alpha_t^1\}.$$

Theorem 2 of Abbasi-yadkori et al. (2011), applied to the actual adaptive observation sequence, gives $\Pr(\mathcal{E}_t^c) \le \delta$ for each t.

Controlling the LP adjustment. Fix a history in $\mathcal{E}_t$ and write

$$r := \max_m\left\{1 - \pi^{\text{TS}}_{a_m^*}(m \mid \mathcal{F}_t)\right\}.$$

Uniqueness implies

$$\sum_{m,a}\beta_m\pi_a^{\text{TS}}(m \mid \mathcal{F}_t)\Delta_{m,a}^2 \ge \Delta_{\min}^2\sum_m\beta_m\left(1 - \pi^{\text{TS}}_{a_m^*}(m \mid \mathcal{F}_t)\right) \ge \beta_{\min}\Delta_{\min}^2 r. \tag{43}$$

Choose $\tau = \min\{\Phi(-2), \Delta_{\min}/(4L)\}$. First suppose $r < \tau$. For $a \ne a_m^*$, TS cannot choose a_m^* if the sampled score of a is larger. Therefore,

$$r \ge \Phi\left(-\frac{\langle\phi(m,a_m^*) - \phi(m,a), \hat{\theta}_t\rangle}{\alpha_t^1\|\phi(m,a_m^*) - \phi(m,a)\|_{V_t^{-1}}}\right).$$

The denominator is positive because $\Delta_{m,a} > 0$. Since $r < \Phi(-2)$,

$$2 \le \frac{\langle\phi(m,a_m^*) - \phi(m,a), \hat{\theta}_t\rangle}{\alpha_t^1\|\phi(m,a_m^*) - \phi(m,a)\|_{V_t^{-1}}}.$$

On $\mathcal{E}_t$, we have $|\langle\phi(m,a_m^*) - \phi(m,a), \hat{\theta}_t - \theta^*\rangle| \le \alpha_t^1\|\phi(m,a_m^*) - \phi(m,a)\|_{V_t^{-1}}$. Consequently,

$$\frac{2}{3}\Delta_{m,a} \le \langle\phi(m,a_m^*) - \phi(m,a), \hat{\theta}_t\rangle \le 2\Delta_{m,a} \le 2L. \tag{44}$$

Thus when $r < \tau$ and on $\mathcal{E}_t$, each true optimal action is also uniquely optimal under the public estimate.

Construct the following distribution that averages the TS policies of types sharing the same optimal action:

$$\bar{\mathbf{q}}_t(m) := \frac{\sum_{i:a_i^*=a_m^*} \beta_i \pi^{\mathrm{TS}}(i \mid \mathcal{F}_t)}{\sum_{i:a_i^*=a_m^*} \beta_i}.$$

We show that this policy satisfies the constraints in (39) on $\mathcal{E}_t$ and $r < \tau$. This policy preserves the action marginals and satisfies retention, since $\bar{q}_t(m,a) \geq \beta_m \pi_a^{\mathrm{TS}}(m \mid \mathcal{F}_t) \geq \gamma \pi_a^{\mathrm{TS}}(m \mid \mathcal{F}_t)$. It assigns probability at least $1-r$ to a_m^* for any m. By (44) and $r < \tau$, type m has estimated regret at most $2Lr \leq \Delta_{\min}/2$ if truthful reports. Reporting a type with a different optimal action gives estimated regret at least $(1-r)(2\Delta_{\min}/3) \geq \Delta_{\min}/2$, because $r < \Phi(-2) < 1/4$. Types sharing the same optimal action receive identical distributions. Thus $\bar{\mathbf{q}}_t$ satisfies every incentive constraint and is feasible for (39).

Both this candidate $\bar{\mathbf{q}}_t$ and TS put mass at least $1-r$ on a_m^*, so their distance in the LP objective is at most r. By optimality, the solution of (39) satisfies

$$\max_m \|\mathbf{q}_t(m) - \pi^{\mathrm{TS}}(m \mid \mathcal{F}_t)\|_\infty \leq r.$$

Hence $q_t(m, a_m^*) \geq 1-2r$, and the conditional regret is at most $2Lr$. If instead $r \geq \tau$, it is at most $L \leq Lr/\tau$. Combining both cases with (43) proves, on $\mathcal{E}_t$,

$$\sum_{m,a} \beta_m q_t(m,a) \Delta_{m,a} \leq Lr/\tau \leq \frac{L}{\beta_{\min}\tau\Delta_{\min}^2} \sum_{m,a} \beta_m \pi_a^{\mathrm{TS}}(m \mid \mathcal{F}_t)\Delta_{m,a}^2. \tag{45}$$

This controls all R-TTS rounds.

Squared regret of a virtual TS action. We adapt the one-round argument in Section 5 of Abeille & Lazaric (2017), including the sampling tail. Fix a type m and a history in $\mathcal{E}_t$, and let A be the action selected by a fresh TS draw. A Gaussian tail bound gives

$$\Pr\left(\|\widetilde{\theta}_t - \hat{\theta}_t\|_{V_t} > \alpha_t^2 \mid \mathcal{F}_t\right) \leq \delta.$$

Define $a_m^* = \arg\max_{a \in \mathcal{K}} \langle \phi(m,a), \theta^* \rangle$ for any $m \in \mathcal{M}$. The sampled optimal value is at least the true optimal value with conditional probability at least p, by Gaussian anti-concentration in the direction $\phi(m, a_m^*)$. If that feature is zero, this holds for every draw. Since $\delta \leq 1/16 < p/2$, the sample is both optimistic and inside the radius-α_t^2 ellipsoid with probability at least $p/2$.

Define

$$g_t := \langle \phi(m, a_m^*), \theta^* \rangle - \min_{\|\theta - \hat{\theta}_t\|_{V_t} \leq \alpha_t^2} \max_a \langle \phi(m,a), \theta \rangle.$$

On $\mathcal{E}_t$, we have $\|\theta^* - \hat{\theta}_t\|_{V_t} \leq \alpha_t^1 \leq \alpha_t^2$, so θ^* is feasible in the minimization defining g_t. The minimum is therefore at most $\max_a \langle \phi(m,a), \theta^* \rangle = \langle \phi(m, a_m^*), \theta^* \rangle$, which shows that $g_t \geq 0$.

Let θ' attain this minimum; such a parameter exists because the ellipsoid is compact and the objective is continuous. For an optimistic sample $\widetilde{\theta}_t$ inside the ellipsoid, its maximizing action A satisfies

$$\langle \phi(m,A), \widetilde{\theta}_t \rangle = \max_a \langle \phi(m,a), \widetilde{\theta}_t \rangle \geq \langle \phi(m, a_m^*), \theta^* \rangle.$$

Consequently,

$$\begin{aligned} g_t &\leq \langle \phi(m,A), \widetilde{\theta}_t \rangle - \max_a \langle \phi(m,a), \theta' \rangle \\ &\leq \langle \phi(m,A), \widetilde{\theta}_t - \theta' \rangle \\ &\leq \|\phi(m,A)\|_{V_t^{-1}} \|\widetilde{\theta}_t - \theta'\|_{V_t} \\ &\leq 2\alpha_t^2 \|\phi(m,A)\|_{V_t^{-1}}. \end{aligned}$$

The last step uses the fact that both $\widetilde{\theta}_t$ and θ' are within distance α_t^2 of $\hat{\theta}_t$ in the V_t-norm.

Conditional on $\mathcal{F}_t$, g_t is fixed and the samples just considered have probability at least $p/2$. Multiplying the preceding inequality by the indicator of these samples and taking conditional expectations gives

$$\frac{p}{2} g_t \leq 2\alpha_t^2 \mathbb{E}[\|\phi(m,A)\|_{V_t^{-1}} \mid \mathcal{F}_t] \Longrightarrow g_t \leq \frac{4\alpha_t^2}{p} \mathbb{E}[\|\phi(m,A)\|_{V_t^{-1}} \mid \mathcal{F}_t].$$

Finally, for any sample inside the ellipsoid, whether optimistic or not, the definition of the minimum gives

$$\langle \phi(m, A), \widetilde{\theta}_t \rangle = \max_a \langle \phi(m, a), \widetilde{\theta}_t \rangle \geq \langle \phi(m, a_m^*), \theta^* \rangle - g_t.$$

Adding and subtracting the sampled reward, we obtain

$$\begin{aligned}\Delta_{m,A} &= \langle \phi(m, a_m^*), \theta^* \rangle - \langle \phi(m, A), \widetilde{\theta}_t \rangle + \langle \phi(m, A), \widetilde{\theta}_t - \theta^* \rangle \\ &\leq g_t + \|\phi(m, A)\|_{V_t^{-1}} \big(\|\widetilde{\theta}_t - \hat{\theta}_t\|_{V_t} + \|\hat{\theta}_t - \theta^*\|_{V_t} \big) \\ &\leq g_t + (\alpha_t^1 + \alpha_t^2) \|\phi(m, A)\|_{V_t^{-1}}.\end{aligned}$$

Adding $L^2\delta$ for samples outside the ellipsoid gives

$$\mathbb{E}[\Delta_{m,A}^2 \mid \mathcal{F}_t] \leq \left(\frac{4\alpha_T^2}{p} + \alpha_T^1 + \alpha_T^2 \right)^2 \mathbb{E}[\|\phi(m, A)\|_{V_t^{-1}}^2 \mid \mathcal{F}_t] + L^2\delta. \tag{46}$$

These expectations concern a virtual TS draw conditional on the actual R-TTS history.

Accumulating information. Averaging (46) over types and applying (45) yields

$$\begin{aligned}R_T(\pi^{\text{R-TTS}}; \nu) \leq \frac{L}{\beta_{\min} \tau \Delta_{\min}^2} &\left(\frac{4\alpha_T^2}{p} + \alpha_T^1 + \alpha_T^2 \right)^2 \\ &\times \sum_{t=1}^{T} \mathbb{E}\left[\sum_{m,a} \beta_m \pi_a^{\text{TS}}(m \mid \mathcal{F}_t) \|\phi(m, a)\|_{V_t^{-1}}^2 \right] \\ &+ \left(L + \frac{L^3}{\beta_{\min} \tau \Delta_{\min}^2} \right) T\delta.\end{aligned} \tag{47}$$

The last line includes both estimation and sampling failures. The retention constraint in (39) gives

$$\begin{aligned}\sum_{m,a} \beta_m \pi_a^{\text{TS}}(m \mid \mathcal{F}_t) \|\phi(m, a)\|_{V_t^{-1}}^2 &\leq \frac{1}{\gamma} \sum_{m,a} \beta_m q_t(m, a) \|\phi(m, a)\|_{V_t^{-1}}^2 \\ &= \frac{1}{\gamma} \mathbb{E}[\|\phi(I_t, A_t)\|_{V_t^{-1}}^2 \mid \mathcal{F}_t].\end{aligned} \tag{48}$$

The expectation on the right is under R-TTS. Since every squared width is at most $1/\lambda$, the elliptical-potential inequality, with the factor $\max\{1, \lambda^{-1}\}$ accounting for widths larger than one, gives

$$\begin{aligned}\sum_{t=1}^{T} \|\phi(I_t, A_t)\|_{V_t^{-1}}^2 &\leq 2 \max\{1, \lambda^{-1}\} \log \frac{\det V_{T+1}}{\det(\lambda I_d)} \\ &\leq 2d \max\{1, \lambda^{-1}\} \log \left(1 + \frac{T}{\lambda d} \right).\end{aligned} \tag{49}$$

Substituting into (47) and taking $\gamma = \beta_{\min}$ proves (41). Finally, $(\alpha_T^2)^2 = O(d^2 \log^2(dT))$ and $1/\tau = O(L/\Delta_{\min})$, giving the stated asymptotic bound. □

Proof of Theorem 9. We first prove the Bayesian regret result. Consider an instance with $M = 3$, $K = 2$, and $d = 3$, where

$$\begin{aligned}\phi(1, 1) &= e_1, \qquad \phi(1, 2) = 0, \\ \phi(2, 1) &= 0, \qquad \phi(2, 2) = e_2, \\ \phi(3, 1) &= 0, \qquad \phi(3, 2) = e_3.\end{aligned}$$

Let

$$\beta_1 = \varepsilon, \qquad \beta_2 = \beta_3 = \frac{1 - \varepsilon}{2}, \qquad 0 < \varepsilon \leq \frac{1}{2T},$$

and take the Bayesian setting with

$$\mu_0 = (1/2, -1/2, 1/2), \qquad \Sigma_0 = \text{diag}(1, 400, 400), \qquad \sigma^2 = 1.$$

Write $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*)$. In this setting, consider the event $\mathcal{B} = \left\{ \left|\theta_2^* + \frac{1}{2}\right| \leq 0.05, \ \left|\theta_3^* - \frac{1}{2}\right| \leq 0.05 \right\}$. This event has probability $\Pr(\mathcal{B}) = (\Phi(1) - \Phi(-1))^2 > 0$.

In this setting, informative features are mutually orthogonal, so observations from one type provide no information about other types. Then, with probability at least $(1-\varepsilon)^T \geq 1 - T\varepsilon \geq 1/2$, no type-1 agent arrives during the horizon, and the posterior estimate $(\hat{\theta}_t)^{(1)}$ remains the same as the prior estimate, so

$$\pi_1^{\mathrm{TS}}(1 \mid \mathcal{F}_t) = c := \Phi(1/2) \approx 0.6915.$$

To control the frequent types uniformly over time, pre-generate independent standard Gaussian noise sequences for the informative pairs $(2,2)$ and $(3,2)$, independently of the prior, type arrivals, and policy randomization. Let $S_{j,n}$ be the sum of the first n noises for pair $(j,2)$, with $S_{j,0} = 0$, and define

$$\mathcal{G} := \{S_{2,n} \leq 100 + n/4,\ S_{3,n} \geq -100 - n/4 \text{ for every } n \geq 0\}.$$

The nonnegative martingales $\exp(S_{2,n}/2 - n/8)$ and $\exp(-S_{3,n}/2 - n/8)$ each start at one. Crossing the corresponding boundary makes the martingale exceed e^{50}, so the maximal inequality and a union bound give $\Pr(\mathcal{G}) \geq 1 - 2e^{-50}$.

After n informative observations of type 2, its posterior mean and standard deviation are $(-200 + n\theta_2^* + S_{2,n})/(400+n)$ and $1/\sqrt{400+n}$. On $\mathcal{B} \cap \mathcal{G}$, their ratio is at most

$$\frac{-200 + n\theta_2^* + S_{2,n}}{\sqrt{400+n}} \leq -\frac{100 + 0.20n}{\sqrt{400+n}} \leq -5.$$

Similarly, the posterior mean divided by its standard deviation for type 3 is at least $(100 + 0.20n)/\sqrt{400+n} \geq 5$. These bounds hold for every observation count, including the counts generated by adaptive action choices. Therefore,

$$\pi_1^{\mathrm{TS}}(2 \mid \mathcal{F}_t) \geq 0.99, \qquad \pi_1^{\mathrm{TS}}(3 \mid \mathcal{F}_t) \leq 0.01$$

throughout the horizon on $\mathcal{B} \cap \mathcal{G}$. When no type-1 agent arrives as well, types 1 and 2 both prefer action 1 under the public estimate, whereas type 3 prefers action 2. The truthfulness constraints therefore imply

$$\pi_1^{\text{R-TTS}}(1 \mid \mathcal{F}_t) = \pi_1^{\text{R-TTS}}(2 \mid \mathcal{F}_t) \geq \pi_1^{\text{R-TTS}}(3 \mid \mathcal{F}_t).$$

Together with the probability-preservation constraint in (39), this gives

$$\frac{1+\varepsilon}{2}\pi_1^{\text{R-TTS}}(2 \mid \mathcal{F}_t) + \frac{1-\varepsilon}{2}\pi_1^{\text{R-TTS}}(3 \mid \mathcal{F}_t) = \varepsilon c + \frac{1-\varepsilon}{2}\left(\pi_1^{\mathrm{TS}}(2 \mid \mathcal{F}_t) + \pi_1^{\mathrm{TS}}(3 \mid \mathcal{F}_t)\right).$$

Then, consider the candidate policy

$$\bar{\pi}_1(1 \mid \mathcal{F}_t) = \bar{\pi}_1(2 \mid \mathcal{F}_t) = \frac{1+c}{2},$$

$$\bar{\pi}_1(3 \mid \mathcal{F}_t) = \pi_1^{\mathrm{TS}}(2 \mid \mathcal{F}_t) + \pi_1^{\mathrm{TS}}(3 \mid \mathcal{F}_t) - \frac{1+c}{2} - \frac{\varepsilon(1-c)}{1-\varepsilon},$$

with $\bar{\pi}_2(m \mid \mathcal{F}_t) = 1 - \bar{\pi}_1(m \mid \mathcal{F}_t)$. For $T \geq 50$, this policy satisfies all constraints in (39), including the retention constraints with $\gamma = \beta_{\min} = \varepsilon$, and achieves an objective value at most $(1-c)/2$. Consequently, every optimal solution satisfies

$$\left|\pi_1^{\text{R-TTS}}(1 \mid \mathcal{F}_t) - c\right| \leq \frac{1-c}{2}.$$

By truthfulness,

$$\pi_2^{\text{R-TTS}}(2 \mid \mathcal{F}_t) = 1 - \pi_1^{\text{R-TTS}}(1 \mid \mathcal{F}_t) \geq \frac{1-c}{2} \approx 0.1543.$$

Thus, on these histories, R-TTS assigns type 2 its suboptimal action with probability at least $(1-c)/2$. Each such allocation incurs regret $-\theta_2^* \geq 0.45$ on $\mathcal{B}$.

The events $\mathcal{B}$, $\mathcal{G}$, and no type-1 arrivals are independent. Conditional on these events, each arriving type is 2 or 3 with probability $1/2$, and the fresh randomization used to sample an action still follows the policy q^t. Hence, for $T \geq 50$,

$$\begin{aligned}\mathrm{BR}_T(\pi^{\text{R-TTS}}) &\geq 0.45\,\frac{1-c}{4}\,T\,\Pr(\mathcal{B})(1-2e^{-50})(1-\varepsilon)^T \\ &\geq 0.45\,\frac{1-c}{8}\,\Pr(\mathcal{B})(1-2e^{-50})T = \Omega(T).\end{aligned}$$

The frequentist lower bound follows an argument analogous to the Bayesian case, with a short initialization event replacing the nonzero prior mean. Use the same features, set $\theta^* = (1/2, -1/2, 1/2)$,

$\lambda = 1$, $L = 2$, $\varepsilon = 1/(2T)$, and take independent $\mathcal{N}(0, 1/400)$ noise. With probability bounded below by a positive constant independent of T, type 2 receives an informative observation in the first round, type 1 arrives exactly once during rounds $2, \dots, \lfloor T/2 \rfloor$, receives action 1 with reward in $[0.3, 0.4]$, and never arrives thereafter. Intersecting this event with suitable Gaussian noise bounds still leaves a probability bounded away from zero and ensures that, throughout the second half of the horizon, the TS probabilities $p_m^t := \pi_1^{\mathrm{TS}}(m \mid \mathcal{F}_t)$ satisfy $1/2 < p_1^t \le p_2^t$, $p_1^t \le 3/4$, and $p_3^t \le 1/2$. Write $q_m^t := \pi_1^{\text{R-TTS}}(m \mid \mathcal{F}_t)$ for the corresponding R-TTS probabilities. Types 1 and 2 then strictly prefer action 1 under the public estimate, so strategyproofness forces $q_1^t = q_2^t$. As in the Bayesian argument, a feasible-policy comparison in the LP gives $q_2^t \le (p_1^t + p_2^t)/2 \le 7/8$. Thus type 2 receives its suboptimal action with probability at least $1/8$, incurring gap $1/2$ on a constant-probability event lasting $\Omega(T)$ rounds. Consequently, $R_T(\pi^{\text{R-TTS}}) = \Omega(T)$, where the worst-case instance has a type distribution depending on T.

□

A.3 Proof of Theorem 3

We construct the following problem instance to prove the regret lower bound. For $L > 2$, set

$$\kappa := \sqrt{\frac{L^2/4 - 1}{d}},$$

and fix $0 < \varepsilon < \min\{1/2, \kappa\}$. Consider the parameter space

$$\Theta := \left\{ \theta \in \mathbb{R}^{d+1} : \theta^{(0)} = 1,\ \theta^{(i)} \in \{-\kappa, \kappa\},\ i = 1, \dots, d \right\}.$$

Then $\|\theta\|_2^2 = 1 + d\kappa^2 = L^2/4$ for every $\theta \in \Theta$. Choose independent $\mathcal{N}(0, \sigma^2)$ observation noise.

Consider M types of agents and K actions, where $2 \le K \le \min\{M, d+1\}$. Set $M_1 = M - K + 1 \ge 1$. Index the agent types by $\{0, \dots, M_1 - 1\} \cup \{M_1, \dots, M_1 + K - 2\}$ and the actions by $\{0, \dots, K-1\}$. For each type $m \in \{0, \dots, M_1 - 1\}$, set

$$\phi(m, 0) = \mathbf{e}_0^{d+1}, \qquad \phi(m, a) = \mathbf{0}^{d+1}, \quad a = 1, \dots, K-1.$$

For $i = 1, \dots, K-1$, set

$$\phi(M_1 + i - 1, i) = \frac{\varepsilon}{\kappa} \mathbf{e}_i^{d+1}, \qquad \phi(M_1 + i - 1, a) = \mathbf{0}^{d+1}, \quad a \ne i.$$

Since $\varepsilon < \kappa$, all feature vectors have Euclidean norm at most one. Thus

$$\langle \hat{\theta}_t, \phi(M_1 + i - 1, i) \rangle = \frac{\varepsilon}{\kappa} \hat{\theta}_t^{(i)}, \qquad \langle \hat{\theta}_t, \phi(m, 0) \rangle = \hat{\theta}_t^{(0)}, \quad m < M_1.$$

Choose $\beta_0 = \cdots = \beta_{M_1 - 1} = \frac{1}{2M_1}$, $\beta_{M_1} = \cdots = \beta_{M_1 + K - 2} = \frac{1}{2K-2}$. Write

$$q_m^t(a) := \mathbb{P}(A_t = a \mid I_t = m, \mathcal{F}_t), \qquad N_{m,a}(T) := \sum_{t=1}^{T} \mathbf{1}\{I_t = m, A_t = a\},$$

and $N_m(T) := \sum_{t=1}^{T} \mathbf{1}\{I_t = m\}$. For $i = 1, \dots, K-1$ and $m = 0, \dots, M_1 - 1$, the strategyproof condition in Definition 1 gives

$$\frac{\varepsilon}{\kappa} \hat{\theta}_t^{(i)} \big(q_{M_1 + i - 1}^t(i) - q_m^t(i) \big) \ge 0.$$

Since $\varepsilon/\kappa > 0$, this implies

$$\hat{\theta}_t^{(i)} < 0 \quad \Longrightarrow \quad q_{M_1 + i - 1}^t(i) \le q_m^t(i), \quad m = 0, \dots, M_1 - 1. \tag{50}$$

Denote the regret incurred by type m agents by $R_T^{\theta}(m)$, and, for $m < M_1$, the regret incurred by type m agents upon choosing action i by $R_T^{\theta}(m, i)$. Fix $i \in \{1, \dots, K-1\}$ and choose $\theta^- \in \Theta$ with $(\theta^-)^{(i)} = -\kappa$.

We first lower-bound $R_T^{\theta^-}(M_1+i-1)+\sum_{m=0}^{M_1-1} R_T^{\theta^-}(m,i)$. Using (50), the identity $\mathbb{E}_{\theta^-}[\mathbf{1}\{I_t = m, A_t = a\} \mid \mathcal{F}_t] = \beta_m q_m^t(a)$, and $\sum_{m=0}^{M_1-1} \beta_m/\beta_{M_1+i-1} = K-1$, we obtain

$$
\begin{aligned}
&R_T^{\theta^-}(M_1+i-1)+\sum_{m=0}^{M_1-1} R_T^{\theta^-}(m,i)\\
=&\mathbb{E}_{\theta^-}\left[\sum_{t=1}^{T}\left(\varepsilon\mathbf{1}\{I_t = M_1+i-1, A_t = i\}+\sum_{m=0}^{M_1-1}\mathbf{1}\{I_t=m, A_t=i\}\right)\right]\\
\geq&\mathbb{E}_{\theta^-}\left[\sum_{t=1}^{T}\left(\beta_{M_1+i-1}\varepsilon q_{M_1+i-1}^t(i)+\sum_{m=0}^{M_1-1}\beta_m q_m^t(i)\right)\mathbf{1}\{\hat{\theta}_t^{(i)}<0\}\right]\\
\geq&(\varepsilon+K-1)\mathbb{E}_{\theta^-}\left[\sum_{t=1}^{T}\mathbf{1}\{I_t=M_1+i-1, A_t=i, \hat{\theta}_t^{(i)}<0\}\right]\\
=&(\varepsilon+K-1)\Big(\mathbb{E}_{\theta^-}[N_{M_1+i-1,i}(T)]-\mathbb{E}_{\theta^-}\left[\sum_{t=1}^{T}\mathbf{1}\{I_t=M_1+i-1, A_t=i, \hat{\theta}_t^{(i)}\geq 0\}\right]\Big). \quad (51)
\end{aligned}
$$

For the pair (M_1+i-1, i), consider a pre-generated sequence of independent rewards $r_1, r_2, \ldots$, and let $\widetilde{\theta}_j^{(i)}$ be the ridge estimate of coordinate i based on its first j rewards. Then

$$
\widetilde{\theta}_j^{(i)} = \frac{(\varepsilon/\kappa)\sum_{\ell=1}^{j} r_\ell}{\lambda + j(\varepsilon/\kappa)^2}.
$$

Under θ^-, these rewards have mean $(\varepsilon/\kappa)(-\kappa) = -\varepsilon$. Since the estimate has the same sign as the sum of rewards, a Gaussian tail bound gives

$$
\begin{aligned}
\mathbb{E}_{\theta^-}\left[\sum_{t=1}^{T}\mathbf{1}\{I_t=M_1+i-1, A_t=i, \hat{\theta}_t^{(i)}\geq 0\}\right] &\leq 1+\sum_{j=1}^{\infty}\mathbb{P}_{\theta^-}\left(\widetilde{\theta}_j^{(i)}\geq 0\right)\\
&\leq 1+\sum_{j=1}^{\infty}\exp\left(-\frac{j\varepsilon^2}{2\sigma^2}\right)\\
&= 1+\frac{1}{\exp(\varepsilon^2/(2\sigma^2))-1}\\
&\leq 1+\frac{2\sigma^2}{\varepsilon^2}. \quad (52)
\end{aligned}
$$

Here the initial 1 accounts for the estimate before the first informative observation. Combining (51) and (52), we have

$$
R_T^{\theta^-}(M_1+i-1)+\sum_{m=0}^{M_1-1} R_T^{\theta^-}(m,i) \geq (\varepsilon+K-1)\left(\mathbb{E}_{\theta^-}[N_{M_1+i-1,i}(T)]-1-\frac{2\sigma^2}{\varepsilon^2}\right). \quad (53)
$$

Now, choose another parameter $\theta^+ \in \Theta$ for which $(\theta^+)^{(i)} = -(\theta^-)^{(i)} = \kappa$ and all other coordinates remain the same.

On the other hand, fix a deterministic threshold $c \in (0, \beta_{M_1+i-1}T/2)$. Directly from the regret incurred on the two instances,

$$
\begin{aligned}
R_T^{\theta^-}(M_1+i-1) &\geq c\varepsilon\,\mathbb{P}_{\theta^-}(N_{M_1+i-1,i}(T)>c), \quad (54)\\
R_T^{\theta^+}(M_1+i-1) &\geq \left(\frac{\beta_{M_1+i-1}T}{2}-c\right)\varepsilon\\
&\quad\times\mathbb{P}_{\theta^+}\left(N_{M_1+i-1,i}(T)\leq c,\ N_{M_1+i-1}(T)>\frac{\beta_{M_1+i-1}T}{2}\right). \quad (55)
\end{aligned}
$$

Consequently,

$$R_T^{\theta^-}(M_1+i-1)+R_T^{\theta^+}(M_1+i-1)$$
$$\geq \min\left\{c\varepsilon,\ \left(\frac{\beta_{M_1+i-1}T}{2}-c\right)\varepsilon\right\}\times\left[\mathbb{P}_{\theta^-}(N_{M_1+i-1,i}(T)>c)+\mathbb{P}_{\theta^+}(N_{M_1+i-1,i}(T)\leq c)\right.$$
$$\left.-\mathbb{P}\left(N_{M_1+i-1}(T)\leq\frac{\beta_{M_1+i-1}T}{2}\right)\right]. \tag{56}$$

Let $\mathbb{P}_{\theta^-}$ and $\mathbb{P}_{\theta^+}$ denote the laws of the full interaction history under the same policy on the two instances. The Bretagnolle–Huber inequality gives

$$\mathbb{P}_{\theta^-}(N_{M_1+i-1,i}(T)>c)+\mathbb{P}_{\theta^+}(N_{M_1+i-1,i}(T)\leq c)\geq\frac{1}{2}\exp(-D_{\mathrm{KL}}(\mathbb{P}_{\theta^-},\mathbb{P}_{\theta^+}))\,. \tag{57}$$

Only observations of the pair (M_1+i-1,i) distinguish the two instances, and the corresponding Gaussian means differ by 2ε. The chain rule for relative entropy therefore yields

$$D_{\mathrm{KL}}(\mathbb{P}_{\theta^-},\mathbb{P}_{\theta^+})=\frac{2\varepsilon^2}{\sigma^2}\mathbb{E}_{\theta^-}[N_{M_1+i-1,i}(T)].$$

Moreover, $N_{M_1+i-1}(T)\sim\mathrm{Binomial}(T,\beta_{M_1+i-1})$ under either instance, so Chernoff's inequality gives

$$\mathbb{P}\left(N_{M_1+i-1}(T)\leq\frac{\beta_{M_1+i-1}T}{2}\right)\leq\exp\left(-\frac{\beta_{M_1+i-1}T}{8}\right).$$

Choosing $c=\beta_{M_1+i-1}T/4$ in (56), we conclude that

$$R_T^{\theta^-}(M_1+i-1)+R_T^{\theta^+}(M_1+i-1)$$
$$\geq\frac{\varepsilon\beta_{M_1+i-1}T}{4}\left[\frac{1}{2}\exp\left(-\frac{2\varepsilon^2\mathbb{E}_{\theta^-}[N_{M_1+i-1,i}(T)]}{\sigma^2}\right)-\exp\left(-\frac{\beta_{M_1+i-1}T}{8}\right)\right]. \tag{58}$$

We now distinguish two cases. If $\mathbb{E}_{\theta^-}[N_{M_1+i-1,i}(T)]\geq 2(1+2\sigma^2/\varepsilon^2)$, then (53) implies

$$R_T^{\theta^-}(M_1+i-1)+\sum_{m=0}^{M_1-1}R_T^{\theta^-}(m,i)\geq(K-1)\left(1+\frac{2\sigma^2}{\varepsilon^2}\right).$$

If instead $\mathbb{E}_{\theta^-}[N_{M_1+i-1,i}(T)]<2(1+2\sigma^2/\varepsilon^2)$, then the first exponential in (58) satisfies

$$\exp\left(-\frac{2\varepsilon^2\mathbb{E}_{\theta^-}[N_{M_1+i-1,i}(T)]}{\sigma^2}\right)\geq\exp\left(-8-\frac{4\varepsilon^2}{\sigma^2}\right)\geq\exp\left(-8-\frac{1}{\sigma^2}\right),$$

where the last inequality uses $\varepsilon<1/2$. For fixed $\sigma>0$ and sufficiently large T, the Chernoff term is at most $\exp(-8-1/\sigma^2)/4$. Hence,

$$R_T^{\theta^-}(M_1+i-1)+R_T^{\theta^+}(M_1+i-1)\geq\frac{\varepsilon\beta_{M_1+i-1}T}{16}\exp\left(-8-\frac{1}{\sigma^2}\right).$$

Since all regret contributions are nonnegative, both cases imply

$$\sum_{\theta\in\{\theta^-,\theta^+\}}\left[R_T^{\theta}(M_1+i-1)+\sum_{m=0}^{M_1-1}R_T^{\theta}(m,i)\right]=\Omega\left(\min\left\{(K-1)\frac{\sigma^2}{\varepsilon^2},\ \beta_{M_1+i-1}\varepsilon T\right\}\right).$$

Now, sum over all $\theta^-\in\Theta$ with $(\theta^-)^{(i)}=-\kappa$, pairing each with the parameter θ^+ obtained by flipping only coordinate i. These 2^{d-1} pairs partition Θ, so

$$\sum_{\substack{\theta^-\in\Theta\\(\theta^-)^{(i)}=-\kappa}}\ \sum_{\theta\in\{\theta^-,\theta^+\}}\left[R_T^{\theta}(M_1+i-1)+\sum_{m=0}^{M_1-1}R_T^{\theta}(m,i)\right]$$
$$=\sum_{\theta\in\Theta}\left[R_T^{\theta}(M_1+i-1)+\sum_{m=0}^{M_1-1}R_T^{\theta}(m,i)\right]$$
$$\geq 2^{d-1}\Omega\left(\min\left\{(K-1)\frac{\sigma^2}{\varepsilon^2},\ \beta_{M_1+i-1}\varepsilon T\right\}\right). \tag{59}$$

Summing over $i = 1, \dots, K-1$ and using $|\Theta| = 2^d$ and $\beta_{M_1+i-1} = 1/[2(K-1)]$, we obtain

$$\begin{aligned}
&\frac{1}{|\Theta|} \sum_{\theta \in \Theta} \sum_{i=1}^{K-1} \left[R_T^\theta(M_1 + i - 1) + \sum_{m=0}^{M_1-1} R_T^\theta(m, i) \right] \\
&= \sum_{i=1}^{K-1} \frac{1}{|\Theta|} \sum_{\theta \in \Theta} \left[R_T^\theta(M_1 + i - 1) + \sum_{m=0}^{M_1-1} R_T^\theta(m, i) \right] \\
&\geq \frac{K-1}{2} \Omega\left(\min\left\{ (K-1)\frac{\sigma^2}{\varepsilon^2},\ \frac{\varepsilon T}{2(K-1)} \right\} \right) \\
&= \Omega\left(\min\left\{ \frac{(K-1)^2\sigma^2}{\varepsilon^2},\ \varepsilon T \right\} \right). \qquad (60)
\end{aligned}$$

Since

$$R_T^\theta = \sum_{i=1}^{K-1} \left[R_T^\theta(M_1 + i - 1) + \sum_{m=0}^{M_1-1} R_T^\theta(m, i) \right],$$

the above inequality becomes

$$\frac{1}{|\Theta|} \sum_{\theta \in \Theta} R_T^\theta = \Omega\left(\min\left\{ \frac{(K-1)^2\sigma^2}{\varepsilon^2},\ \varepsilon T \right\} \right),$$

indicating that there exists a $\theta \in \Theta$ such that

$$R_T^\theta = \Omega\left(\min\left\{ \frac{(K-1)^2\sigma^2}{\varepsilon^2},\ \varepsilon T \right\} \right).$$

Finally, choose $K = \min\{M, d+1\}$ and $M_1 = M - K + 1$, so that $K - 1 = \min\{M-1, d\}$. Balancing the two terms by setting $\varepsilon = (\sigma^2(K-1)^2/T)^{1/3}$ gives a lower bound of $\Omega(\sigma^{2/3}(K-1)^{2/3}T^{2/3})$. For fixed $\sigma > 0$ and $L > 2$, assume

$$T > \max\left\{ 8\sigma^2(K-1)^2,\ \sigma^2(K-1)^2 \left(\frac{d}{L^2/4 - 1} \right)^{3/2},\ 16(K-1)\left(8 + \frac{1}{\sigma^2} + \log 4 \right) \right\}.$$

Then $\varepsilon = (\sigma^2(K-1)^2/T)^{1/3} < \min\{1/2, \kappa\}$, and

$$\exp\left(-\frac{\beta_{M_1+i-1} T}{8} \right) \leq \frac{1}{4} \exp\left(-8 - \frac{1}{\sigma^2} \right),$$

so all the preceding bounds apply. Since the policy was arbitrary, this establishes the minimax lower bound for sufficiently large T.

B Missing details from Section 4

Proof of Lemma 1. If $\mathcal{M}_t \neq \varnothing$, assign every reported type the same distribution $|\mathcal{M}_t|^{-1} \sum_{m \in \mathcal{M}_t} p_m^t$. Identical distributions satisfy all incentive constraints and retain probability at least $1/|\mathcal{M}_t|$ on each unresolved type's proposed action. If $\mathcal{M}_t = \varnothing$, the retention constraints are omitted and every proposal is estimated-optimal. This menu is already strategyproof, so $q^t = p^t$ is feasible and achieves the minimum objective value of zero. □

B.1 Proof of Theorem 4

Firstly, since the action is always chosen under q^t output from LP (8), the policy satisfies Definition 1 at each time step.

We then first prove the problem-independent regret upper bound (15) under the frequentist setting.

Define the following event in Algorithm 1:

$$\mathcal{B}_t := \left\{ \text{there exists } m \in \mathcal{M} \text{ such that } U_{m,t} \geq \varepsilon_t / \sqrt{\beta_m} \right\},$$

and define the following event in Algorithm 1:

$$\mathcal{E}_t := \left\{ \left| \langle \theta^* - \hat{\theta}_t, \phi(m,a) \rangle \right| \leq \alpha_t \|\phi(m,a)\|_{V_t^{-1}},\quad \forall (m,a) \in \mathcal{M} \times \mathcal{K} \right\}.$$

Define $\mathcal{E} = \cap_t \mathcal{E}_t$.

Frequentist regret. We first perform the following regret decomposition:

$$R_T(\pi^{\text{CGT-UCB}}) = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{I_t,A_t}\right] \tag{61}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\mathcal{B}_t, \mathcal{E}\}\Delta_{I_t,A_t} + \mathbf{1}\{\mathcal{B}_t^c, \mathcal{E}\}\Delta_{I_t,A_t} + \mathbf{1}\{\mathcal{E}^c\}\Delta_{I_t,A_t}\right]. \tag{62}$$

By Lemma 3, the first term above is upper bounded by

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\mathcal{B}_t, \mathcal{E}_t\}\Delta_{I_t,A_t}\right] &\le \Delta_{\max}\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\mathcal{B}_t, \mathcal{E}_t\}\right] \\
&\le \Delta_{\max}\mathbb{E}\left[\sum_{t=1}^{T} \frac{4\alpha_t^2}{\varepsilon_t^2}\mathbb{E}\Big[\,\|\phi(I_t,A_t)\|^2_{V_t^{-1}}\,\Big|\,\mathcal{F}_t\Big]\right] \\
&\le \Delta_{\max}\frac{4\alpha_T^2}{\varepsilon_T^2}\mathbb{E}\left[\sum_{t=1}^{T} \|\phi(I_t,A_t)\|^2_{V_t^{-1}}\right] \\
&\le \Delta_{\max}\frac{8\alpha_T^2 \max\{1,\lambda^{-1}\}}{\varepsilon_T^2} d\log\left(1+\frac{T}{\lambda d}\right),
\end{aligned} \tag{63}$$

where the third inequality is by the monotonicity of ε_t and α_t, and the last line is by the elliptical-potential theorem in Abbasi-yadkori et al. (2011). The factor $\max\{1,\lambda^{-1}\}$ accounts for squared widths exceeding one when $\lambda < 1$.

For the second term in (62), under the event $\mathcal{B}_t^c \cap \mathcal{E}$, $p_m^t = e_{\hat{a}_m^t}$; thus, p_m^t is naturally truthful and thus $q_m^t = e_{\hat{a}_m^t}$ for all m is the solution output from LP (8) that achieves an objective value of 0. On $\mathcal{E}_t$, for every m, we have

$$\begin{aligned}
\langle \theta^*, \phi(m, a_m^*)\rangle &\le \left\langle \hat{\theta}_t, \phi(m, a_m^*)\right\rangle + \alpha_t\|\phi(m,a_m^*)\|_{V_t^{-1}} \\
&\le \left\langle \hat{\theta}_t, \phi(m, b_{m,t})\right\rangle + \alpha_t\|\phi(m,b_{m,t})\|_{V_t^{-1}},
\end{aligned}$$

where the second inequality follows from the definition of $b_{m,t}$. Moreover,

$$\left\langle \theta^*, \phi(m, \hat{a}_m^t)\right\rangle \ge \left\langle \hat{\theta}_t, \phi(m, \hat{a}_m^t)\right\rangle - \alpha_t\|\phi(m,\hat{a}_m^t)\|_{V_t^{-1}}.$$

Subtracting the two inequalities gives

$$\Delta_{m,\hat{a}_m^t} \le U_{m,t}. \tag{64}$$

Consequently, on $\mathcal{B}_t^c \cap \mathcal{E}$,

$$\Delta_{m,\hat{a}_m^t} \le U_{m,t} \le \varepsilon_t/\sqrt{\beta_m} \qquad \forall m.$$

Then,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\mathcal{B}_t^c, \mathcal{E}\}\Delta_{I_t,A_t}\right] \le \sum_{t=1}^{T}\sum_{m=1}^{M} \sqrt{\beta_m}\varepsilon_t. \tag{65}$$

Putting them together, we have

$$\begin{aligned}
R_T &\le \frac{8\alpha_T^2 L\max\{1,\lambda^{-1}\}}{\varepsilon_T^2} d\log\left(1+\frac{T}{\lambda d}\right) + \sum_{t=1}^{T}\varepsilon_t\sum_{m=1}^{M}\sqrt{\beta_m} + \mathbb{P}(\mathcal{E}^c)T\Delta_{\max} \\
&\le \frac{8\alpha_T^2 L\max\{1,\lambda^{-1}\}}{\varepsilon_T^2} d\log\left(1+\frac{T}{\lambda d}\right) + \sum_{t=1}^{T}\varepsilon_t\sum_{m=1}^{M}\sqrt{\beta_m} + L,
\end{aligned}$$

where the last line uses Theorem 2 of Abbasi-yadkori et al. (2011).

Since

$$
\begin{aligned}
&(\sum_{m=1}^{M}\sqrt{\beta_m})\sum_{t=1}^{T}\varepsilon_t \\
&\le (\sum_{m=1}^{M}\sqrt{\beta_m})\sum_{t=1}^{T}(8\alpha_T^2 L\max\{1,\lambda^{-1}\}d\log\left(1+\frac{T}{\lambda d}\right)/(\sum_{m=1}^{M}\sqrt{\beta_m}t))^{1/3} \\
&\le \frac{3}{2}(8\alpha_T^2 L\max\{1,\lambda^{-1}\}d\log\left(1+\frac{T}{\lambda d}\right))^{1/3}T^{2/3}(\sum_{m=1}^{M}\sqrt{\beta_m})^{2/3},
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{8\alpha_T^2 L\max\{1,\lambda^{-1}\}}{\varepsilon_T^2}d\log\left(1+\frac{T}{\lambda d}\right) \\
&=\left(8\alpha_T^2 L\max\{1,\lambda^{-1}\}d\log\left(1+\frac{T}{\lambda d}\right)\right)^{1/3}T^{2/3}(\sum_{m=1}^{M}\sqrt{\beta_m})^{2/3},
\end{aligned}
$$

and $\sum_{m=1}^{M}\sqrt{\beta_m}\le\sqrt{M}$, we obtain the worst-case regret upper bound.

This proves (15).

Bayesian regret. Under the Bayesian setting, the definition of the good event becomes

$$
\mathcal{E}_t := \Big\{\big|\langle\theta^*-\hat{\theta}_t,\phi(m,a)\rangle\big|\le\alpha_T'\|\phi(m,a)\|_{\Sigma_t^{-1}},\quad\forall(m,a)\in\mathcal{M}\times\mathcal{K}\Big\}.
$$

At any time t, we have $\Pr(\mathcal{E}_t^c|\mathcal{F}_t)<1/T^2$ by Lemma 6. We then slightly rewrite the regret decomposition in (62) as follows:

$$
\mathrm{BR}_T(\pi^{\text{CGT-UCB}})=\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{B}_t,\mathcal{E}_t\}\Delta_{I_t,A_t}+\mathbf{1}\{\mathcal{B}_t^c,\mathcal{E}_t\}\Delta_{I_t,A_t}+\mathbf{1}\{\mathcal{E}_t^c\}\Delta_{I_t,A_t}\right]. \tag{66}
$$

For the first term, we have

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{B}_t,\mathcal{E}_t\}\Delta_{I_t,A_t}\right]&=\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{\mathcal{B}_t,\mathcal{E}_t\}\Delta_{I_t,A_t}|\mathcal{F}_t\right]\right] \\
&\le\mathbb{E}\left[\sum_{t=1}^{T}\frac{4(\alpha_T')^2}{\varepsilon_t^2}\mathbb{E}\Big[\Delta_{\max}\,\|\phi(I_t,A_t)\|^2_{\Sigma_t^{-1}}\Big|\,\mathcal{F}_t\Big]\right] \\
&\le\frac{4(\alpha_T')^2}{\varepsilon_T^2}\mathbb{E}\left[\sum_{t=1}^{T}\Delta_{\max}\|\phi(I_t,A_t)\|^2_{\Sigma_t^{-1}}\right] \\
&\le\frac{8(\alpha_T')^2 d\max\{\sigma^2,\lambda_{\max}(\Sigma_0^{-1})\}}{\varepsilon_T^2}\log\left(\frac{\mathrm{tr}(\Sigma_0)+T/\sigma^2}{d\,\det(\Sigma_0)^{1/d}}\right)\mathbb{E}_{\theta^*}[\Delta_{\max}],
\end{aligned}
$$

For the second term, we still have

$$
\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{B}_t^c,\mathcal{E}_t\}\Delta_{I_t,A_t}\right]\le\sum_{t=1}^{T}\sum_{m=1}^{M}\sqrt{\beta_m}\varepsilon_t. \tag{67}
$$

For the third term, we have

$$
\sum_{t=1}^{T}\mathbb{E}[\mathbf{1}\{\mathcal{E}_t^c\}\Delta_{I_t,A_t}]\le\sum_{t=1}^{T}\sqrt{\mathbb{E}[\mathbf{1}\{\mathcal{E}_t^c\}]}\sqrt{\mathbb{E}[\Delta_{\max}^2]}\le\sqrt{\mathbb{E}[\Delta_{\max}^2]}.
$$

Putting them together, we have

$$
\begin{aligned}
&\mathrm{BR}_T(\pi^{\text{CGT-UCB}}) \\
&\le\frac{8(\alpha_T')^2 d\max\{\sigma^2,\lambda_{\max}(\Sigma_0^{-1})\}}{\varepsilon_T^2}\log\left(\frac{\mathrm{tr}(\Sigma_0)+T/\sigma^2}{d\,\det(\Sigma_0)^{1/d}}\right)\mathbb{E}_{\theta^*}[\Delta_{\max}] \\
&+\sum_{t=1}^{T}\sum_{m=1}^{M}\sqrt{\beta_m}\varepsilon_t+\sqrt{\mathbb{E}[\Delta_{\max}^2]}.
\end{aligned} \tag{68}
$$

Substituting the value of ε_t completes the proof.

B.2 Proof of Theorem 5

Use the events $\mathcal{B}_t$, $\mathcal{E}_t$, and $\mathcal{E} = \bigcap_{t=1}^{T} \mathcal{E}_t$ from Appendix B.1.

Define

$$t_0 := \max\left\{t \le T : \varepsilon_t/\sqrt{\beta_{\min}} \ge \Delta_{\min}\right\},$$

with the convention $t_0 = 0$ if the set is empty.

We first control the prefix $t \le t_0$. Suppose $t_0 \ge 1$. Using the proof in Appendix B.1, we obtain

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{t_0}\big(\mathbf{1}\{\mathcal{B}_t, \mathcal{E}\} + \mathbf{1}\{\mathcal{B}_t^c, \mathcal{E}\}\big)\Delta_{I_t,A_t}\right]\\ &\le \frac{8\alpha_{t_0}^2 L \max\{1,\lambda^{-1}\}}{\varepsilon_{t_0}^2} d\log\left(1+\frac{t_0}{\lambda d}\right) + \sum_{t=1}^{t_0}\varepsilon_t \sum_{m=1}^{M}\sqrt{\beta_m}.\end{aligned}$$

Since $\varepsilon_{t_0}/\sqrt{\beta_{\min}} \ge \Delta_{\min}$,

$$\frac{8dL\alpha_{t_0}^2 \max\{1,\lambda^{-1}\}\log\left(1+\frac{t_0}{\lambda d}\right)}{\varepsilon_{t_0}^2} \le \frac{8dL\alpha_{t_0}^2 \max\{1,\lambda^{-1}\}\log\left(1+\frac{t_0}{\lambda d}\right)}{\Delta_{\min}^2\beta_{\min}}.$$

Moreover,

$$\varepsilon_{t_0} = \left(\frac{8dL\alpha_T^2 \max\{1,\lambda^{-1}\}\log\left(1+\frac{T}{\lambda d}\right)}{t_0\sum_{m=1}^{M}\sqrt{\beta_m}}\right)^{1/3} \ge \Delta_{\min}\sqrt{\beta_{\min}}$$

implies

$$t_0 \le \frac{8dL\alpha_T^2 \max\{1,\lambda^{-1}\}\log\left(1+\frac{T}{\lambda d}\right)}{\beta_{\min}^{3/2}\Delta_{\min}^3\sum_{m=1}^{M}\sqrt{\beta_m}}.$$

Consequently,

$$\begin{aligned}\sum_{t=1}^{t_0}\varepsilon_t\left(\sum_{m=1}^{M}\sqrt{\beta_m}\right) &\le \frac{3}{2}\left(8dL\alpha_T^2\max\{1,\lambda^{-1}\}\log\left(1+\frac{T}{\lambda d}\right)\right)^{1/3} t_0^{2/3}\left(\sum_{m=1}^{M}\sqrt{\beta_m}\right)^{2/3}\\ &\le \frac{12dL\alpha_T^2\max\{1,\lambda^{-1}\}\log\left(1+\frac{T}{\lambda d}\right)}{\Delta_{\min}^2\beta_{\min}}.\end{aligned}$$

Thus, using $t_0 \le T$ and $\alpha_{t_0} \le \alpha_T$,

$$\mathbb{E}\left[\sum_{t=1}^{t_0}\mathbf{1}\{\mathcal{E}\}\Delta_{I_t,A_t}\right] \le \frac{20dL\alpha_T^2\max\{1,\lambda^{-1}\}\log\left(1+\frac{T}{\lambda d}\right)}{\Delta_{\min}^2\beta_{\min}}. \tag{69}$$

The same bound holds trivially when $t_0 = 0$, since the sum is empty.

It remains to control $t > t_0$. On this interval,

$$\varepsilon_t < \Delta_{\min}\sqrt{\beta_m},\ m \in \mathcal{M}.$$

By (64), on $\mathcal{B}_t^c \cap \mathcal{E}$,

$$0 \le \Delta_{m,\hat{a}_m^t} \le U_{m,t} \le \varepsilon_t/\sqrt{\beta_m} < \Delta_{\min}.$$

By the definition of $\Delta_{\min}$, this implies

$$\hat{a}_m^t = a_m^* \qquad \forall m.$$

On $\mathcal{B}_t^c$, the proposal menu is estimated-optimal, so $q^t = p^t$. Hence $\mathbb{E}[\mathbf{1}\{\mathcal{B}_t^c, \mathcal{E}, t > t_0\}\Delta_{I_t,A_t}] = 0$. According to (62), we next control $\mathbb{E}[\mathbf{1}\{\mathcal{B}_t, \mathcal{E}, t > t_0\}\Delta_{I_t,A_t}]$.

Fix a history $\mathcal{F}_t$. By Lemma 4, when $\mathbb{E}[\mathbf{1}\{\mathcal{B}_t, \mathcal{E}_t, t > t_0\}\Delta_{I_t,A_t} \mid \mathcal{F}_t] > 0$, we have

$$\alpha_t\|\phi(m, b_{m,t})\|_{V_t^{-1}} \ge \frac{\Delta_{\min}}{2}$$

for some $m \in \mathcal{M}_t$. Then, we have

$$\begin{aligned}
\frac{\Delta_{\min}^2}{4} \leq & \alpha_t^2 \|\phi(m, b_{m,t})\|_{V_t^{-1}}^2 \\
= & \sum_a \alpha_t^2 p_m^t(a) \|\phi(m,a)\|_{V_t^{-1}}^2 \\
\leq & M \sum_a \alpha_t^2 \pi_a^{\text{CGT-UCB}}(m|\mathcal{F}_t) \|\phi(m,a)\|_{V_t^{-1}}^2 \\
\leq & M\alpha_t^2 \sum_{m \in \mathcal{M}} \sum_a \pi_a^{\text{CGT-UCB}}(m|\mathcal{F}_t) \|\phi(m,a)\|_{V_t^{-1}}^2 \\
\leq & \alpha_t^2 M \sum_{m \in \mathcal{M}} \sum_a \frac{\beta_m}{\beta_{\min}} \pi_a^{\text{CGT-UCB}}(m|\mathcal{F}_t) \|\phi(m,a)\|_{V_t^{-1}}^2 \\
= & \frac{\alpha_t^2 M}{\beta_{\min}} \mathbb{E}[\|\phi(I_t, A_t)\|_{V_t^{-1}}^2 | \mathcal{F}_t],
\end{aligned}$$

where the first inequality after the equality follows from constraint (10) and $|\mathcal{M}_t| \leq M$.

Therefore, using $\mathcal{E} \subseteq \mathcal{E}_t$ and conditioning on $\mathcal{F}_t$,

$$\begin{aligned}
& \sum_{t=1}^T \mathbb{E}[\mathbf{1}\{\mathcal{B}_t, \mathcal{E}, t > t_0\} \Delta_{I_t, A_t}] \\
& \leq \sum_{t=1}^T \mathbb{E}\left[\mathbf{1}\left\{\frac{4\alpha_t^2 M}{\beta_{\min}} \mathbb{E}[\|\phi(I_t, A_t)\|_{V_t^{-1}}^2 \mid \mathcal{F}_t] \geq \Delta_{\min}^2\right\} \Delta_{\max}\right] \\
& \leq \sum_{t=1}^T \mathbb{E}\left[\frac{4\alpha_t^2 M}{\Delta_{\min}^2 \beta_{\min}} \mathbb{E}[\|\phi(I_t, A_t)\|_{V_t^{-1}}^2 \mid \mathcal{F}_t] \Delta_{\max}\right] \\
& \leq \frac{LM}{\beta_{\min} \Delta_{\min}^2} 8\alpha_T^2 \max\{1, \lambda^{-1}\} d \log\left(1 + \frac{T}{\lambda d}\right), \qquad (70)
\end{aligned}$$

where the last inequality uses $\Delta_{\max} \leq L$, the tower property, and the elliptical-potential bound (49).

Finally, adding the term $\mathbb{E}\left[\sum_{t=1}^T \mathbf{1}\{\mathcal{E}^c\} \Delta_{I_t, A_t}\right] \leq L$, we obtain

$$R_T(\pi^{\text{CGT-UCB}}; \nu) \leq \frac{20L + 8LM}{\beta_{\min} \Delta_{\min}^2} \alpha_T^2 \max\{1, \lambda^{-1}\} d \log\left(1 + \frac{T}{\lambda d}\right) + L,$$

completing the proof.

Lemma 3. *In the frequentist setting, take $\rho_t = \alpha_t$ and $\varepsilon_t > 0$. Define the following events in Algorithm 1:*

$$\mathcal{B}_t := \left\{\text{there exists } m \in \mathcal{M} \text{ such that } U_{m,t} \geq \varepsilon_t / \sqrt{\beta_m}\right\},$$

$$\mathcal{E}_t := \left\{\left|\langle \theta^* - \hat{\theta}_t, \phi(m,a)\rangle\right| \leq \alpha_t \|\phi(m,a)\|_{V_t^{-1}}, \quad \forall (m,a) \in \mathcal{M} \times \mathcal{K}\right\}.$$

Then

$$\mathbf{1}\{\mathcal{B}_t, \mathcal{E}_t\} \leq \frac{4\alpha_t^2}{\varepsilon_t^2} \mathbb{E}\left[\|\phi(I_t, A_t)\|_{V_t^{-1}}^2 \,\middle|\, \mathcal{F}_t\right]. \qquad (71)$$

Moreover, consider the following event in Algorithm 3:

$$\mathcal{B}_t := \left\{\text{there exists } m \in \mathcal{M} \text{ such that } U_{m,t} \geq \varepsilon_t\right\}.$$

We have

$$\mathbf{1}\{\mathcal{B}_t, \mathcal{E}_t\} \leq \frac{4\alpha_t^2}{\beta_{\min} \varepsilon_t^2} \mathbb{E}\left[\|\phi(I_t, A_t)\|_{V_t^{-1}}^2 \,\middle|\, \mathcal{F}_t\right]. \qquad (72)$$

Proof of Lemma 3. We first show that, for every $m \in \mathcal{M}$,

$$U_{m,t} \leq 2\alpha_t \|\phi(m, b_{m,t})\|_{V_t^{-1}}. \qquad (73)$$

When $b_{m,t} = \hat{a}_m^t$, this follows immediately from the definition of $U_{m,t}$. Otherwise,

$$\begin{aligned}\left\langle \hat{\theta}_t, \phi(m, b_{m,t}) \right\rangle + \alpha_t \|\phi(m, b_{m,t})\|_{V_t^{-1}} &\geq \left\langle \hat{\theta}_t, \phi(m, \hat{a}_m^t) \right\rangle + \alpha_t \|\phi(m, \hat{a}_m^t)\|_{V_t^{-1}} \\ &\geq \left\langle \hat{\theta}_t, \phi(m, b_{m,t}) \right\rangle + \alpha_t \|\phi(m, \hat{a}_m^t)\|_{V_t^{-1}}.\end{aligned}$$

Consequently,

$$\alpha_t \|\phi(m, b_{m,t})\|_{V_t^{-1}} \geq \alpha_t \|\phi(m, \hat{a}_m^t)\|_{V_t^{-1}}.$$

Using this inequality together with $\langle \hat{\theta}_t, \phi(m, \hat{a}_m^t) \rangle \geq \langle \hat{\theta}_t, \phi(m, b_{m,t}) \rangle$, we obtain

$$\begin{aligned}U_{m,t} &= \left\langle \hat{\theta}_t, \phi(m, b_{m,t}) \right\rangle + \alpha_t \|\phi(m, b_{m,t})\|_{V_t^{-1}} - \left\langle \hat{\theta}_t, \phi(m, \hat{a}_m^t) \right\rangle + \alpha_t \|\phi(m, \hat{a}_m^t)\|_{V_t^{-1}} \\ &\leq 2\alpha_t \|\phi(m, b_{m,t})\|_{V_t^{-1}},\end{aligned}$$

which proves (73).

Now suppose that $\mathcal{B}_t$ holds, and choose any $m \in \mathcal{M}_t$, so that $U_{m,t} \geq \varepsilon_t / \sqrt{\beta_m}$. By (73),

$$\varepsilon_t \leq \sqrt{\beta_m} U_{m,t} \leq 2\sqrt{\beta_m} \alpha_t \|\phi(m, b_{m,t})\|_{V_t^{-1}}. \tag{74}$$

Since $p_m^t = e_{b_{m,t}}$ for such an m, squaring (74) gives

$$\varepsilon_t^2 \leq 4\alpha_t^2 \beta_m \|\phi(m, b_{m,t})\|_{V_t^{-1}}^2 = 4\alpha_t^2 \beta_m \sum_a p_m^t(a) \|\phi(m, a)\|_{V_t^{-1}}^2. \tag{75}$$

Since the above holds for any $m \in \mathcal{M}_t$, we have

$$\begin{aligned}\varepsilon_t^2 &\leq \frac{1}{|\mathcal{M}_t|} \sum_{m' \in \mathcal{M}_t} 4\alpha_t^2 \beta_{m'} \sum_a p_{m'}^t(a) \|\phi(m', a)\|_{V_t^{-1}}^2 \\ &\leq 4\alpha_t^2 \sum_{m' \in \mathcal{M}_t} \beta_{m'} \sum_a \pi_a^{\text{CGT-UCB}}(m' | \mathcal{F}_t) \|\phi(m', a)\|_{V_t^{-1}}^2,\end{aligned}$$

where the second inequality is from constraint (10) in the LP. Under truthful reporting, conditionally on $\mathcal{F}_t$, I_t has distribution $(\beta_m)_{m \in \mathcal{M}}$; conditionally on $\mathcal{F}_t$ and $I_t = m'$, A_t has distribution $\pi^{\text{CGT-UCB}}(m' | \mathcal{F}_t)$. Hence

$$\sum_{m' \in \mathcal{M}} \beta_{m'} \sum_a \pi_a^{\text{CGT-UCB}}(m' | \mathcal{F}_t) \|\phi(m', a)\|_{V_t^{-1}}^2 = \mathbb{E}\Big[\|\phi(I_t, A_t)\|_{V_t^{-1}}^2 \Big| \mathcal{F}_t \Big].$$

Therefore, on $\mathcal{B}_t$, we have

$$\mathbf{1}\{\mathcal{B}_t\} \leq 1 \leq \frac{4\alpha_t^2}{\varepsilon_t^2} \mathbb{E}\Big[\|\phi(I_t, A_t)\|_{V_t^{-1}}^2 \Big| \mathcal{F}_t \Big],$$

On $\mathcal{B}_t^c$, the indicator is zero. Thus the bound holds for $\mathbf{1}\{\mathcal{B}_t\}$, and hence also for $\mathbf{1}\{\mathcal{B}_t, \mathcal{E}_t\}$, proving (71).

For the second part, $\mathcal{M}_t = \{m \in \mathcal{M} : U_{m,t} \geq \varepsilon_t\}$. On $\mathcal{B}_t$, this set is nonempty, and for every $m \in \mathcal{M}_t$, we have

$$\begin{aligned}&\varepsilon_t^2 \leq 4\alpha_t^2 \|\phi(m, b_{m,t})\|_{V_t^{-1}}^2 = 4\alpha_t^2 \sum_a p_m^t(a) \|\phi(m, a)\|_{V_t^{-1}}^2 \implies \\ &\varepsilon_t^2 \leq \frac{1}{|\mathcal{M}_t| \beta_{\min}} \sum_{m' \in \mathcal{M}_t} 4\alpha_t^2 \beta_{m'} \sum_a p_{m'}^t(a) \|\phi(m', a)\|_{V_t^{-1}}^2 \\ &\leq \frac{4\alpha_t^2}{\beta_{\min}} \sum_{m' \in \mathcal{M}} \beta_{m'} \sum_a \pi_a^{\text{CGT-UCB}}(m' | \mathcal{F}_t) \|\phi(m', a)\|_{V_t^{-1}}^2.\end{aligned}$$

Therefore, on $\mathcal{B}_t$, we have

$$\mathbf{1}\{\mathcal{B}_t\} \leq 1 \leq \frac{4\alpha_t^2}{\beta_{\min} \varepsilon_t^2} \mathbb{E}\Big[\|\phi(I_t, A_t)\|_{V_t^{-1}}^2 \Big| \mathcal{F}_t \Big],$$

As above, the bound is trivial on $\mathcal{B}_t^c$, proving (72). □

Lemma 4. *When $\mathbb{E}[\mathbf{1}\{\mathcal{B}_t, \mathcal{E}_t, t > t_0\} \Delta_{I_t, A_t} \mid \mathcal{F}_t] > 0$, there exists a type m such that*

$$\alpha_t \|\phi(m, b_{m,t})\|_{V_t^{-1}} \geq \frac{\Delta_{\min}}{2} \quad \textit{and} \quad U_{m,t} \geq \frac{\varepsilon_t}{\sqrt{\beta_m}}. \tag{76}$$

Proof. Fix such a history $\mathcal{F}_t$, so that $t > t_0$ and $\mathcal{B}_t \cap \mathcal{E}_t$ holds. Define

$$\mathcal{M}_t := \{m \in \mathcal{M} : U_{m,t} \geq \varepsilon_t/\sqrt{\beta_m}\}.$$

Consider $\mathbb{E}[\Delta_{I_t,A_t} \mid \mathcal{F}_t] = \sum_{m,a} \beta_m q_m^t(a)\Delta_{m,a}$. We claim that if $\mathbb{E}[\mathbf{1}\{\mathcal{B}_t, \mathcal{E}_t, t > t_0\}\Delta_{I_t,A_t} \mid \mathcal{F}_t] > 0$, there exists an $m \in \mathcal{M}_t$ such that $\{\hat{a}_m^t \neq a_m^*\} \cup \{\hat{a}_m^t = a_m^*, b_{m,t} \neq \hat{a}_m^t\}$ holds. Otherwise, $\hat{a}_m^t = a_m^* = b_{m,t}$ for all $m \in \mathcal{M}_t$. Since $a_{m,t}^{\text{prop}} = \hat{a}_m^t$ for all $m \notin \mathcal{M}_t$, the entire proposal menu is estimated-optimal and feasible, implying $q^t = p^t$, and

$$\begin{aligned}\mathbb{E}[\Delta_{I_t,A_t} \mid \mathcal{F}_t] &= \sum_{m,a} \beta_m q_m^t(a)\Delta_{m,a} \\ &= \sum_m \beta_m \Delta_{m,\hat{a}_m^t} \\ &= \sum_{m\in\mathcal{M}_t} \beta_m \Delta_{m,\hat{a}_m^t} + \sum_{m\notin\mathcal{M}_t} \beta_m \Delta_{m,\hat{a}_m^t}.\end{aligned}$$

The first term is zero since $\hat{a}_m^t = a_m^* = b_{m,t}$ for all $m \in \mathcal{M}_t$. In the second summation, we have $\Delta_{m,\hat{a}_m^t} \leq U_{m,t} < \varepsilon_t/\sqrt{\beta_m} < \Delta_{\min}$ for each $m \notin \mathcal{M}_t$, which implies that the second term is also zero, forming a contradiction.

Then, we show that when $\{\hat{a}_m^t \neq a_m^*\} \cup \{\hat{a}_m^t = a_m^*, b_{m,t} \neq \hat{a}_m^t\}$ holds, we must have $\Delta_{\min} \leq 2\alpha_t \|\phi(m, b_{m,t})\|_{V_t^{-1}}$ for such m.

If

$$b_{m,t} \neq a_m^*,$$

then $b_{m,t}$ is suboptimal and hence

$$\Delta_{m,b_{m,t}} \geq \Delta_{\min}.$$

On $\mathcal{E}_t$, the optimism of the UCB action gives

$$\begin{aligned}\mu_{m,a_m^*} &\leq \mathrm{UCB}_{m,a_m^*}(t) \\ &\leq \mathrm{UCB}_{m,b_{m,t}}(t) \\ &\leq \mu_{m,b_{m,t}} + 2\alpha_t \|\phi(m, b_{m,t})\|_{V_t^{-1}}.\end{aligned}$$

Consequently,

$$\Delta_{\min} \leq \Delta_{m,b_{m,t}} \leq 2\alpha_t \|\phi(m, b_{m,t})\|_{V_t^{-1}}.$$

Suppose instead that

$$\hat{a}_m^t \neq a_m^*.$$

Therefore,

$$\Delta_{m,\hat{a}_m^t} \geq \Delta_{\min}.$$

By the greedy simple-regret certificate,

$$\Delta_{m,\hat{a}_m^t} \leq U_{m,t} \leq 2\alpha_t \|\phi(m, b_{m,t})\|_{V_t^{-1}}.$$

Thus,

$$\Delta_{\min} \leq 2\alpha_t \|\phi(m, b_{m,t})\|_{V_t^{-1}}.$$

□

B.3 Algorithm 1 for unknown $\boldsymbol{\beta}$

For unknown $\boldsymbol{\beta}$, we revise Algorithm 1 to Algorithm 3. Then, the resulting regret becomes the following.

Theorem 10. *In the frequentist setting, if we choose*

$$\rho_t = \alpha_t := \sigma\sqrt{2\log T + d\log\left(1 + \frac{t}{\lambda d}\right)} + \sqrt{\lambda}L,$$

$$\varepsilon_t = \left(\frac{8\alpha_T^2 L \max\{1, \lambda^{-1}\} d \log\left(1 + \frac{T}{\lambda d}\right)}{t}\right)^{1/3},$$

Algorithm 3 Confidence-Gated Truthful UCB (CGT-UCB) with unknown β

Require: Feature vectors $\{\phi(m,a) : m \in \mathcal{M}, a \in \mathcal{K}\}$; regularization parameter $\lambda > 0$; confidence radii $(\rho_t)_{t\geq 1}$; resolution schedule $(\varepsilon_t)_{t\geq 1}$.

Initialize $(\hat{\theta}_1, W_1) \leftarrow (0_d, \lambda I_d)$ (frequentist) or $(\hat{\theta}_1, W_1) \leftarrow (\mu_0, \Sigma_0)$ (Bayesian).
for $t = 1, \ldots, T$ **do**
Phase 1: Construct proposal policy
  for each type $m \in \mathcal{M}$ **do**
    Compute the estimated-optimal action $\hat{a}_m^t \in \arg\max_{a\in\mathcal{K}} \left\langle \hat{\theta}_t, \phi(m,a) \right\rangle$.
    Compute the exploration action

$$b_{m,t} \in \arg\max_{a\in\mathcal{K}} \left\{ \left\langle \hat{\theta}_t, \phi(m,a) \right\rangle + \rho_t \|\phi(m,a)\|_{W_t^{-1}} \right\}.$$

    Compute the greedy simple-regret certificate

$$U_{m,t} \leftarrow \left\langle \hat{\theta}_t, \phi(m,b_{m,t}) - \phi(m,\hat{a}_m^t) \right\rangle + \rho_t \left(\|\phi(m,b_{m,t})\|_{W_t^{-1}} + \|\phi(m,\hat{a}_m^t)\|_{W_t^{-1}} \right). \quad (77)$$

    if $U_{m,t} \geq \varepsilon_t$ **then**
      Set the proposal action to the exploration action: $a_{m,t}^{\text{prop}} \leftarrow b_{m,t}$.
    else
      Set the proposal action to the estimated-optimal action: $a_{m,t}^{\text{prop}} \leftarrow \hat{a}_m^t$.
    end if
    Define the deterministic proposal distribution $p_m^t \leftarrow e_{a_{m,t}^{\text{prop}}} \in \Delta_K$.
  end for

Phase 2: Construct CGT-UCB policy
  Define $\mathcal{M}_t := \{m \in \mathcal{M} : U_{m,t} \geq \varepsilon_t\}$. Compute $q^t = (q_m^t)_{m\in\mathcal{M}}$ by solving

$$\begin{aligned}
\underset{(q_m^t)_{m\in\mathcal{M}}}{\text{minimize}} \quad & \sum_{m\in\mathcal{M}} \|q_m^t - p_m^t\|_1 && (78)\\
\text{subject to} \quad & (q_i^t - q_j^t)^\top \Phi_i \hat{\theta}_t \geq 0, \qquad \forall i,j \in \mathcal{M}, && (79)\\
& q_m^t(b_{m,t}) \geq \frac{1}{|\mathcal{M}_t|} p_m^t(b_{m,t}), \qquad \forall m \in \mathcal{M}_t, && (80)\\
& q_m^t \in \Delta_K, \ \forall m \in \mathcal{M}. && (81)
\end{aligned}$$

  Omit (80) if $\mathcal{M}_t = \varnothing$.
  $\pi^{\text{CGT-UCB}}(m|\mathcal{F}_t) \leftarrow q_m^t$, for all $m \in \mathcal{M}$.
  Observe the reported type $I_t' \in \mathcal{M}$, and sample $A_t \sim \pi^{\text{CGT-UCB}}(I_t'|\mathcal{F}_t)$.
  Observe the reward r_t, and update $\hat{\theta}_{t+1}$ and W_{t+1} using (1) (frequentist) or (2) (Bayesian).
end for

then the policy $\pi^{\text{CGT-UCB}}$ is strategyproof and satisfies

$$R_T(\pi^{\text{CGT-UCB}}) \leq \left(\frac{2}{\beta_{\min}} + 3 \right) \left(dL\alpha_T^2 \max\{1, \lambda^{-1}\} \log\left(1 + \frac{T}{\lambda d}\right) \right)^{1/3} T^{2/3} + L. \quad (82)$$

Proof. Decomposing the regret as in (62), we deal with the three terms separately. Following a derivation similar to that of (63) and using Lemma 3, we obtain the following upper bound for the first term:

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\mathcal{B}_t, \mathcal{E}_t\} \Delta_{I_t, A_t} \right] \leq \Delta_{\max} \frac{8\alpha_T^2 \max\{1, \lambda^{-1}\}}{\beta_{\min} \varepsilon_T^2} d \log\left(1 + \frac{T}{\lambda d}\right). \quad (83)$$

For the second term, on $\mathcal{B}_t^c$, the proposal menu is estimated-optimal and hence $q^t = p^t$. On $\mathcal{B}_t^c \cap \mathcal{E}$, we have

$$\Delta_{m,\hat{a}_m^t} \leq U_{m,t} \leq \varepsilon_t \qquad \forall m.$$

Then,

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{B}_t^c,\mathcal{E}\}\Delta_{I_t,A_t}\right]\le\sum_{t=1}^{T}\sum_{m=1}^{M}\beta_m\varepsilon_t=\sum_{t=1}^{T}\varepsilon_t. \tag{84}$$

Putting them together, we have

$$\begin{aligned}R_T&\le\frac{8\alpha_T^2L\max\{1,\lambda^{-1}\}}{\beta_{\min}\varepsilon_T^2}d\log\left(1+\frac{T}{\lambda d}\right)+\sum_{t=1}^{T}\varepsilon_t+\mathbb{P}(\mathcal{E}^c)T\Delta_{\max}\\&\le\frac{8\alpha_T^2L\max\{1,\lambda^{-1}\}}{\beta_{\min}\varepsilon_T^2}d\log\left(1+\frac{T}{\lambda d}\right)+\sum_{t=1}^{T}\varepsilon_t+L.\end{aligned}$$

Substituting the value of ε_t, we obtain the results.

□

C Technical Lemmas

Lemma 5. *Let $x_1,\dots,x_t\in\mathbb{R}^d$ be independent random vectors supported on a finite set $\{y_1,\dots,y_k\}$. Suppose that*

$$\mathbb{E}[x_sx_s^\top]=\Sigma\succeq 0,\qquad s\in[t],$$

and let $r:=\operatorname{rank}(\Sigma)\ge 1$. Define $X_s:=x_sx_s^\top$. Then, for every $\delta\in[0,1)$,

$$\Pr\left(\sum_{s=1}^{t}X_s\not\succeq(1-\delta)t\Sigma\right)\le r\exp\left(-\frac{\delta^2t}{2\max_{i\in[k]}\|y_i\|_{\Sigma^+}^2}\right), \tag{85}$$

where Σ^+ denotes the Moore–Penrose pseudoinverse of Σ.

Proof of Lemma 5. Let

$$\mathcal{U}:=\operatorname{range}(\Sigma),\qquad P_{\mathcal{U}}:=\Sigma^{+/2}\Sigma\Sigma^{+/2},$$

where $P_{\mathcal{U}}$ is the orthogonal projection onto $\mathcal{U}$.

We first observe that $x_s\in\mathcal{U}$ almost surely. Indeed, for every $v\in\ker(\Sigma)$,

$$0=v^\top\Sigma v=\mathbb{E}\left[(v^\top x_s)^2\right].$$

Therefore, $v^\top x_s=0$ almost surely for every $v\in\ker(\Sigma)$, which implies

$$x_s\in\ker(\Sigma)^\perp=\operatorname{range}(\Sigma)=\mathcal{U}$$

almost surely. In particular, every point y_i in the support of x_s belongs to $\mathcal{U}$.

Define the whitened random matrices

$$Z_s:=\frac{1}{t}\Sigma^{+/2}X_s\Sigma^{+/2}.$$

Restricted to the r-dimensional subspace $\mathcal{U}$, these matrices are independent and positive semidefinite. Moreover,

$$\sum_{s=1}^{t}\mathbb{E}[Z_s]=\Sigma^{+/2}\Sigma\Sigma^{+/2}=P_{\mathcal{U}}. \tag{86}$$

Since $P_{\mathcal{U}}$ acts as the identity on $\mathcal{U}$, we have

$$\lambda_{\min}\left(\sum_{s=1}^{t}\mathbb{E}[Z_s]\Big|_{\mathcal{U}}\right)=1.$$

For every realization $x_s=y_i$,

$$\begin{aligned}\lambda_{\max}(Z_s)&=\frac{1}{t}x_s^\top\Sigma^+x_s\\&\le\frac{1}{t}\max_{i\in[k]}\|y_i\|_{\Sigma^+}^2.\end{aligned} \tag{87}$$

Applying the matrix Chernoff inequality on $\mathcal{U}$, whose dimension is r, gives

$$\Pr\left(\lambda_{\min}\left(\sum_{s=1}^{t} Z_s\Big|_{\mathcal{U}}\right) \le 1-\delta\right) \le r\exp\left(-\frac{\delta^2 t}{2\max_{i\in[k]}\|y_i\|_{\Sigma^+}^2}\right). \tag{88}$$

Finally, since X_s and Σ are both supported on $\mathcal{U}$,

$$\sum_{s=1}^{t} X_s \succeq (1-\delta)t\Sigma$$

if and only if

$$\left(\sum_{s=1}^{t} Z_s\right)\Big|_{\mathcal{U}} \succeq (1-\delta)I_{\mathcal{U}}.$$

Consequently,

$$\left\{\sum_{s=1}^{t} X_s \not\succeq (1-\delta)t\Sigma\right\} \subseteq \left\{\lambda_{\min}\left(\sum_{s=1}^{t} Z_s\Big|_{\mathcal{U}}\right) \le 1-\delta\right\}.$$

Combining this inclusion with (88) proves the result. □

Lemma 6. *Suppose that, conditional on $\mathcal{F}_t$, the posterior distribution satisfies*

$$\theta^* \mid \mathcal{F}_t \sim \mathcal{N}\left(\hat{\theta}_t, \Sigma_t^{-1}\right),$$

where $\Sigma_t \succ 0$ is the posterior precision matrix in (2)*. Then, almost surely,*

$$\Pr\left(\left|\left\langle u, \hat{\theta}_t - \theta^*\right\rangle\right| \le \sqrt{d + 2\sqrt{d\log(1/\delta)} + 2\log(1/\delta)}\|u\|_{\Sigma_t^{-1}} \text{ for every } u \in \mathbb{R}^d \;\middle|\; \mathcal{F}_t\right) \ge 1-\delta. \tag{89}$$

Proof of Lemma 6. By the duality of Mahalanobis norms,

$$\sup_{u\neq 0} \frac{\left|\langle u, \hat{\theta}_t - \theta^*\rangle\right|}{\|u\|_{\Sigma_t^{-1}}} = \|\hat{\theta}_t - \theta^*\|_{\Sigma_t}.$$

Under the assumed Gaussian posterior,

$$\|\hat{\theta}_t - \theta^*\|_{\Sigma_t}^2 \;\Big|\; \mathcal{F}_t \sim \chi_d^2.$$

The Laurent–Massart inequality Laurent & Massart (2000), with $x = \log(1/\delta)$, therefore yields

$$\Pr\left(\|\hat{\theta}_t - \theta^*\|_{\Sigma_t}^2 > d + 2\sqrt{dx} + 2x \;\Big|\; \mathcal{F}_t\right) \le e^{-x} = \delta,$$

which proves the claim. □

D Experimental Details

Evaluation protocol. We compare CGT-UCB with ETC, truthful Thompson sampling (TTS), and revised truthful Thompson sampling (R-TTS). All experiments use $T = 6000$ rounds and 1000 seeds. Within each seed, the methods share the same true-type arrivals and environmental randomness.

Algorithm parameters. We use $\sigma = \lambda = 1$, with $L = 3$ for synthetic experiments and the working bound $L = 2$ for Warfarin. CGT-UCB uses

$$\alpha_t = \sigma\sqrt{2\log T + d\log\left(1 + \frac{t}{\lambda d}\right)} + \sqrt{\lambda}L,$$

$$\varepsilon_t = c_\varepsilon\left(\frac{8dL\alpha_T^2\log\left(1+\frac{T}{\lambda d}\right)}{t\sum_{m=1}^{M}\sqrt{\beta_m}}\right)^{1/3}.$$

We set $c_\varepsilon = (2/3)^{1/6}$ in the dimension experiments, $c_\varepsilon = 0.7$ in the worst-case experiment, and $c_\varepsilon = 0.85$ in the Warfarin experiment. ETC explores uniformly for $T_0 = \min\{T, \lceil T^{2/3} \rceil\}$ rounds and then commits using its frozen estimate. Both Thompson-sampling methods use 128 Monte Carlo samples per round. Unknown parameter coordinates are initialized at zero, and action ties are resolved by the first action index. Synthetic rewards have independent standard Gaussian noise.

Varying the feature dimension. We fix $M = K = 5$ and vary $d \in \{6, 8, 10\}$. Each type has a private coordinate block, and its action features are supported only on that block. The block sizes are $(1, 1, 1, 1, 1)$, $(2, 2, 2, 1, 1)$, and $(2, 2, 2, 2, 2)$, respectively. Within each block, action directions are obtained by normalizing Gaussian vectors, with radii sampled uniformly from $[0.5, 1]$. Features for types $2, \ldots, 5$ are additionally multiplied by 0.3. The corresponding parameter blocks have random Gaussian directions and norms 1 for type 1 and $\sqrt{1.25/4}$ for each remaining type, giving $\|\theta^*\|_2 = 1.5$. Type-arrival probabilities are $\boldsymbol{\beta} = (0.6, 0.1, 0.1, 0.1, 0.1)$. Thus, increasing d expands the type-specific parameter spaces while keeping the numbers of types and actions fixed.

Worst-case instance. We instantiate the orthogonal information structure from the proof of Theorem 3. We set $d = M = K = 10$ and

$$\phi(m, a) = \begin{cases} e_m, & a = m, \\ 0, & a \neq m. \end{cases}$$

Type 1 arrives with probability $1/2$, and each remaining type with probability $1/18$. The coordinate $\theta_1^* = 1$ is publicly known to all methods and excluded from estimation updates and uncertainty calculations. The remaining coordinates are

$$(\theta_2^*, \ldots, \theta_{10}^*) = \delta(1, -1, 1, -1, 1, -1, 1, -1, -1), \qquad \delta = 0.3434142728.$$

Only a type's diagonal action reveals its unknown coordinate. We keep both the instance and the gap δ fixed throughout the experiment.

Warfarin dose selection. We use 5,528 usable patient records from the Warfarin dataset (Consortium, 2009), split into 3,316 development and 2,212 evaluation records. Using development covariates only, we impute and standardize age, height, weight, and VKORC1 genotype and fit eight clusters. Each type is represented by its cluster centroid with an intercept. Action-specific feature blocks give $M = 8$, $K = 3$, and $d = 15$. Actions correspond to weekly dose classes below 21, between 21 and 49 inclusive, and above 49 mg.

Types arrive according to their empirical evaluation-set frequencies. Within each arriving type, we sample an evaluation patient with replacement. The reward is one if the selected class matches the patient's recorded dose class and zero otherwise; only the selected action's reward is revealed. Regret is measured relative to the best empirical dose class for each type. No offline fitted reward model generates feedback or initializes the learner. The resulting type-level reward means need not be exactly linear in the features. Rewards measure agreement with recorded dose classes rather than clinical efficacy, and strategic reporting follows the paper's behavioral model.