LOW-REGRET LEARNING IN STRATEGIC BANDITS VIA INCENTIVE DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Online platforms often recommend products based on information reported by providers, who are self-interested and may misreport information to maximize their own profits. To help platforms make decisions, we study how the learner can minimize its regret against arms' misreporting features in strategic bandits. We propose the PALinUCB algorithm whose efficacy roots from a delicately designed contract utility. We show that, under PALinUCB, the regret is largely determined by the deviation incentive of the optimal arm. In particular, when the optimal arm behaves appropriately, the resulting joint strategy profile forms a $\mathrm{poly}(\log T)$-approximate Nash equilibrium, where each suboptimal arm is selected only $\mathrm{poly}(\log T)$ times. Moreover, under Nash equilibrium, PALinUCB achieves a $\mathrm{poly}(\log T)$ regret. Further, numerical results corroborate our theoretical findings and indicate that our contract design can substantially shape arms' strategic behavior for the interest of the learner, and PALinUCB remains robust under strategic reporting.

1 INTRODUCTION

Online platforms often recommend products based on contextual information reported by self-interested providers. Since the platform cannot directly observe the true product features, providers may strategically misreport them, creating a conflict of interest between the platform and the providers. More specifically, the platform aims to maximize rewards by recommending high-quality products. However, to improve the recommendation opportunity, providers may use different marketing strategies to shape how their product features are presented to the platform, which may deviate from the true features of the product, as illustrated in Figure 1. We model this game as a strategic contextual bandit problem. At each round, product providers, acting as arms, strategically report contextual features to the platform, acting as the learner. Then the learner selects one arm based on the reported information and receives the corresponding reward.

Recent literature investigates the strategic contextual bandits, including Kleine Buening et al. (2024) and Verma et al. (2025), which respectively designed the OptGTM and COBRA algorithms for regret minimization, but their guarantees apply only under restrictive conditions. In specific, although OptGTM allows arms to misreport their features, its regret analysis assumes that such reports preserve each arm's true mean reward under every context. COBRA, on the other hand, assumes that the UCB estimate computed from each arm's reported feature is always no smaller than the arm's true mean reward. These conditions substantially restrict the strategic reporting behavior that can be analyzed. In this work, we relax these assumptions and study a broader class of strategic reporting strategies.

We also observe that the learner can influence arms' strategic behavior through incentive design. For example, Kleine Buening et al. (2024); Verma et al. (2025) assign a fixed contract of 1 to the selected arm. In our work, we propose a novel contract design as $r_{t,i} - \eta$ where $r_{t,i}$ is the noisy reward of arm i at round t and η is a constant. To illustrate the intuition behind this contract design, consider again a food-delivery platform. The platform may charge a fixed promotion fee whenever a restaurant is recommended, representing the cost of obtaining exposure. The restaurant's payoff from being recommended is then its realized reward minus this fee. As a result, restaurants with poor realized performance may no longer benefit from additional recommendations and may even prefer to avoid being selected. By linking the value of selection to realized performance, this contract design can

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.

weaken arms' incentives to misreport merely to gain exposure and thereby mitigate the effect of strategic reporting on the platform's learning process. Using this contract design, we propose the Per-Arm Linear Upper Confidence Bound (PALinUCB) algorithm. Our theoretical and empirical results show that, compared with existing incentive designs, the proposed contract better aligns the interests of the learner and the arms under information asymmetry.

Incentive design has been extensively studied in economics, where contracts are designed to shape the behavior of self-interested agents. Walton & Carroll (2022) proposed a general framework for robust contracting models, showing how contract structures can influence agents' decisions under strategic interactions. Burkett & Rosenthal (2024) further studied data-driven contract design and demonstrated how contracts can be optimized to improve decision-making under information asymmetry. Motivated by these ideas, we introduce contract design into strategic linear contextual bandits and study how the contract associated with arm selection shapes arms' incentives, strategic behaviors, and the learner's regret.



Figure 1: **Interaction protocol.** Self-interested providers may strategically report product information to improve their recommendation opportunities. The platform selects a product based on the reported information and observes the realized reward. The selected provider receives a contract equal to the realized reward minus a platform fee, which shapes its incentives for future strategic behavior.

1.1 Contributions

Our main contributions are summarized as follows.

- **Novel incentive design and PALinUCB algorithm.** Motivated by scenarios in which an arm is charged a fixed fee upon selection, we design the selected arm's contract as its realized reward minus the fee. To minimize regret under strategic reporting, we further propose the Per-Arm Linear Upper Confidence Bound (PALinUCB) algorithm, which maintains a separate estimator for each arm using only observations collected from that arm. Consequently, an arm's misreported features affect only its own reward estimate rather than directly altering the estimates of other arms. The separate estimators also help curb substantial strategic deviations in reported features.
- **Theoretical results for regret and strategic behavior.** We establish theoretical guarantees for PALinUCB under our proposed contract design. We first show that the regret of PALinUCB is largely determined by the deviation incentive of the optimal arm, providing a quantitative characterization of how the optimal arm's deviation incentive affects PALinUCB's regret. In particular, when the optimal arm adopts a non-pessimistic reporting strategy, any admissible strategy profile forms a $\mathrm{poly}(\log T)$-approximate Nash equilibrium, where no arm can significantly improve its utility through deviating from its current strategy. Moreover, at Nash equilibrium, the deviation incentive of the optimal arm is zero, and PALinUCB achieves $\mathrm{poly}(\log T)$ regret.

- **Empirical insights on incentive-driven strategic behavior.** Experiments with adaptive strategic arms demonstrate how incentive design shapes strategic behavior of arms and the regret of algorithms. Compared with OptGTM with the contract design $\mathbf{1}(i_t = i)$, PALinUCB with our proposed contract design $r_{t,i} - \eta$ generally achieves lower regret, showing that our contract design helps algorithms shape arms' incentives become favorable to the learner's objective. We further observe the impact of contract parameter η on PALinUCB's regret, which shows the intended role of our contract design in shaping arms' strategic incentives.

1.2 Related Work

Linear Contextual Bandits Stochastic linear bandits assume a shared linear reward model across arms, where the learner can exploit observations from selected arms to infer reward information about other arms. In this setting, Auer (2002) proposed the LinRel and SupLinRel algorithms for regret minimization and established a $\tilde{O}(\sqrt{dT})$ regret upper bound for SupLinRel. Dani et al. (2008) proposed and analyzed the CONFIDENCEBALL1 and CONFIDENCEBALL2 algorithms, whose regret bounds are $\tilde{O}(d\sqrt{T})$ and $\tilde{O}(d^{3/2}\sqrt{T})$ respectively. Dani et al. (2008) also established a generic regret lower bound for all algorithms. Furthermore, in linear contextual bandits, Li et al. (2010) introduced LinUCB for regret minimization, which is an upper confidence bound (UCB)-type algorithm that utilizes the linear similarity among arms. Li et al. (2010) also established a $\widetilde{\mathcal{O}}(\sqrt{KdT})$ regret upper bound for LinUCB. Chu et al. (2011) proposed SupLinUCB, another efficient UCB-type algorithm, and established a near-optimal upper bound $\tilde{O}(\sqrt{dT})$ on its regret. Meanwhile, Abbasi-Yadkori et al. (2011) introduced and analyzed OFUL, which uses self-normalized concentration inequalities to construct ellipsoidal confidence sets. Agrawal & Goyal (2013) designed a Thompson Sampling algorithm for linear contextual bandits and established a $\tilde{O}(d^2\sqrt{T})$ regret upper bound for it. Despite their different algorithm design, these baseline algorithms target in bandit instances with truthful arms. In our model, by contrast, arms are self-interested and may misreport contextual features to maximize their own utility. As a result, the learner's decisions are affected by arms' strategic behavior.

Mechanism Design and Strategic Learning Mechanism design studies how allocation and payment rules can be designed to shape the behavior of self-interested agents. Vickrey (1961) introduced the Vickrey auction, illustrating how incentive-compatible mechanisms can influence agents' strategic decisions. Groves (1973) further developed the Clarke–Groves mechanism, showing how carefully designed payment rules can encourage truthful information revelation. Holmström (1979) studied incentive contracts under moral hazard, while Walton & Carroll (2022) and Burkett & Rosenthal (2024) further investigated robust and data-driven contract design under strategic interactions and information asymmetry. Motivated by these ideas, we design the contract associated with arm selection to shape arms' incentives for strategic reporting and align their behaviors with the learner's regret-minimization objective.

Bandits With Strategic Arms In strategic bandits, Braverman et al. (2019) assumed arms can strategically manipulate the rewards reported to the learner. Feng et al. (2020) subsequently showed that standard stochastic bandit algorithms can retain sublinear regret guarantees when arms have limited manipulation capabilities. Esmaeili et al. (2025) introduced sharply adaptive, monotonic, and fair (SAMF) algorithms that incentivize strategic arms to improve the learner's reward at asymptotic equilibrium. They further established a sublinear regret upper bound when arms deviate from equilibrium. In strategic linear contextual bandits, recent works have introduced mechanism design to handle both the regret minimization and the best arm identification (BAI) problems. On one hand, Kleine Buening et al. (2024) proposed OptGTM, which augmented a variant of the LinUCB algorithm with a Grim Trigger Condition to constraint strategic feature misreporting, and established a sublinear upper bound on its regret under a Nash Equilibrium. However, the regret upper bound of OptGTM is only applied for strategic bandits where arms' reporting features preserve their true mean rewards. Meanwhile, Verma et al. (2025) proposed COBRA, another mechanism-based algorithm designed to enforce truthful behavior among strategic arms. However, the regret upper bound of COBRA assumes that the UCB estimate computed from each arm's reported feature is always no smaller than the arm's true mean reward. On the other hand, Li & Zhong (2026) proposed the MESHA algorithm, which is equipped with a uniform sampling rule and an epoch-wise elimination

rule, to identify the best arm given a limited time budget in strategic linear bandits. Li & Zhong (2026) also established guarantee for MESHA's identification performance under strategic behavior and analyzed failures of optimal-design-based algorithms when being applied to strategic linear bandits . While OptGTM and COBRA established regret guarantees under restrictive conditions and MESHA focuses on BAI, we relax these restrictions, introduce a novel incentive design, and develop PALinUCB for regret minimization under a broader class of strategic behaviors.

2 Problem Formulation

In this paper, we denote the set of positive integers up to K by $[K] = \{1, \dots, K\}$. The inner product of two vectors is denoted by $\langle \cdot, \cdot \rangle$. For any positive definite matrix V, we define the weighted matrix norm $\|x\|_V$ as $\sqrt{x^\top V x}$. The indicator function is denoted by $\mathbf{1}(\cdot)$.

2.1 Linear Contextual Bandits

We begin with a linear contextual bandit of K arms over a horizon of T rounds and a context space $\mathcal{C}$. For each context $c \in \mathcal{C}$ and arm $i \in [K]$, let $x_i^c \in \mathbb{R}^d$ denote the true feature vector with $\|x_i^c\|_2 \leq 1$. The mean reward of arm i under context c is $\mu_i(c) := \langle \theta^*, x_i^c \rangle$, where θ^* is a latent parameter and $\|\theta^*\|_2 \leq 1$. The unit-norm bounds are imposed only for notational convenience and can be replaced by any constant bounds.

At each round $t \in [T]$, under context $c_t \in \mathcal{C}$, the learner observes features $\{x_i(c_t)\}_{i\in[K]}$, selects an arm i_t and receives its corresponding reward

$$r_{t,i_t} = \mu_{i_t}(c_t) + \xi_{t,i_t},$$

where ξ_{t,i_t} is zero-mean and 1-sub-Gaussian conditioned on $\{(x_{i_s}(c_s), r_{s,i_s})\}_{s<t}$. Let $\mu_{t,i} := \mu_i(c_t)$ denote the mean reward of arm i and $i_t^* := \arg\max_{i\in[K]} \mu_{t,i}$ denote the unique optimal arm at round t, respectively. The optimality gap of arm i is $\Delta_{t,i} := \mu_{t,i_t^*} - \mu_{t,i}$.

2.2 Strategic Arms and Nash Equilibrium

We now describe a strategic variant of linear contextual bandits, where the learner cannot observe true features of each arm. In this case, each arm may strategically report their features to maximize their utility. The context sequence c_t may be stochastic or deterministic, but is exogenous: its generation does not depend on the arms' reporting strategies or the learner's past selections.

Strategic Profile. Unlike the standard linear contextual bandit model, each arm i reports a potentially manipulated features $x_{t,i}$ to the learner in strategic linear contextual bandits and the learner selects an arm based on reported features $\{x_{t,i}\}_{i\in[K]}$ at each round t. For all arms, the mean reward of arm i under context c_t remains as $\mu_i(c_t) = \langle \theta^\star, x_i^{c_t} \rangle$, which depends on the arm's true contextual feature rather than its reported feature. We impose the same boundedness condition as in the standard linear contextual bandit model, i.e., $\|x_{t,i}\|_2 \leq 1, \forall t \in [T],\ i \in [K]$.

A strategy of arm i, denoted by σ_i, is a possibly randomized and history-dependent reporting policy. At round t, after observing the public history $h_{t-1} = \{(x_{s,i_s}, r_{s,i_s})\}_{s<t}$ and its true contextual feature $x_i^{c_t}$, arm i selects a reported feature. That is,

$$x_{t,i} \sim \sigma_i(\cdot \mid h_{t-1}, x_i^{c_t}), \qquad x_{t,i} \in \{x : \|x\|_2 \leq 1\}.$$

We write $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_K)$ for a *joint strategy profile*.

Utility Function of Arms When arm i is selected at round t, its contract is defined as $r_{t,i} - \eta$, where $\eta \in \mathbb{R}$ is a constant, which is different from existing strategic linear bandit literature. For example, Kleine Buening et al. (2024); Verma et al. (2025) define 1 as the contract for the selected arm. Later, our study indicates that this contract design is fundamental to align interests between the learner and the arms under information asymmetry. Given the contract $r_{t,i_t} - \eta$, under a joint strategy profile $\boldsymbol{\sigma}$ and an algorithm $\mathcal{A}$, the *expected cumulative utility* of arm i is given by

$$U_i^{\mathcal{A}}(\boldsymbol{\sigma}) := \mathbb{E}_{\boldsymbol{\sigma},\mathcal{A}}\left[\sum_{t=1}^{T} \mathbf{1}\{i_t = i\}(r_{t,i} - \eta)\right].$$

We abbreviate $U_i^{\mathcal{A}}(\boldsymbol{\sigma})$ as $U_i(\boldsymbol{\sigma})$ when there is no ambiguity.

Nash Equilibrium Given an algorithm $\mathcal{A}$, each arm strategically reports its features to maximize its own expected cumulative utility $U_i^{\mathcal{A}}(\boldsymbol{\sigma})$. We define the following *Nash equilibrium* to characterize the equilibrium of the arms' strategical behaviors in this game.

Definition 1 (Nash equilibrium)**.** *A strategy profile $\boldsymbol{\sigma}^* = (\sigma_1^*, \ldots, \sigma_K^*)$ forms a* Nash equilibrium *(NE) if, for any admissible alternative strategy σ_i' of arm $i \in [K]$,*

$$U_i(\sigma_i^*, \boldsymbol{\sigma}_{-i}^*) \geq U_i(\sigma_i', \boldsymbol{\sigma}_{-i}^*),$$

where $\boldsymbol{\sigma}_{-i} = \boldsymbol{\sigma} \setminus \{\sigma_i\}$.

In other words, under a Nash equilibrium, any arm does not have the incentive to deviate from its current strategy. In the following, we define *Approximate Nash equilibrium* to describe approximate equilibrium behaviors of arms.

Definition 2 (ρ-Nash equilibrium)**.** *For $\rho \geq 0$, a strategy profile $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_K)$ is an* ρ-Nash equilibrium *(ρ-NE) if, for any admissible unilateral deviation σ_i' of arm $i \in [K]$,*

$$U_i(\sigma_i', \boldsymbol{\sigma}_{-i}) - U_i(\sigma_i, \boldsymbol{\sigma}_{-i}) \leq \rho.$$

In other words, under a ρ-NE, no arm can gain more than ρ in expected cumulative utility $U_i(\boldsymbol{\sigma})$ through any admissible strategic deviation. In particular, an NE is a 0-NE.

2.3 Strategic Regret

Given any arms' joint strategy profile $\boldsymbol{\sigma}$, the learner applies an algorithm $\mathcal{A}$ to minimize the following *strategic regret*

$$R_T(\boldsymbol{\sigma}, \mathcal{A}) := \mathbb{E}_{\boldsymbol{\sigma}, \mathcal{A}} \left[\sum_{t=1}^{T} \left(\mu_{t, i_t^*} - \mu_{t, i_t} \right) \right].$$

We focus on instances where there exists a uniformly optimal arm over the context space, that is, there exists an arm i^* such that $\mu_{i^*}(c) > \mu_j(c)$ under any context $c \in \mathcal{C}$ for any arm $j \neq i^*$. For brevity, we define $\underline{\Delta}_i = \inf_{c \in \mathcal{C}} (\mu_{i^*}(c) - \mu_i(c))$ and $\bar{\Delta}_i = \sup_{c \in \mathcal{C}} (\mu_{i^*}(c) - \mu_i(c))$. We denote the largest optimality gap by $\Delta_{\max}$, i.e., $\Delta_{\max} = \max_{i \neq i^*} \bar{\Delta}_i$.

Given our contract design $r_{t,i} - \eta$ for arm i, the optimal arm i^* receives expected utility $\mu_{i^*}(c_t) - \eta$ when it is selected under context c_t. For the validity of the game, we focus on the regime where $\mu_{i^*}(c) - \eta$ is positive over the context space $\mathcal{C}$. Specifically, this guarantees that the optimal arm does not receive negative utility so that there exists an arm that has incentive to be selected. Thereby, we impose the following assumption 3.

Assumption 3. *For the optimal arm i^*, there exist constants $0 < g_1 \leq g_2 < \infty$ such that $g_1 \leq \mu_{i^*}(c) - \eta \leq g_2, \forall c \in \mathcal{C}$.*

Remark 4. *Assumption 3 can be satisfied with a positive choice of η when the reward is with a known positive lower bound. Specifically, suppose that $0 < \mu_{\min} \leq \mu_i(c) \leq \mu_{\max}$ for all $i \in [K]$ and $c \in \mathcal{C}$. Then any $\eta \in (0, \mu_{\min})$ makes Assumption 3 hold, with $g_1 = \mu_{\min} - \eta$ and $g_2 = \mu_{\max} - \eta$. For example, the learner may choose $\eta = \mu_{\min}/2$. Hence, the existence of positive contract parameter η does not require knowledge of the latent parameter $\theta^\star$ or the context-dependent mean rewards; it only requires a valid positive lower bound on the reward range.*

3 Learner's Algorithm: PALinUCB

We now present *Per-Arm LinUCB* (PALinUCB), as in Algorithm 1. A key design of PALinUCB is to estimate the mean reward of each arm separately, which localizes the impact of each arm's misreporting behavior. Let $n_t(i) := \sum_{s<t} \mathbf{1}\{i_s = i\}$ denote the number of times arm i has been selected before round t. For each arm i, PALinUCB maintains the regularized design matrix and the corresponding least-squares estimator

$$V_{t,i} := I_d + \sum_{s<t : i_s = i} x_{s,i} x_{s,i}^\top, \quad \widehat{\theta}_{t,i} := V_{t,i}^{-1} \sum_{s<t : i_s = i} x_{s,i} r_{s,i}.$$

After observing reported features $\{x_{t,i}\}_{i=1}^K$ at round t, PALinUCB computes an upper confidence bound on each arm i's mean reward

$$\mathrm{UCB}_{t,i} := \langle \widehat{\theta}_{t,i}, x_{t,i} \rangle + \beta_{t,i} \|x_{t,i}\|_{V_{t,i}^{-1}},$$

where $\beta_{t,i} := \sqrt{d \log\left(1 + \frac{n_t(i)}{d}\right) + 2\log(2T)} + 1$. Lastly, PALinUCB selects arm i_t with the largest $\mathrm{UCB}_{t,i}$, observes its noisy reward r_{t,i_t} and updates the statistics associated with arm i_t.

Algorithm 1 Per-Arm Linear Upper Confidence Bound (PALinUCB)

Require: Horizon T

1: Initialize $V_{1,i} \leftarrow I_d$, $b_{1,i} \leftarrow \mathbf{0}$, $n_1(i) \leftarrow 0$ for all $i \in [K]$

2: **for** $t = 1, \ldots, T$ **do**

3: Observe reported features $\{x_{t,i}\}_{i=1}^K$.

4: Compute the upper confidence bound for each arm $i \in [K]$ by

$$\mathrm{UCB}_{t,i} \leftarrow \langle \widehat{\theta}_{t,i}, x_{t,i} \rangle + \beta_{t,i} \|x_{t,i}\|_{V_{t,i}^{-1}}.$$

5: Select $i_t \in \arg\max_{i \in [K]} \mathrm{UCB}_{t,i}$ and observe r_{t,i_t}. Update statistics:

$$V_{t+1,i_t} \leftarrow V_{t,i_t} + x_{t,i_t} x_{t,i_t}^\top, \quad b_{t+1,i_t} \leftarrow b_{t,i_t} + x_{t,i_t} r_{t,i_t}, \quad n_{t+1}(i_t) \leftarrow n_t(i_t) + 1,$$

$$\widehat{\theta}_{t+1,i_t} \leftarrow V_{t+1,i_t}^{-1} b_{t+1,i_t}, \quad \beta_{t+1,i_t} \leftarrow \sqrt{d \log\left(1 + \frac{n_{t+1}(i_t)}{d}\right) + 2\log(2T)} + 1.$$

6: **end for**

4 Theoretical Results

We first show that PALinUCB's regret is largely determined by the strategy of the optimal arm. For any admissible strategy profile $\boldsymbol{\sigma} = (\sigma_{i^*}, \boldsymbol{\sigma}_{-i^*})$, we describe the deviation incentive of the optimal arm with

$$\mathrm{Expl}_{i^*}(\boldsymbol{\sigma}) := \sup_{\sigma'_{i^*}} \left\{ U_{i^*}(\sigma'_{i^*}, \boldsymbol{\sigma}_{-i^*}) - U_{i^*}(\boldsymbol{\sigma}) \right\},$$

where the supremum is taken over all admissible alternative strategies σ'_{i^*} of arm i^*. $\mathrm{Expl}_{i^*}(\boldsymbol{\sigma})$ measures how much incentive the optimal arm has to deviate from the current strategy to any other alternative admissible strategies. We show that $\mathrm{Expl}_{i^*}(\boldsymbol{\sigma})$ largely determines the regret of PALinUCB. The regret guarantee is formalized in Theorem 5. Note that we also introduce an admissible strategy space that constrains the extent of misreporting, reflecting arms' reputation concerns or platform verification requirements. The proof can be found in Appendix D.

Theorem 5. *The admissible strategy space $\Sigma_i(\varepsilon_i, \theta_i^*)$ is defined as*

$$\Sigma_i(\varepsilon_i, \theta_i^*) := \left\{ \sigma_i = \{x_{t,i}\}_{t=1}^T : \left| \langle \theta_i^*, x_{t,i} \rangle - \mu_i(c_t) \right| \le \varepsilon_i, \quad \forall t \in [T] \right\}$$

with $0 \le \varepsilon_i < \underline{\Delta}_i$ and $\theta_i^* \in \{\theta_i^* \in \mathbb{R}^d : \|\theta_i^*\|_2 \le 1\}$ for all $i \ne i^*$. Then under Assumption 3, given any joint strategy profile $\boldsymbol{\sigma} \in \prod_{i \in [K]} \Sigma_i(\varepsilon_i, \theta_i^*)$, Algorithm 1 satisfies

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[T - n_{T+1}(i^*)\right] \in \widetilde{O}\left(\frac{g_2}{g_1} d^2 \sum_{i \ne i^*} \frac{1}{\kappa_i^3 \underline{\Delta}_i^2} + \frac{\mathrm{Expl}_{i^*}(\boldsymbol{\sigma})}{g_1} \right),$$

and consequently,

$$R_T(\boldsymbol{\sigma}, \mathrm{PALinUCB}) \in \widetilde{O}\left(\Delta_{\max} \frac{g_2}{g_1} d^2 \sum_{i \ne i^*} \frac{1}{\kappa_i^3 \underline{\Delta}_i^2} + \frac{\Delta_{\max}}{g_1} \mathrm{Expl}_{i^*}(\boldsymbol{\sigma}) \right),$$

where $\kappa_i = 1 - \varepsilon_i / \underline{\Delta}_i$, $\quad \Delta_{\max} = \max_{i \ne i^*} \overline{\Delta}_i$.

Remark 6. *Note that the parameters θ_i^* and ε_i are initially determined for each arm i. Consider a product provider on a recommendation platform that may strategically report its product features in order to increase its chance of being recommended. In this case, $x_i(c_t)$ represents the true feature vector of product i under context c_t, and $x_{t,i}$ represents the feature vector reported by provider i at round t. The strategic deviation of arm i at round t is defined as $\varepsilon_{t,i} := \langle \theta_i^\star, x_{t,i} \rangle - \langle \theta^\star, x_i(c_t) \rangle$. Here, $\langle \theta^\star, x_i(c_t) \rangle$ represents the true expected reward generated when product i is selected, where $\theta^\star$ captures the true market response to product features. We assume that each provider knows the true market state $\theta^\star$. However, providers may adopt heterogeneous advertising strategies, which induce different market beliefs θ_i^* about how product features translate into rewards. Given this induced belief θ_i^*, provider i may strategically report its feature as $x_{t,i}$ to induce the perceived reward $\langle \theta_i^\star, x_{t,i} \rangle$, thereby increasing its likelihood of being recommended. Hence, $\varepsilon_{t,i}$ captures the gap between the reward perceived by the market under provider i's advertising strategy and strategically reported features, and the product's true expected reward under the actual market state. Then ε_i upper bounds the gap, which describes the strength of the strategic deviation and is determined by each provider's ethical considerations or reputation concerns.*

Extending the observation of Theorem 5, we further show that the optimal arm i^*'s strategy remains the key factor determining PALinUCB's regret under more general settings. Details are provided in Appendix H. Next, we show that an approximate Nash Equilibrium can be formed when the optimal arm i^* adopts some certain strategies.

Theorem 7. *The non-pessimistic strategy class Σ_i^{np} of arm i is defined as*

$$\Sigma_i^{\mathrm{np}} := \left\{ \sigma_i \in \Sigma_i(\varepsilon_i) : \mathbb{P}_{\sigma_i, \boldsymbol{\sigma}_{-i}} \left(m_{t,i} \geq 0,\ \forall t \in [T] \right) = 1,\ \forall \boldsymbol{\sigma}_{-i} \in \prod_{j \neq i} \Sigma_j(\varepsilon_j, \theta_j^*) \right\}$$

with $m_{t,i} = \varepsilon_{t,i} - x_{t,i}^\top V_{t,i}^{-1} \sum_{s<t} \varepsilon_{s,i} x_{s,i} \cdot \mathbf{1}(i_s = i)$, $0 \leq \varepsilon_i < \underline{\Delta}_i$, and $\theta_i^* \in \{\theta_i^* \in \mathbb{R}^d :\ \|\theta_i^*\|_2 \leq 1\}$ for all $i \neq i^*$. Then under Assumption 3, given Algorithm 1, any admissible joint strategy profile $\boldsymbol{\sigma}$ satisfying $\sigma_{i^*} \in \Sigma_{i^*}^{\mathrm{np}}$ forms a ρ_T-Nash equilibrium, where

$$\rho_T \in \widetilde{O}\left(d^2 \max\left\{ g_2 \sum_{j \neq i^*} \frac{1}{\kappa_j^3 \underline{\Delta}_j^2},\ \max_{i \neq i^*} \frac{(g_2 - \underline{\Delta}_i)_+ + (\overline{\Delta}_i - g_1)_+}{\kappa_i^3 \underline{\Delta}_i^2} \right\} \right),$$

$(x)_+ := \max\{x, 0\}$, and $\kappa_i = 1 - \varepsilon_i / \underline{\Delta}_i$.

To demonstrate that the condition in Theorem 7 is non-vacuous, we verify that $\Sigma_{i^*}^{\mathrm{np}}$ is non-empty. Let $\sigma_i^{\mathrm{tr}} = \{x_i(c_t)\}_{t \in [T]}$ denote the truthful reporting strategy of arm i. Since truthful reporting implies $m_{t,i^*} = 0$ for all $t \in [T]$, we have $\sigma_{i^*}^{\mathrm{tr}} \in \Sigma_{i^*}^{\mathrm{np}}$. Therefore, Theorem 7 implies that any joint strategy profile $\boldsymbol{\sigma}$ in which the optimal arm adopts $\sigma_{i^*}^{\mathrm{tr}}$ forms a ρ-Nash equilibrium. Furthermore, for any joint strategy profile $\boldsymbol{\sigma}$ that can form a ρ-Nash equilibrium, the definition of ρ-Nash equilibrium implies that $\mathrm{Expl}_{i^*}(\boldsymbol{\sigma}) \leq \rho$. Therefore, Theorem 5 indicates that

$$R_T(\boldsymbol{\sigma}, \mathrm{PALinUCB}) \in \widetilde{O}\left(\Delta_{\max} \frac{g_2}{g_1} d^2 \sum_{i \neq i^*} \frac{1}{\kappa_i^3 \underline{\Delta}_i^2} + \frac{\Delta_{\max}}{g_1} \rho \right).$$

Particularly, under Nash equilibrium, we have $\mathrm{Expl}_{i^*}(\boldsymbol{\sigma}) = 0$ and therefore the regret of PALinUCB is upper bounded by $\mathrm{poly}(\log T)$.

Remark 8. *Existing analytical results in strategic bandits are typically conditioned on the existence of Nash equilibria. For example, the regret guarantees of OptGTM (Kleine Buening et al., 2024) and COBRA (Verma et al., 2025) are established under NE, while the existence of such equilibria is generally not characterized. In contrast, we can show that given PALinUCB, there exists a joint strategy profile $\boldsymbol{\sigma}$ that can form a NE in some instances. Details are provided in Appendix D.*

In summary, Theorem 5 shows that PALinUCB's regret is determined by the deviation incentive of the optimal arm, providing a quantitative characterization of how incentive misalignment affects the learner's regret. Besides, Theorem 7 further shows that any admissible strategy profile with a non-pessimistic optimal arm forms a ρ_T-Nash equilibrium. Since truthful reporting is a non-pessimistic strategy, such equilibria exist for a broad class of strategy profiles. Moreover, at any Nash equilibrium, the deviation incentive of the optimal arm is zero, and Theorem 5 implies a $\mathrm{poly}(\log T)$ regret guarantee.

5 NUMERICAL EXPERIMENTS

In this section, we examine how incentive design shapes arms' strategic behavior and affects the learner's regret. We evaluate PALinUCB in a synthetic strategic linear contextual bandit environment with the following parameters: arms $K = 7$, feature dimension $d = 2$, latent parameter $\theta^\star = (1, 0)^\top$, and reward noise $\xi_{t,i} \sim \mathcal{N}(0, 0.1^2)$. For each arm i and each context c, we set the true feature of arm i under context c with $x_i^c = y_c \cdot z_i$, where $y_c \sim \mathrm{Unif}[0.5, 1]$ and z_i is a fixed parameter of arm i. We further set $\inf_{c\in\mathcal{C}} \mu_{i^*}(c) = 1$ and $\sup_{c\in\mathcal{C}} \mu_{i^*}(c) = 2$. In Figure 2, we compare PALinUCB's regret under different values of η with that of OptGTM by varying the horizon T. We further vary η to show that how the contract parameter shapes arms' strategic incentives and impacts regret in Figures 3 and 4. Although our theoretical analysis relies on Assumption 3, we evaluate PALinUCB with different values of the contract parameter η that either satisfy or violate this assumption. The detailed implementation of strategic behavior is deferred to Appendix F.

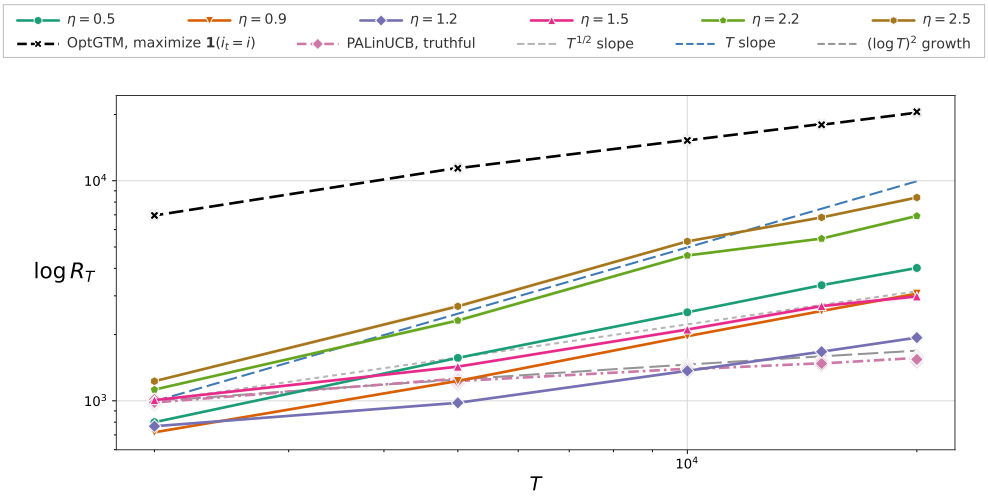
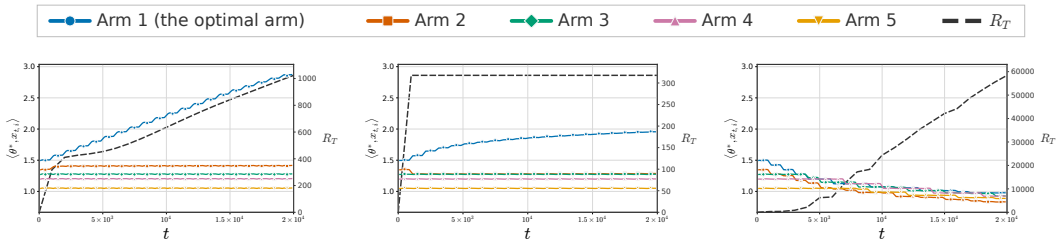


Figure 2: Cumulative regret R_T of PALinUCB (with $r_{t,i} - \eta$ as the contract design) and that of OptGTM (with $\mathbf{1}(i_t = i)$ as the contract design) across varying horizons T on log scales.

Contract design helps minimize learner's regret. Figure 2 illustrates the impact of the contract parameter η and time horizon T on the cumulative regret $R(T)$. We evaluate PALinUCB when arms maximize their utility under various values of η, and compare its performance against OptGTM, where arms seek to maximize their total pull counts. First, PALinUCB consistently achieves lower regret than OptGTM across all evaluated values of η. This highlights the effectiveness of shaping strategic incentives through our proposed contract utility $r_{t,i} - \eta$, which drives a substantial performance advantage over OptGTM's baseline contract utility of $\mathbf{1}\{i_t = i\}$. Note that PALinUCB's regret is always smaller than that of OptGTM regardless of choice of contract parameter η, which indicates the robustness of PALinUCB against the choice of η.

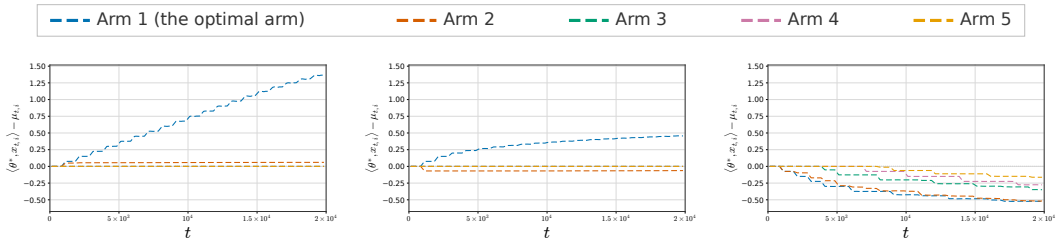
Second, Figure 2 demonstrates that the impact of η persists as T grows. Comparing the regret trajectories against reference curves ($T^{1/2}$, T, and $(\log T)^2$), we observe that PALinUCB achieves sublinear regret when $\eta \leq 2$, but incurs linear regret when $\eta > 2$. This divergence underscores the necessity of Assumption 3 for guaranteeing sublinear regret. Furthermore, the linear regret for $\eta > 2$ aligns with our theoretical analysis in Appendix E, which proves that PALinUCB suffers linear regret at Nash equilibrium whenever the optimal arm yields negative utility upon selection.

Contract design shapes strategic reporting behavior. In Figure 3, we plot $\langle \theta^*, x_{t,i} \rangle$, the reward induced by the reported feature of arm i, together with the regret of PALinUCB when t grows. In Figure 4, we plot $\langle \theta^*, x_{t,i} \rangle - \mu_{t,i}$, which is the gap between the reward induced by the reported feature and the true mean reward for arm i, reflecting the direction and magnitude of each arm's strategic deviation when t grows. We set the time horizon $T = 2 \times 10^4$ and vary the contract parameter η in both figures. Without loss of generality, we set arm 1 as the optimal arm. For clarity, Arms 6 and 7 are omitted since their reported values remain nearly unchanged across different values of η.



(a) (b) (c)

Figure 3: Reported reward $\langle \theta^*, x_{t,i} \rangle - \mu_{t,i}$ (solid curves corresponding to the left y-axis) and regret $R(t)$ (dashed curves corresponding to the right y-axis) with different η: **(a)** $\eta = 0.8$. **(b)** $\eta = 1.4$. **(c)** $\eta = 2.2$.



(a) (b) (c)

Figure 4: Reward deviation $\langle \theta^*, x_{t,i} \rangle - \mu_{t,i}$ with different η: **(a)** $\eta = 0.8$. **(b)** $\eta = 1.4$. **(c)** $\eta = 2.2$.

Figures 4a and 4b indicate that, for $\eta \leq 2$, the strategic deviations of arm 1 are positive throughout the horizon, indicating that the optimal arm tends to over-report its reward in order to be selected by the learner. Correspondingly, Figures 3a and 3b show that the reward induced by the optimal arm's reported feature remains higher than those of the other arms, making the optimal arm more likely to be selected and resulting in lower regret over the horizon T. In contrast, Figure 4c shows that, for $\eta > 2$, the strategic deviations of all arms become negative, indicating that arms tend to under-report their reward and avoid selection. Correspondingly, as shown in Figure 3c, the reward induced by the optimal arm's reported feature is then no longer the largest, making it less likely to be selected and leading to larger regret over the horizon T. This divergence again underscores the necessity of Assumption 3 for guaranteeing sublinear regret.

Moreover, Figures 3a, 3b, 4a and 3b altogether indicate that compared with $\eta = 0.8$, $\eta = 1.4$ yields lower regret of PALinUCB because it furhter weakens the incentives of suboptimal arms to get selected while preserving sufficient incentive for the optimal arm to be selected. Altogether, We observe a non-monotone dependence of PALinUCB's regret on η, which corroborates with our contract design. In particular, Figure 4 suggests that an arm tends to over-report its reward and gets selected when its instantaneous reward exceeds the contract parameter η, whereas it tends to under-report its reward and avoid selection when its reward falls below η. This illustrates the intended role of the contract $r_{t,i} - \eta$ in shaping arms' strategic incentives.

6 Conclusion

In this work, we introduce contract design to strategic linear contextual bandits and propose PALinUCB, which maintains separate estimators for each arm with the contract design $r_{t,i} - \eta$ to shape arms' strategic incentives. Our theoretical results show that, PALinUCB's regret is largely determined by the strategic behavior of the optimal arm and achieves $\mathrm{poly}(\log T)$ under Nash equilibrium. We further show that approximate Nash equilibrium can be formed when the optimal arm adopts a non-pessimistic reporting strategy. Our experiments demonstrate that the contract we design shapes arms' strategic incentives and there is a non-monotone dependence of PALinUCB's regret on η. These results motivate future research on applying contract design to broader strategic bandit settings, including more general contract structures and richer strategic interactions.

ETHICS STATEMENT

This work uses synthetic experiments, involving no human subjects or personal data. Our study concerns utility design for shaping strategic behavior in online decision-making systems, which may affect the incentives and welfare of participating providers. Practical applications should take into account their potential impact on participating providers.

AI USE STATEMENT

In this work, we used ChatGPT 5.6 to assist with language editing, improving clarity, and organizing parts of the manuscript. We did not use generative AI tools to generate theoretical results, proofs, experimental data, or empirical conclusions. All AI-assisted content was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

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.

Mark Braverman, Jieming Mao, Jon Schneider, and S Matthew Weinberg. Multi-armed bandit problems with strategic arms. In *Conference on Learning Theory*, pp. 383–416. PMLR, 2019.

Justin Burkett and Maxwell Rosenthal. Data-driven contract design. *Journal of Economic Theory*, 221:105900, 2024.

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.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, pp. 355–366, 2008.

Seyed A Esmaeili, Suho Shin, and Aleksandrs Slivkins. Robust performance incentivizing algorithms for multi-armed bandits with strategic agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13814–13822, 2025.

Zhe Feng, David Parkes, and Haifeng Xu. The intrinsic robustness of stochastic bandits to strategic manipulation. In *International Conference on Machine Learning*, pp. 3092–3101. PMLR, 2020.

Theodore Groves. Incentives in teams. *Econometrica: Journal of the Econometric Society*, pp. 617–631, 1973.

Bengt Holmström. Moral hazard and observability. *The Bell journal of economics*, pp. 74–91, 1979.

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

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

Xin Li and Zixin Zhong. MESHA: Mechanism-enforced sequential halving for strategic linear bandits. *arXiv preprint arXiv:2607.14706*, 2026.

Arun Verma, Indrajit Saha, Makoto Yokoo, and Bryan Kian Hsiang Low. COBRA: Contextual bandit algorithm for ensuring truthful strategic agents. *arXiv preprint arXiv:2505.23720*, 2025.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

Daniel Walton and Gabriel Carroll. A general framework for robust contracting models. *Econometrica*, 90(5):2129–2159, 2022.

APPENDIX OVERVIEW

The appendix is organized as follows. Appendix A collects the notation and technical lemmas used throughout the proofs. Appendix B establishes the upper bound on the pull counts of suboptimal arms when PALinUCB is applied. Appendices C and D present the proofs of Theorems 7 and 5, respectively, together with instances where PALinUCB with the proposed contract design reaches a Nash equilibrium. Appendix E presents instances where the Nash equilibrium with linear regret exist when Assumption 3 is violated. Appendix F provides additional experimental details, including the experimental setup, the simulation of arms' strategic behaviors, and additional experimental results. Finally, Appendix G compares our theoretical guarantees with those of existing algorithms for strategic contextual bandits, and Appendix H derives a more general regret upper bound for PALinUCB without Assumption 3.

A NOTATIONS AND TECHNICAL LEMMAS

For all arm $i \in [K]$ and $t \in [T]$, we set

$$S_{t,i} = \sum_{s<t, i_s=i} \varepsilon_{s,i} x_{s,i}, \quad P_{t,i} = S_{t,i}^\top V_{t,i}^{-1} S_{t,i},$$

$$q_{t,i} = x_{t,i}^\top V_{t,i}^{-1} S_{t,i}, \quad m_{t,i} = \varepsilon_{t,i} - q_{t,i}, \quad a_{t,i}^2 = x_{t,i}^\top V_{t,i}^{-1} x_{t,i}.$$

Throughout the proofs, we assume without loss of generality that arm 1 is the optimal arm, i.e., $i^* = 1$. Here are some useful facts applied in our proofs.

Lemma 9. *Fix any arm $i \in [K]$. Let*

$$V_{1,i} = I_d, \qquad V_{t+1,i} = V_{t,i} + \mathbf{1}\{i_t = i\} x_{t,i} x_{t,i}^\top,$$

where $\|x_{t,i}\|_2 \le 1$ for all $t \in [T]$. Then

$$\sum_{\substack{t=1 \\ i_t=i}}^{T} \min\{1, a_{t,i}^2\} \le 2d \log\left(1 + \frac{n_{T+1}(i)}{d}\right) \le 2d \log\left(1 + \frac{T}{d}\right).$$

Proof. When $i_t = i$, we have

$$V_{t+1,i} = V_{t,i} + x_{t,i} x_{t,i}^\top.$$

Therefore,

$$\begin{aligned} \det(V_{t+1,i}) &= \det\left(V_{t,i} + x_{t,i} x_{t,i}^\top\right) \\ &= \det(V_{t,i}) \det\left(I + V_{t,i}^{-1} x_{t,i} x_{t,i}^\top\right) \\ &= \det(V_{t,i}) \left(1 + x_{t,i}^\top V_{t,i}^{-1} x_{t,i}\right) \\ &= \det(V_{t,i})(1 + a_{t,i}^2). \end{aligned}$$

When $i_t \neq i$, we have $V_{t+1,i} = V_{t,i}$. Therefore,

$$\log \frac{\det(V_{T+1,i})}{\det(V_{1,i})} = \sum_{\substack{t=1 \\ i_t=i}}^{T} \log(1 + a_{t,i}^2).$$

For every $z \ge 0$,

$$\min\{1, z\} \le 2 \log(1 + z),$$

and hence

$$\sum_{\substack{t=1 \\ i_t=i}}^{T} \min\{1, a_{t,i}^2\} \le 2 \log \frac{\det(V_{T+1,i})}{\det(V_{1,i})}.$$

Since

$$V_{T+1,i} = I_d + \sum_{\substack{t=1 \\ i_t=i}}^{T} x_{t,i} x_{t,i}^\top,$$

and $\|x_{t,i}\|_2 \le 1$,

$$\mathrm{tr}(V_{T+1,i}) = d + \sum_{\substack{t=1 \\ i_t=i}}^{T} \|x_{t,i}\|_2^2 \le d + n_{T+1}(i).$$

Applying the arithmetic–geometric mean inequality to the eigenvalues of $V_{T+1,i}$ yields

$$\det(V_{T+1,i}) \le \left(\frac{\mathrm{tr}(V_{T+1,i})}{d}\right)^d \le \left(1 + \frac{n_{T+1}(i)}{d}\right)^d.$$

Since $V_{1,i} = I_d$ and therefore $\det(V_{1,i}) = 1$, we obtain

$$2\log\frac{\det(V_{T+1,i})}{\det(V_{1,i})} \le 2d\log\left(1 + \frac{n_{T+1}(i)}{d}\right).$$

Finally, $n_{T+1}(i) \le T$ gives the last inequality. □

Lemma 10. *For any arm $i \in [K]$ and round $t \in [T]$, if arm i is pulled at round t, i.e., $i_t = i$, then*

$$P_{t+1,i} - P_{t,i} = \varepsilon_{t,i}^2 - \frac{m_{t,i}^2}{1 + a_{t,i}^2}.$$

If $i_t \neq i$, then

$$P_{t+1,i} = P_{t,i}.$$

Proof. Suppose that $i_t = i$. By definition,

$$V_{t+1,i} = V_{t,i} + x_{t,i}x_{t,i}^\top, \quad S_{t+1,i} = S_{t,i} + \varepsilon_{t,i}x_{t,i}.$$

By the Sherman–Morrison formula,

$$V_{t+1,i}^{-1} = V_{t,i}^{-1} - \frac{V_{t,i}^{-1}x_{t,i}x_{t,i}^\top V_{t,i}^{-1}}{1 + x_{t,i}^\top V_{t,i}^{-1}x_{t,i}}.$$

Recalling that

$$a_{t,i}^2 := x_{t,i}^\top V_{t,i}^{-1}x_{t,i}, \qquad q_{t,i} := x_{t,i}^\top V_{t,i}^{-1}S_{t,i},$$

we obtain

$$\begin{aligned} P_{t+1,i} &= (S_{t,i} + \varepsilon_{t,i}x_{t,i})^\top V_{t+1,i}^{-1}(S_{t,i} + \varepsilon_{t,i}x_{t,i}) \\ &= P_{t,i} + 2\varepsilon_{t,i}q_{t,i} + \varepsilon_{t,i}^2 a_{t,i}^2 - \frac{\left(q_{t,i} + \varepsilon_{t,i}a_{t,i}^2\right)^2}{1 + a_{t,i}^2}. \end{aligned}$$

Therefore,

$$\begin{aligned} P_{t+1,i} - P_{t,i} &= \frac{2\varepsilon_{t,i}q_{t,i} + \varepsilon_{t,i}^2 a_{t,i}^2 + \varepsilon_{t,i}^2 a_{t,i}^2 - q_{t,i}^2 - \varepsilon_{t,i}^2 a_{t,i}^4}{1 + a_{t,i}^2} \\ &= \varepsilon_{t,i}^2 - \frac{\varepsilon_{t,i}^2 - 2\varepsilon_{t,i}q_{t,i} + q_{t,i}^2}{1 + a_{t,i}^2}. \end{aligned}$$

Since

$$m_{t,i} = \varepsilon_{t,i} - q_{t,i},$$

we have

$$P_{t+1,i} - P_{t,i} = \varepsilon_{t,i}^2 - \frac{m_{t,i}^2}{1 + a_{t,i}^2}.$$

If $i_t \neq i$, then neither $V_{t,i}$ nor $S_{t,i}$ is updated, and hence

$$V_{t+1,i} = V_{t,i}, \qquad S_{t+1,i} = S_{t,i}.$$

Therefore,

$$P_{t+1,i} = P_{t,i}.$$

□

Corollary 11. *For any arm $i \in [K]$ and any $t \in [T+1]$,*

$$\sum_{\substack{s<t \\ i_s=i}} \frac{m_{s,i}^2}{1+a_{s,i}^2} \le \sum_{\substack{s<t \\ i_s=i}} \varepsilon_{s,i}^2 \le \varepsilon_i^2 n_{T+1}(i).$$

Proof. By Lemma 10, for every $s < t$,

$$P_{s+1,i} - P_{s,i} = \mathbf{1}\{i_s = i\} \left(\varepsilon_{s,i}^2 - \frac{m_{s,i}^2}{1+a_{s,i}^2} \right).$$

Summing over $s = 1, \dots, t-1$ gives

$$P_{t,i} - P_{1,i} = \sum_{\substack{s<t \\ i_s=i}} \left(\varepsilon_{s,i}^2 - \frac{m_{s,i}^2}{1+a_{s,i}^2} \right).$$

Since $S_{1,i} = 0$, we have $P_{1,i} = 0$. Moreover,

$$P_{t,i} = S_{t,i}^\top V_{t,i}^{-1} S_{t,i} \ge 0.$$

Therefore,

$$\sum_{\substack{s<t \\ i_s=i}} \frac{m_{s,i}^2}{1+a_{s,i}^2} \le \sum_{\substack{s<t \\ i_s=i}} \varepsilon_{s,i}^2.$$

Finally, since $|\varepsilon_{s,i}| \le \varepsilon_i$ for every s,

$$\sum_{\substack{s<t \\ i_s=i}} \varepsilon_{s,i}^2 \le \varepsilon_i^2 \sum_{s<t} \mathbf{1}\{i_s = i\} = \varepsilon_i^2 n_t(i).$$

This completes the proof. □

B Strategy-Robust Pull-Count Bound

We first establish a pull-count bound for suboptimal arms, which serves as the main technical ingredient in the proofs of Theorems 5 and 7 .

Lemma 12. *Fix any arm $i \in [K]$. Let*

$$b_{t,i}(\delta) := \sqrt{d \log\left(1 + \frac{n_t(i)}{d}\right) + 2\log\frac{1}{\delta}} + 1.$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $t \in [T]$,

$$\left\| \widehat{\theta}_{t,i} - \theta_i^\star + V_{t,i}^{-1} S_{t,i} \right\|_{V_{t,i}} \le b_{t,i}(\delta).$$

Hence, for every $x \in \mathbb{R}^d$,

$$\left| \left\langle \widehat{\theta}_{t,i} - \theta_i^\star, x \right\rangle + x^\top V_{t,i}^{-1} S_{t,i} \right| \le b_{t,i}(\delta) \|x\|_{V_{t,i}^{-1}}$$

holds with probability at least $1-\delta$.

Proof. Recall that the reward model is

$$r_{s,i} = \langle \theta_i^\star, x_{s,i} \rangle - \varepsilon_{s,i} + \xi_{s,i}.$$

Hence,

$$\begin{aligned} \widehat{\theta}_{t,i} &= V_{t,i}^{-1} \sum_{s<t:i_s=i} x_{s,i} r_{s,i} \\ &= V_{t,i}^{-1} \left[\sum_{s<t:i_s=i} x_{s,i} x_{s,i}^\top \theta_i^\star - S_{t,i} + \sum_{s<t:i_s=i} x_{s,i} \xi_{s,i} \right]. \end{aligned}$$

Since

$$\sum_{s<t:i_s=i} x_{s,i}x_{s,i}^\top = V_{t,i} - I_d,$$

we have that

$$\widehat{\theta}_{t,i} - \theta_i^\star = -V_{t,i}^{-1}S_{t,i} + V_{t,i}^{-1}\left(\sum_{s<t:i_s=i} x_{s,i}\xi_{s,i} - \theta_i^\star\right).$$

Therefore,

$$\widehat{\theta}_{t,i} - \theta_i^\star + V_{t,i}^{-1}S_{t,i} = V_{t,i}^{-1}\left(\sum_{s<t:i_s=i} x_{s,i}\xi_{s,i} - \theta_i^\star\right).$$

Taking the $V_{t,i}$-norm and applying the triangle inequality in both sides gives

$$\begin{aligned}&\left\|\widehat{\theta}_{t,i} - \theta_i^\star + V_{t,i}^{-1}S_{t,i}\right\|_{V_{t,i}} \\ &\le \left\|\sum_{s<t:i_s=i} x_{s,i}\xi_{s,i}\right\|_{V_{t,i}^{-1}} + \|\theta_i^\star\|_{V_{t,i}^{-1}}.\end{aligned}$$

Since $V_{t,i} \succeq I_d$,

$$\|\theta_i^\star\|_{V_{t,i}^{-1}} \le \|\theta_i^\star\|_2 \le 1.$$

Since the reported features and the selection decision at round t are determined before observing the reward noise, the sequence of selected reported features is predictable with respect to the pre-reward filtration. Hence the self-normalized martingale inequality implies that, with probability at least $1-\delta$, simultaneously for all $t \in [T]$,

$$\left\|\sum_{s<t:i_s=i} x_{s,i}\xi_{s,i}\right\|_{V_{t,i}^{-1}}^2 \le \log\det(V_{t,i}) + 2\log\frac{1}{\delta}.$$

Moreover,

$$\mathrm{tr}(V_{t,i}) = d + \sum_{s<t:i_s=i} \|x_{s,i}\|_2^2 \le d + n_t(i).$$

By the arithmetic–geometric mean inequality applied to the eigenvalues of $V_{t,i}$,

$$\det(V_{t,i}) \le \left(1 + \frac{n_t(i)}{d}\right)^d.$$

Thus,

$$\left\|\sum_{s<t:i_s=i} x_{s,i}\xi_{s,i}\right\|_{V_{t,i}^{-1}} \le \sqrt{d\log\left(1 + \frac{n_t(i)}{d}\right) + 2\log\frac{1}{\delta}}.$$

It follows that

$$\left\|\widehat{\theta}_{t,i} - \theta_i^\star + V_{t,i}^{-1}S_{t,i}\right\|_{V_{t,i}} \le b_{t,i}(\delta).$$

For every $x \in \mathbb{R}^d$, the generalized Cauchy–Schwarz inequality yields

$$\left|\left\langle \widehat{\theta}_{t,i} - \theta_i^\star + V_{t,i}^{-1}S_{t,i}, x\right\rangle\right| \le b_{t,i}(\delta)\|x\|_{V_{t,i}^{-1}},$$

which proves

$$\left|\left\langle \widehat{\theta}_{t,i} - \theta_i^\star, x\right\rangle + x^\top V_{t,i}^{-1}S_{t,i}\right| \le b_{t,i}(\delta)\|x\|_{V_{t,i}^{-1}}.$$

□

We now derive Lemma 13 by Lemma 12.

Lemma 13. *Suppose $0 \le \varepsilon_i < \underline{\Delta}_i$ holds for all arm $i \neq 1$. For any non-pessimistic strategy $\sigma_1^{\mathrm{np}} \in \Sigma_1^{\mathrm{np}}$ of the optimal arm and any admissible joint strategy profile $\boldsymbol{\sigma}_{-1}$ of the remaining arms,*

$$\mathbb{E}_{\sigma_1^{\mathrm{np}},\boldsymbol{\sigma}_{-1}}\left[n_{T+1}(i)\right] \le L_i(T) + 1, \qquad \forall i \neq 1,$$

where

$$L_i(T) := \frac{2d(4-\kappa_i)^2 \log\left(1+\frac{T}{d}\right)}{\kappa_i(20-7\kappa_i-4\kappa_i^2)} \left[\frac{64\overline{\beta}_T^2}{\kappa_i^2 \underline{\Delta}_i^2} + \frac{4}{\kappa_i}\right],$$

with

$$\overline{\beta}_T := \sqrt{d\log\left(1+\frac{T}{d}\right) + 2\log(2T)} + 1.$$

In particular,

$$\mathbb{E}_{\sigma_1^{\mathrm{np}},\boldsymbol{\sigma}_{-1}}\left[n_{T+1}(i)\right] \in \widetilde{O}\left(\frac{d^2}{\kappa_i^3 \underline{\Delta}_i^2}\right), \qquad \forall i \neq 1.$$

Moreover,

$$\mathbb{E}_{\sigma_1^{\mathrm{np}},\boldsymbol{\sigma}_{-1}}\left[T - n_{T+1}(1)\right] \le \sum_{i=2}^{K}\big(L_i(T)+1\big),$$

and consequently

$$\mathbb{E}_{\sigma_1^{\mathrm{np}},\boldsymbol{\sigma}_{-1}}\left[T - n_{T+1}(1)\right] \in \widetilde{O}\left(d^2 \sum_{i=2}^{K} \frac{1}{\kappa_i^3 \underline{\Delta}_i^2}\right).$$

Proof. We first establish an upper bound on $n_{T+1}(i)$ conditioned on an event $\mathcal{E}_i$, which is defined as

$$\mathcal{E}_i := \bigcap_{j\in\{1,i\}} \bigcap_{t=1}^{T} \left\{ \left\| \widehat{\theta}_{t,j} - \theta_j^\star + V_{t,j}^{-1} S_{t,j} \right\|_{V_{t,j}} \le \beta_{t,j} \right\}.$$

By Lemma 12 with confidence level $\delta = 1/(2T)$ for each $j \in \{1, i\}$,

$$\mathbb{P}(\mathcal{E}_i^c) \le \frac{1}{T}.$$

Conditioned on $\mathcal{E}_i$, Lemma 12 implies

$$\left| \left\langle \widehat{\theta}_{t,i} - \theta_i^\star, x_{t,i} \right\rangle + q_{t,i} \right| \le \beta_{t,i} a_{t,i}.$$

Hence,

$$\left\langle \widehat{\theta}_{t,i} - \theta_i^\star, x_{t,i} \right\rangle \le -q_{t,i} + \beta_{t,i} a_{t,i}.$$

Since

$$\langle \theta_i^\star, x_{t,i} \rangle = \mu_i(c_t) + \varepsilon_{t,i},$$

we obtain

$$\begin{aligned} \langle \widehat{\theta}_{t,i}, x_{t,i} \rangle &\le \mu_i(c_t) + \varepsilon_{t,i} - q_{t,i} + \beta_{t,i} a_{t,i} \\ &= \mu_i(c_t) + m_{t,i} + \beta_{t,i} a_{t,i}. \end{aligned} \tag{1}$$

Therefore,

$$\mathrm{UCB}_{t,i} \le \mu_i(c_t) + m_{t,i} + 2\beta_{t,i} a_{t,i}. \tag{2}$$

On the other hand, arm 1 follows a non-pessimistic strategy $\sigma_1^{\mathrm{np}} \in \Sigma_1^{\mathrm{np}}$, hence

$$m_{t,1} \ge 0.$$

Again by Lemma 12,

$$\left\langle \widehat{\theta}_{t,1} - \theta_1^\star, x_{t,1} \right\rangle + q_{t,1} \ge -\beta_{t,1} a_{t,1},$$

and therefore

$$\begin{aligned}\mathrm{UCB}_{t,1} &= \langle \widehat{\theta}_{t,1}, x_{t,1} \rangle + \beta_{t,1} a_{t,1} \\ &\geq \langle \theta_1^\star, x_{t,1} \rangle - q_{t,1} \qquad (3) \\ &= \mu_1(c_t) + \varepsilon_{t,1} - q_{t,1} \qquad (4) \\ &= \mu_1(c_t) + m_{t,1} \qquad (5)\end{aligned}$$

Suppose arm i is selected at round t. By the selection rule of PALinUCB,

$$\mathrm{UCB}_{t,i} \geq \mathrm{UCB}_{t,1}\,.$$

Combining equation 2 and equation 5 gives

$$\mu_1(c_t) - \mu_i(c_t) \leq m_{t,i} + 2\beta_{t,i} a_{t,i}.$$

Since

$$\underline{\Delta}_i \leq \mu_1(c_t) - \mu_i(c_t),$$

we conclude that whenever $i_t = i$,

$$\underline{\Delta}_i \leq m_{t,i} - m_{t,1} + 2\beta_{t,i} a_{t,i} \leq m_{t,i} + 2\beta_{t,i} a_{t,i}. \qquad (6)$$

Since $V_{t,i} \succeq I_d$ and $\|x_{t,i}\|_2 \leq 1$,

$$a_{t,i}^2 = x_{t,i}^\top V_{t,i}^{-1} x_{t,i} \leq \|x_{t,i}\|_2^2 \leq 1.$$

Therefore, Lemma 9 gives

$$\sum_{t \leq T : i_t = i} a_{t,i}^2 \leq 2d \log\left(1 + \frac{n_{T+1}(i)}{d}\right) \leq 2d \log\left(1 + \frac{T}{d}\right). \qquad (7)$$

For convenience, let

$$\Phi_T := 2d \log\left(1 + \frac{T}{d}\right).$$

Moreover,

$$\beta_{t,i} \leq \overline{\beta}_T := \sqrt{d \log\left(1 + \frac{T}{d}\right) + 2\log(2T)} + 1.$$

Let $\alpha_i \in (0,1)$ and $r_i \in (0,1]$, whose values will be specified later. Define

$$\mathcal{U}_i := \{t \in [T] : i_t = i,\ 2\beta_{t,i} a_{t,i} > \alpha_i \underline{\Delta}_i\}\,,$$

$$\mathcal{H}_i := \{t \in [T] : i_t = i,\ a_{t,i} > r_i\}\,,$$

and

$$\mathcal{M}_i := \{t \in [T] : i_t = i\} \setminus (\mathcal{U}_i \cup \mathcal{H}_i).$$

Bounding $|\mathcal{U}_i|$. For every $t \in \mathcal{U}_i$,

$$a_{t,i} > \frac{\alpha_i \underline{\Delta}_i}{2\beta_{t,i}} \geq \frac{\alpha_i \underline{\Delta}_i}{2\overline{\beta}_T}.$$

Hence,

$$|\mathcal{U}_i| \frac{\alpha_i^2 \underline{\Delta}_i^2}{4\overline{\beta}_T^2} \leq \sum_{t \in \mathcal{U}_i} a_{t,i}^2 \leq \Phi_T,$$

which gives

$$|\mathcal{U}_i| \leq \frac{4\overline{\beta}_T^2 \Phi_T}{\alpha_i^2 \underline{\Delta}_i^2}. \qquad (8)$$

Bounding $|\mathcal{H}_i|$. For every $t \in \mathcal{H}_i$, we have $a_{t,i}^2 > r_i^2$. Thus,

$$|\mathcal{H}_i| r_i^2 \leq \sum_{t \in \mathcal{H}_i} a_{t,i}^2 \leq \Phi_T,$$

and therefore

$$|\mathcal{H}_i| \leq \frac{\Phi_T}{r_i^2}. \tag{9}$$

Bounding $|\mathcal{M}_i|$. For every $t \in \mathcal{M}_i$, since $t \notin \mathcal{U}_i$,

$$2\beta_{t,i} a_{t,i} \leq \alpha_i \underline{\Delta}_i.$$

Combining this with equation 6 yields

$$m_{t,i} \geq (1 - \alpha_i)\underline{\Delta}_i. \tag{10}$$

Moreover, since $t \notin \mathcal{H}_i$,

$$a_{t,i} \leq r_i.$$

Hence,

$$\frac{m_{t,i}^2}{1 + a_{t,i}^2} \geq \frac{(1 - \alpha_i)^2 \underline{\Delta}_i^2}{1 + r_i^2}.$$

Summing over $t \in \mathcal{M}_i$ and applying Corollary 11 gives

$$\begin{aligned} |\mathcal{M}_i| \frac{(1 - \alpha_i)^2 \underline{\Delta}_i^2}{1 + r_i^2} &\leq \sum_{t \in \mathcal{M}_i} \frac{m_{t,i}^2}{1 + a_{t,i}^2} \\ &\leq \varepsilon_i^2 n_{T+1}(i). \end{aligned}$$

Therefore,

$$|\mathcal{M}_i| \leq c_i n_{T+1}(i), \tag{11}$$

where

$$c_i := \frac{(1 + r_i^2)\varepsilon_i^2}{(1 - \alpha_i)^2 \underline{\Delta}_i^2}. \tag{12}$$

Since

$$\{t \in [T] : i_t = i\} = \mathcal{U}_i \cup \mathcal{H}_i \cup \mathcal{M}_i,$$

we have

$$n_{T+1}(i) \leq |\mathcal{U}_i| + |\mathcal{H}_i| + |\mathcal{M}_i|.$$

Applying equation 8, equation 9, and equation 11, we obtain

$$n_{T+1}(i) \leq \Phi_T \left[\frac{4\overline{\beta}_T^2}{\alpha_i^2 \underline{\Delta}_i^2} + \frac{1}{r_i^2} \right] + c_i n_{T+1}(i).$$

Thus, whenever $c_i < 1$,

$$n_{T+1}(i) \leq \frac{\Phi_T}{1 - c_i} \left[\frac{4\overline{\beta}_T^2}{\alpha_i^2 \underline{\Delta}_i^2} + \frac{1}{r_i^2} \right]. \tag{13}$$

Recall that

$$\frac{\varepsilon_i}{\underline{\Delta}_i} = 1 - \kappa_i.$$

Choose

$$\alpha_i := \frac{\kappa_i}{4}, \qquad r_i := \sqrt{\frac{\kappa_i}{4}}.$$

Then

$$c_i = \frac{(1 + \kappa_i/4)(1 - \kappa_i)^2}{(1 - \kappa_i/4)^2},$$

and hence

$$1 - c_i = \frac{\kappa_i(20 - 7\kappa_i - 4\kappa_i^2)}{(4 - \kappa_i)^2}.$$

Since $\kappa_i \in (0, 1]$,

$$20 - 7\kappa_i - 4\kappa_i^2 \geq 9,$$

so $c_i < 1$.

Moreover,

$$\frac{1}{r_i^2} = \frac{4}{\kappa_i}, \qquad \frac{4\overline{\beta}_T^2}{\alpha_i^2 \underline{\Delta}_i^2} = \frac{64\overline{\beta}_T^2}{\kappa_i^2 \underline{\Delta}_i^2}.$$

Substituting these expressions into equation 13 yields, on $\mathcal{E}_i$,

$$n_{T+1}(i) \leq L_i(T),$$

where

$$L_i(T) = \frac{2d(4 - \kappa_i)^2 \log\left(1 + \frac{T}{d}\right)}{\kappa_i(20 - 7\kappa_i - 4\kappa_i^2)} \left[\frac{64\overline{\beta}_T^2}{\kappa_i^2 \underline{\Delta}_i^2} + \frac{4}{\kappa_i}\right].$$

Since

$$20 - 7\kappa_i - 4\kappa_i^2 \geq 9, \qquad (4 - \kappa_i)^2 \leq 16,$$

and

$$\Phi_T \in \widetilde{O}(d), \qquad \overline{\beta}_T^2 \in \widetilde{O}(d),$$

we have

$$L_i(T) \in \widetilde{O}\left(\frac{d^2}{\kappa_i^3 \underline{\Delta}_i^2} + \frac{d}{\kappa_i^2}\right).$$

Furthermore, since $d \geq 1$, $\kappa_i \leq 1$, and $\underline{\Delta}_i \leq 2$, the second term is absorbed by the first, and therefore

$$L_i(T) = \widetilde{O}\left(\frac{d^2}{\kappa_i^3 \underline{\Delta}_i^2}\right). \tag{14}$$

Since $n_{T+1}(i) \leq T$ deterministically,

$$\begin{aligned}
\mathbb{E}_{\sigma_1^{\mathrm{np}}, \boldsymbol{\sigma}_{-1}}[n_{T+1}(i)] &= \mathbb{E}_{\sigma_1^{\mathrm{np}}, \boldsymbol{\sigma_{-1}}}\left[n_{T+1}(i)\mathbf{1}\{\mathcal{E}_i\}\right] + \mathbb{E}_{\sigma_1^{\mathrm{np}}, \boldsymbol{\sigma_{-1}}}\left[n_{T+1}(i)\mathbf{1}\{\mathcal{E}_i^c\}\right] \\
&\leq L_i(T) + T\mathbb{P}(\mathcal{E}_i^c) \\
&\leq L_i(T) + 1.
\end{aligned}$$

Together with equation 14, this gives

$$\mathbb{E}_{\sigma_1^{\mathrm{np}}, \boldsymbol{\sigma}_{-1}}[n_{T+1}(i)] \in \widetilde{O}\left(\frac{d^2}{\kappa_i^3 \underline{\Delta}_i^2}\right).$$

Finally, since exactly one arm is selected in each round,

$$T - n_{T+1}(1) = \sum_{i=2}^{K} n_{T+1}(i).$$

Taking expectations and summing the preceding arm-wise bounds yields

$$\mathbb{E}_{\sigma_1^{\mathrm{np}}, \boldsymbol{\sigma}_{-1}}\left[T - n_{T+1}(1)\right] \leq \sum_{i=2}^{K} \big(L_i(T) + 1\big),$$

and consequently

$$\mathbb{E}_{\sigma_1^{\mathrm{np}}, \boldsymbol{\sigma}_{-1}}\left[T - n_{T+1}(1)\right] \in \widetilde{O}\left(d^2 \sum_{i=2}^{K} \frac{1}{\kappa_i^3 \underline{\Delta}_i^2}\right).$$

This completes the proof. □

C Proof of Theorem 7

Theorem 7. *The non-pessimistic strategy class Σ_i^{np} of arm i is defined as*

$$\Sigma_i^{\mathrm{np}} := \left\{ \sigma_i \in \Sigma_i(\varepsilon_i) : \mathbb{P}_{\sigma_i, \boldsymbol{\sigma}_{-i}} \left(m_{t,i} \geq 0,\ \forall t \in [T] \right) = 1,\ \forall \boldsymbol{\sigma}_{-i} \in \prod_{j \neq i} \Sigma_j(\varepsilon_j, \theta_j^*) \right\}$$

with $m_{t,i} = \varepsilon_{t,i} - x_{t,i}^\top V_{t,i}^{-1} \sum_{s<t} \varepsilon_{s,i} x_{s,i} \cdot \mathbf{1}(i_s = i)$, $0 \leq \varepsilon_i < \underline{\Delta}_i$, and $\theta_i^* \in \{\theta_i^* \in \mathbb{R}^d :\ \|\theta_i^*\|_2 \leq 1\}$ for all $i \neq i^*$. Then under Assumption 3, given Algorithm 1, any admissible joint strategy profile $\boldsymbol{\sigma}$ satisfying $\sigma_{i^*} \in \Sigma_{i^*}^{\mathrm{np}}$ forms a ρ_T-Nash equilibrium, where

$$\rho_T \in \widetilde{O}\left(d^2 \max\left\{ g_2 \sum_{j \neq i^*} \frac{1}{\kappa_j^3 \underline{\Delta}_j^2},\ \max_{i \neq i^*} \frac{(g_2 - \underline{\Delta}_i)_+ + (\overline{\Delta}_i - g_1)_+}{\kappa_i^3 \underline{\Delta}_i^2} \right\} \right),$$

$(x)_+ := \max\{x, 0\}$, *and* $\kappa_i = 1 - \varepsilon_i / \underline{\Delta}_i$.

Proof. For convenience, define

$$B_i(T) := L_i(T) + 1, \qquad i \neq 1,$$

where $L_i(T)$ is given in Lemma 13. We first prove the following bound on deviation incentive of arms for any admissible strategy profile $\boldsymbol{\sigma}$ with $\sigma_1 \in \Sigma_1^{\mathrm{np}}$:

$$\bar{\rho}_T := \max\left\{ g_2 \sum_{j=2}^{K} B_j(T),\ \max_{i \neq 1} \left[(g_2 - \underline{\Delta}_i)_+ + (\overline{\Delta}_i - g_1)_+ \right] B_i(T) \right\}.$$

Specifically, we show that for every arm $i \in [K]$ and every admissible alternative strategy σ_i',

$$U_i(\sigma_i', \boldsymbol{\sigma}_{-i}) - U_i(\boldsymbol{\sigma}) \leq \bar{\rho}_T. \tag{15}$$

For any strategy profile $\boldsymbol{\tau}$,

$$\begin{aligned} U_i(\boldsymbol{\tau}) &= \mathbb{E}_{\boldsymbol{\tau}} \left[\sum_{t=1}^{T} \mathbf{1}\{i_t = i\} (r_{t,i} - \eta) \right] \\ &= \mathbb{E}_{\boldsymbol{\tau}} \left[\sum_{t=1}^{T} \mathbf{1}\{i_t = i\} \big(\mu_i(c_t) - \eta \big) \right]. \end{aligned} \tag{16}$$

Since the arm selection i_t is determined before the reward noise ξ_{t,i_t} is observed, and the reward noise is conditionally 0-sub-gaussian. Hence,

$$\mathbb{E}_{\boldsymbol{\tau}} \left[\mathbf{1}\{i_t = i\} \xi_{t,i} \right] = 0.$$

We bound the incentive deviation of the optimal arm 1 and other arms separately.

The optimal arm $i^* = 1$. Fix any admissible strategy σ_1' of arm 1. By Assumption 3,

$$0 < g_1 \leq \mu_1(c) - \eta \leq g_2, \qquad \forall c \in \mathcal{C}. \tag{17}$$

Define the expected utility that arm 1 would obtain if it were selected at every round:

$$U_1^{\mathrm{full}} := \mathbb{E} \left[\sum_{t=1}^{T} \big(\mu_1(c_t) - \eta \big) \right].$$

Since the context process is sampled by external world, the distribution of $(c_t)_{t=1}^T$ and hence U_1^{full} do not depend on the arms' strategic behaviors.

Because $\mu_1(c_t) - \eta > 0$ at every round, selecting arm 1 in every round maximizes its cumulative utility conditioned on any realization of the context sequence. Therefore, for every admissible strategy σ_1',

$$U_1(\sigma_1', \boldsymbol{\sigma}_{-1}) \le U_1^{\text{full}}.$$

It follows that

$$\begin{aligned} &U_1(\sigma_1', \boldsymbol{\sigma}_{-1}) - U_1(\boldsymbol{\sigma}) \\ &\le U_1^{\text{full}} - U_1(\boldsymbol{\sigma}) \\ &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T} \mathbf{1}\{i_t \neq 1\}\big(\mu_1(c_t) - \eta\big)\right] \\ &\le g_2\, \mathbb{E}_{\boldsymbol{\sigma}}\left[T - n_{T+1}(1)\right]. \end{aligned} \tag{18}$$

Since the optimal arm follows a non-pessimistic strategy, $\sigma_1 \in \Sigma_1^{\text{np}}$, while the strategies of all suboptimal arms are arbitrary admissible strategies, Lemma 13 gives

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[T - n_{T+1}(1)\right] \le \sum_{j=2}^{K} B_j(T).$$

Substituting this bound into equation 18 yields

$$U_1(\sigma_1', \boldsymbol{\sigma}_{-1}) - U_1(\boldsymbol{\sigma}) \le g_2 \sum_{j=2}^{K} B_j(T). \tag{19}$$

A suboptimal arm $i \neq 1$. Fix any suboptimal arm $i \neq 1$ and any admissible strategy σ_i'. For every context $c \in \mathcal{C}$,

$$\mu_i(c) - \eta = \mu_1(c) - \eta - \big(\mu_1(c) - \mu_i(c)\big).$$

By Assumption 3 and the definitions of $\underline{\Delta}_i$ and $\overline{\Delta}_i$,

$$g_1 \le \mu_1(c) - \eta \le g_2,$$

and

$$\underline{\Delta}_i \le \mu_1(c) - \mu_i(c) \le \overline{\Delta}_i.$$

Hence,

$$g_1 - \overline{\Delta}_i \le \mu_i(c) - \eta \le g_2 - \underline{\Delta}_i, \qquad \forall c \in \mathcal{C}. \tag{20}$$

Define

$$u_i^+ := (g_2 - \underline{\Delta}_i)_+, \qquad u_i^- := (\overline{\Delta}_i - g_1)_+.$$

Then equation 20 implies

$$-u_i^- \le \mu_i(c) - \eta \le u_i^+, \qquad \forall c \in \mathcal{C}. \tag{21}$$

Consequently, for any strategy profile $\boldsymbol{\tau}$,

$$-u_i^-\, \mathbb{E}_{\boldsymbol{\tau}}[n_{T+1}(i)] \le U_i(\boldsymbol{\tau}) \le u_i^+\, \mathbb{E}_{\boldsymbol{\tau}}[n_{T+1}(i)]. \tag{22}$$

Let

$$\boldsymbol{\sigma}' := (\sigma_i', \boldsymbol{\sigma}_{-i}).$$

Since $i \neq 1$, the optimal arm does not deviate. Thus, $\sigma_1 \in \Sigma_1^{\text{np}}$ under both $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$, while the remaining arms follow arbitrary admissible strategies. Apply Lemma 13 to both strategy profiles and gives

$$\mathbb{E}_{\boldsymbol{\sigma}'}[n_{T+1}(i)] \le B_i(T), \qquad \mathbb{E}_{\boldsymbol{\sigma}}[n_{T+1}(i)] \le B_i(T). \tag{23}$$

Applying the upper bound in equation 22 to $\boldsymbol{\sigma}'$, and the lower bound to $\boldsymbol{\sigma}$, yields

$$U_i(\boldsymbol{\sigma}') \le u_i^+ B_i(T), \quad U_i(\boldsymbol{\sigma}) \ge -u_i^- B_i(T).$$

Therefore,

$$U_i(\sigma_i', \boldsymbol{\sigma}_{-i}) - U_i(\boldsymbol{\sigma}) \le (u_i^+ + u_i^-)\, B_i(T) \\ = \left[(g_2 - \underline{\Delta}_i)_+ + (\overline{\Delta}_i - g_1)_+\right] B_i(T). \tag{24}$$

Combining equation 19 and equation 24, we conclude that for every arm $i \in [K]$ and every admissible strategy profile σ_i',

$$U_i(\sigma_i', \boldsymbol{\sigma}_{-i}) - U_i(\boldsymbol{\sigma}) \le \bar{\rho}_T.$$

Hence, $\boldsymbol{\sigma}$ is a $\bar{\rho}_T$-Nash equilibrium.

It remains to relate this exact bound to the rate stated in the theorem. By Lemma 13, for every $i \neq 1$,

$$B_i(T) = L_i(T) + 1 \in \widetilde{O}\left(\frac{d^2}{\kappa_i^3 \underline{\Delta}_i^2}\right).$$

Therefore,

$$\bar{\rho}_T \in \widetilde{O}\left(d^2 \max\left\{ g_2 \sum_{j=2}^{K} \frac{1}{\kappa_j^3 \underline{\Delta}_j^2},\ \max_{i \neq 1} \frac{(g_2 - \underline{\Delta}_i)_+ + (\overline{\Delta}_i - g_1)_+}{\kappa_i^3 \underline{\Delta}_i^2} \right\}\right).$$

Taking $\rho_T := \bar{\rho}_T$ proves the claimed result. □

D Proof of Theorem 5

Theorem 5. *The admissible strategy space $\Sigma_i(\varepsilon_i, \theta_i^*)$ is defined as*

$$\Sigma_i(\varepsilon_i, \theta_i^*) := \left\{\sigma_i = \{x_{t,i}\}_{t=1}^T : \left|\langle \theta_i^*,\, x_{t,i}\rangle - \mu_i(c_t)\right| \le \varepsilon_i, \quad \forall t \in [T]\right\}$$

with $0 \le \varepsilon_i < \underline{\Delta}_i$ and $\theta_i^* \in \{\theta_i^* \in \mathbb{R}^d :\ \|\theta_i^*\|_2 \le 1\}$ for all $i \neq i^*$. Then under Assumption 3, given any joint strategy profile $\boldsymbol{\sigma} \in \prod_{i \in [K]} \Sigma_i(\varepsilon_i, \theta_i^*)$, Algorithm 1 satisfies

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[T - n_{T+1}(i^*)\right] \in \widetilde{O}\left(\frac{g_2}{g_1} d^2 \sum_{i \neq i^*} \frac{1}{\kappa_i^3 \underline{\Delta}_i^2} + \frac{\mathrm{Expl}_{i^*}(\boldsymbol{\sigma})}{g_1}\right),$$

and consequently,

$$R_T(\boldsymbol{\sigma}, \mathrm{PALinUCB}) \in \widetilde{O}\left(\Delta_{\max} \frac{g_2}{g_1} d^2 \sum_{i \neq i^*} \frac{1}{\kappa_i^3 \underline{\Delta}_i^2} + \frac{\Delta_{\max}}{g_1} \mathrm{Expl}_{i^*}(\boldsymbol{\sigma})\right),$$

where $\kappa_i = 1 - \varepsilon_i / \underline{\Delta}_i$, $\quad \Delta_{\max} = \max_{i \neq i^*} \overline{\Delta}_i$.

Proof. Let $\boldsymbol{\sigma}$ be an arbitrary admissible strategy profile of the strategic game induced by Algorithm 1. Define

$$B_T := \sum_{i=2}^{K} \big(L_i(T) + 1\big),$$

where $L_i(T)$ is defined in Lemma 13.

We call an admissible strategy σ_i^{cal} of arm i calibrated if it induces zero strategic deviation, i.e., $\varepsilon_{t,i} = 0$ for all $t \in [T]$. We define the set of all calibrated strategies for arm i as

$$\Sigma_i^{\mathrm{cal}} = \{\sigma_i \in \Sigma_i(\varepsilon_i, \theta_i^*) : \varepsilon_{t,i} = 0,\ \forall t \in [T]\}.$$

It follows directly from the definitions of Σ_i^{cal} and Σ_i^{np} that $\Sigma_i^{\mathrm{cal}} \subseteq \Sigma_i^{\mathrm{np}}$.

The proof proceeds by using a calibrated strategy of the optimal arm as a benchmark strategy. Fix any

$$\widetilde{\sigma}_1 \in \Sigma_1^{\mathrm{cal}},$$

and let

$$\widetilde{\boldsymbol{\sigma}} := \big(\widetilde{\sigma}_1, \boldsymbol{\sigma}_{-1}\big).$$

Since arm 1 follows a calibrated strategy under $\widetilde{\boldsymbol{\sigma}}$, while the strategies of all suboptimal arms remain arbitrary, Lemma 13 implies

$$\mathbb{E}_{\widetilde{\boldsymbol{\sigma}}}\left[T - n_{T+1}(1)\right] \le B_T. \tag{25}$$

We next translate this pull-count guarantee into a utility guarantee for arm 1. Define its full-selection utility as

$$U_1^{\mathrm{full}} := \mathbb{E}\left[\sum_{t=1}^{T}\big(\mu_1(c_t) - \eta\big)\right].$$

This is the expected utility that arm 1 would obtain if it were selected at every round. Since the context process is exogenous, its distribution does not depend on the reporting strategies. Hence U_1^{full} is independent of the strategy profile.

For any strategy profile $\boldsymbol{\tau}$, the definition of utility gives

$$U_1(\boldsymbol{\tau}) = \mathbb{E}_{\boldsymbol{\tau}}\left[\sum_{t=1}^{T}\mathbf{1}\{i_t = 1\}(r_{t,1} - \eta)\right].$$

Because i_t is determined before the reward noise at round t is observed and the reward noise is conditionally zero-mean,

$$\mathbb{E}_{\boldsymbol{\tau}}\left[\mathbf{1}\{i_t = 1\}\xi_{t,1}\right] = 0.$$

Therefore,

$$U_1(\boldsymbol{\tau}) = \mathbb{E}_{\boldsymbol{\tau}}\left[\sum_{t=1}^{T}\mathbf{1}\{i_t = 1\}\big(\mu_1(c_t) - \eta\big)\right].$$

Consequently,

$$U_1^{\mathrm{full}} - U_1(\boldsymbol{\tau}) = \mathbb{E}_{\boldsymbol{\tau}}\left[\sum_{t=1}^{T}\mathbf{1}\{i_t \neq 1\}\big(\mu_1(c_t) - \eta\big)\right]. \tag{26}$$

Applying equation 26 to $\widetilde{\boldsymbol{\sigma}}$ and using Assumption 3,

$$\mu_1(c) - \eta \le g_2, \qquad \forall c \in \mathcal{C},$$

we obtain

$$\begin{aligned} U_1^{\mathrm{full}} - U_1(\widetilde{\boldsymbol{\sigma}}) &= \mathbb{E}_{\widetilde{\boldsymbol{\sigma}}}\left[\sum_{t=1}^{T}\mathbf{1}\{i_t \neq 1\}\big(\mu_1(c_t) - \eta\big)\right] \\ &\le g_2\mathbb{E}_{\widetilde{\boldsymbol{\sigma}}}\left[T - n_{T+1}(1)\right] \\ &\le g_2 B_T, \end{aligned} \tag{27}$$

where the last inequality follows from equation 25.

By the definition of $\mathrm{Expl}_1(\boldsymbol{\sigma})$, the gain of arm 1 by altering its strategy from σ_1 to $\widetilde{\sigma}_1$ is bounded by

$$U_1(\widetilde{\boldsymbol{\sigma}}) - U_1(\boldsymbol{\sigma}) \le \mathrm{Expl}_1(\boldsymbol{\sigma}).$$

Equivalently,

$$U_1(\boldsymbol{\sigma}) \ge U_1(\widetilde{\boldsymbol{\sigma}}) - \mathrm{Expl}_1(\boldsymbol{\sigma}).$$

Therefore,

$$\begin{aligned} U_1^{\mathrm{full}} - U_1(\boldsymbol{\sigma}) &\le U_1^{\mathrm{full}} - U_1(\widetilde{\boldsymbol{\sigma}}) + \mathrm{Expl}_1(\boldsymbol{\sigma}) \\ &\le g_2 B_T + \mathrm{Expl}_1(\boldsymbol{\sigma}), \end{aligned} \tag{28}$$

where the last inequality follows from equation 27.

On the other hand, applying equation 26 to $\boldsymbol{\sigma}$ and using the lower utility bound

$$\mu_1(c) - \eta \ge g_1, \qquad \forall c \in \mathcal{C},$$

gives

$$U_1^{\text{full}} - U_1(\boldsymbol{\sigma}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T} \mathbf{1}\{i_t \neq 1\}\big(\mu_1(c_t) - \eta\big)\right] \geq g_1 \mathbb{E}_{\boldsymbol{\sigma}}\left[T - n_{T+1}(1)\right]. \tag{29}$$

Combining equation 28 and equation 29, we obtain the pull-count bound

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[T - n_{T+1}(1)\right] \leq \frac{g_2 B_T + \text{Expl}_1(\boldsymbol{\sigma})}{g_1} = \frac{g_2}{g_1}\sum_{i=2}^{K}\big(L_i(T) + 1\big) + \frac{\text{Expl}_1(\boldsymbol{\sigma})}{g_1}. \tag{30}$$

By Lemma 13,

$$L_i(T) + 1 \in \widetilde{O}\left(\frac{d^2}{\kappa_i^3 \underline{\Delta}_i^2}\right).$$

Therefore,

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[T - n_{T+1}(1)\right] \in \widetilde{O}\left(\frac{g_2}{g_1} d^2 \sum_{i=2}^{K} \frac{1}{\kappa_i^3 \underline{\Delta}_i^2} + \frac{\text{Expl}_1(\boldsymbol{\sigma})}{g_1}\right). \tag{31}$$

It remains to bound the strategic regret. Since arm 1 is uniformly optimal, for every suboptimal arm $i \neq 1$ and every $c \in \mathcal{C}$,

$$0 < \mu_1(c) - \mu_i(c) \leq \overline{\Delta}_i \leq \Delta_{\max}.$$

Hence,

$$\begin{aligned} R_T(\boldsymbol{\sigma}, \text{PALinUCB}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T}\big(\mu_1(c_t) - \mu_{i_t}(c_t)\big)\right] \\ &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T}\mathbf{1}\{i_t \neq 1\}\big(\mu_1(c_t) - \mu_{i_t}(c_t)\big)\right] \\ &\leq \Delta_{\max}\mathbb{E}_{\boldsymbol{\sigma}}\left[T - n_{T+1}(1)\right]. \end{aligned}$$

Combining this with equation 31 gives

$$R_T(\boldsymbol{\sigma}, \text{PALinUCB}) \in \widetilde{O}\left(\Delta_{\max}\frac{g_2}{g_1} d^2 \sum_{i=2}^{K} \frac{1}{\kappa_i^3 \underline{\Delta}_i^2} + \frac{\Delta_{\max}}{g_1}\text{Expl}_1(\boldsymbol{\sigma})\right),$$

which proves the claimed result. □

The equilibrium-regret guarantee in Theorem 5 is conditioned on equilibrium existence. We next prove that a NE exists on a non-trivial class of instances

Proposition 14. *Fix any $K \geq 2$, $d \geq 2$, and horizon T. Suppose the reward noise is zero and*

$$0 < \inf_{c \in \mathcal{C}} \mu_1(c),$$

and there exists η such that

$$\max_{i \neq 1} \sup_{c \in \mathcal{C}} \mu_i(c) \leq \eta < \inf_{c \in \mathcal{C}} \mu_1(c).$$

Suppose further that, for every $i \neq 1$,

$$\sup_{c \in \mathcal{C}} |\mu_i(c)| \leq \varepsilon_i < \Delta_i.$$

Then the strategic game induced by PALinUCB reaches a Nash equilibrium $\boldsymbol{\sigma}^\star$. Moreover, PALinUCB selects arm 1 in every round under $\boldsymbol{\sigma}^\star$, and hence

$$R_T(\boldsymbol{\sigma}^\star, \text{PALinUCB}) = 0.$$

Proof. Let $v \in \mathbb{R}^d$ be any unit vector. Consider a strategy profile $\boldsymbol{\sigma}^\star$ defined as follows. Arm 1 chooses

$$\theta_1^\star = v$$

and, at every round t, reports

$$x_{t,1} = \mu_1(c_t)v.$$

Every suboptimal arm $i \neq 1$ chooses

$$\theta_i^\star = \mathbf{0}$$

and always reports

$$x_{t,i} = \mathbf{0}.$$

We first verify that $\sigma^\star$ is admissible. Since $\|\theta^\star\|_2 \leq 1$ and $\|x_1^c\|_2 \leq 1$, we have $|\mu_1(c)| \leq 1$ for every $c \in \mathcal{C}$. Hence,

$$\|x_{t,1}\|_2 = |\mu_1(c_t)| \leq 1.$$

Moreover,

$$\langle \theta_1^\star, x_{t,1} \rangle = \langle v, \mu_1(c_t)v \rangle = \mu_1(c_t),$$

so arm 1 is calibrated:

$$\left| \langle \theta_1^\star, x_{t,1} \rangle - \mu_1(c_t) \right| = 0.$$

For every $i \neq 1$,

$$\left| \langle \theta_i^\star, \mathbf{0} \rangle - \mu_i(c_t) \right| = |\mu_i(c_t)| \leq \varepsilon_i,$$

and $\|\mathbf{0}\|_2 \leq 1$. Therefore all reports under $\sigma^\star$ are admissible.

We next show that PALinUCB selects arm 1 at every round. Suppose inductively that arm 1 has been selected in the first $t-1$ rounds. Define

$$A_{t-1} := \sum_{s=1}^{t-1} \mu_1(c_s)^2.$$

Since the instance is noiseless,

$$r_{s,1} = \mu_1(c_s),$$

and arm 1 reports $x_{s,1} = \mu_1(c_s)v$. Hence,

$$V_{t,1} = I_d + A_{t-1} v v^\top,$$

and

$$b_{t,1} = \sum_{s=1}^{t-1} x_{s,1} r_{s,1} = A_{t-1} v.$$

Therefore,

$$\widehat{\theta}_{t,1} = V_{t,1}^{-1} b_{t,1} = \frac{A_{t-1}}{1 + A_{t-1}} v.$$

Furthermore,

$$\|x_{t,1}\|_{V_{t,1}^{-1}} = \frac{\mu_1(c_t)}{\sqrt{1 + A_{t-1}}},$$

where we used $\mu_1(c_t) > 0$. Thus,

$$\begin{aligned} \mathrm{UCB}_{t,1} &= \langle \widehat{\theta}_{t,1}, x_{t,1} \rangle + \beta_{t,1} \|x_{t,1}\|_{V_{t,1}^{-1}} \\ &= \mu_1(c_t) \left(\frac{A_{t-1}}{1 + A_{t-1}} + \frac{\beta_{t,1}}{\sqrt{1 + A_{t-1}}} \right) > 0. \end{aligned}$$

On the other hand, every suboptimal arm reports the zero vector, and hence

$$\mathrm{UCB}_{t,i} = 0, \qquad i \neq 1.$$

Therefore arm 1 is uniquely selected at round t. Since the same argument also applies at $t = 1$, induction shows that arm 1 is selected in every round.

It remains to verify that no arm has a profitable deviation. Under $\boldsymbol{\sigma}^\star$, arm 1 is selected at every round, so

$$U_1(\boldsymbol{\sigma}^\star) = \mathbb{E}\left[\sum_{t=1}^{T}\big(\mu_1(c_t) - \eta\big)\right].$$

By assumption,

$$\mu_1(c) - \eta > 0, \qquad \forall c \in \mathcal{C}.$$

Hence, for any admissible alternative strategy σ_1',

$$\begin{aligned} U_1(\sigma_1', \boldsymbol{\sigma}^\star_{-1}) &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{i_t = 1\}\big(\mu_1(c_t) - \eta\big)\right] \\ &\le \mathbb{E}\left[\sum_{t=1}^{T}\big(\mu_1(c_t) - \eta\big)\right] = U_1(\boldsymbol{\sigma}^\star). \end{aligned}$$

Thus arm 1 cannot profitably deviate.

For every suboptimal arm $i \neq 1$, its equilibrium utility is zero because it is never selected. Under any admissible unilateral deviation σ_i',

$$\begin{aligned} U_i(\sigma_i', \boldsymbol{\sigma}^\star_{-i}) &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{i_t = i\}\big(\mu_i(c_t) - \eta\big)\right] \\ &\le 0 = U_i(\boldsymbol{\sigma}^\star), \end{aligned}$$

where the inequality follows from

$$\mu_i(c) - \eta \le 0, \qquad \forall c \in \mathcal{C}.$$

Therefore no suboptimal arm has a profitable alternative strategy.

It follows that $\boldsymbol{\sigma}^\star$ is an Nash equilibrium. Since PALinUCB selects the uniformly optimal arm 1 in every round,

$$R_T(\boldsymbol{\sigma}^\star) = 0.$$

□

E Negative Results: When PALinUCB Fails

The regret guarantee at Nash equilibrium in Theorem 5 relies on the assumption that the optimal arm uniformly benefits from being selected. We now show that this incentive condition cannot simply be removed. In particular, consider the regime

$$\eta > \sup_{c \in \mathcal{C}} \mu_1(c),$$

under which even the uniformly optimal arm receives negative utility whenever it is selected. Since arm 1 is uniformly optimal, this further implies $\mu_i(c) - \eta < 0$ for every arm i and context c. In this all-negative-utility regime, all arms prefer to avoid being selected. We show that, despite keeping the strict separation condition $\varepsilon_i < \underline{\Delta}_i$ for every suboptimal arm, PALinUCB can admit exact Nash equilibria with linear regret.

Example 15. *Consider a noiseless instance with fixed $K \ge 2$, $d \ge 2$, and horizon T. Let $v, u \in \mathbb{R}^d$ be orthogonal unit vectors and set $\theta^\star = v$. Choose constants $0 < \mu_2 < \mu_1 < \eta < 1$ and, for $i \ge 3$, $0 < \mu_i < \min\{\mu_2, \mu_1/2\}$. Consider a single context c with true features $x_i^c = \mu_i v + \sqrt{1-\mu_i^2}\, u$. Then $\mu_i(c) = \mu_i$ and $\underline{\Delta}_i = \mu_1 - \mu_i$.*

Set $\varepsilon_1 = \mu_1$ and $\theta_1^\star = 0$. For arm 2, let $\varepsilon_2 := \min\left\{\frac{\underline{\Delta}_2}{2}, \frac{\mu_2}{2(1+\sqrt{T})}\right\}, \theta_2^\star = (\mu_2 - \varepsilon_2)v$. For each $i \ge 3$, set $\varepsilon_i = \mu_i$ and $\theta_i^\star = 0$. Then $0 < \varepsilon_i < \underline{\Delta}_i$ for every $i \neq 1$.

Consider the strategy profile $\boldsymbol{\sigma}^\star$ under which, at every round,

$$x_{t,1} = 0, \qquad x_{t,2} = v, \qquad x_{t,i} = 0 \text{ for } i \ge 3.$$

Under PALinUCB, $\boldsymbol{\sigma}^\star$ is an Nash equilibrium with $n_{T+1}(2) = T$. Consequently,

$$R_T(\boldsymbol{\sigma}^\star) = \underline{\Delta}_2 T = \Theta(T).$$

Proof. We first verify that the strategy profile $\boldsymbol{\sigma}^\star$ is admissible. For arm 1 and every $i \geq 3$, since $\theta_i^\star = 0$ and $x_{t,i} = 0$,

$$|\langle \theta_i^\star, x_{t,i} \rangle - \mu_i| = \mu_i = \varepsilon_i.$$

For arm 2,

$$|\langle \theta_2^\star, v \rangle - \mu_2| = \varepsilon_2.$$

Hence all prescribed reports are admissible.

We next show that arm 2 cannot avoid being selected under any alternative strategy. Consider any admissible alternative strategy $\sigma_2' = (\phi, \pi_2')$, where $\|\phi\|_2 \leq 1$. Conditioning on any realization of the internal randomization of π_2', let z_t denote its report at round t, and define

$$e_t := \langle \phi, z_t \rangle - \mu_2, \qquad |e_t| \leq \varepsilon_2.$$

Let

$$S_t := \sum_{s<t : i_s = 2} e_s z_s, \qquad a_t := \|z_t\|_{V_{t,2}^{-1}}.$$

Since the instance is noiseless and the reward of arm 2 is always μ_2,

$$\widehat{\theta}_{t,2} = \phi - V_{t,2}^{-1}\phi - V_{t,2}^{-1} S_t.$$

Therefore,

$$\mathrm{UCB}_{t,2} = \mu_2 + e_t - z_t^\top V_{t,2}^{-1}\phi - z_t^\top V_{t,2}^{-1} S_t + \beta_{t,2} a_t. \tag{32}$$

Because $V_{t,2} \succeq I_d$, $\|\phi\|_2 \leq 1$, and $\beta_{t,2} \geq 1$,

$$-z_t^\top V_{t,2}^{-1}\phi + \beta_{t,2} a_t \geq 0.$$

Moreover,

$$\|S_t\|_{V_{t,2}^{-1}} \leq \varepsilon_2 \sqrt{n_t(2)}.$$

Indeed, if Z_t denotes the matrix whose rows are the previously selected reports of arm 2, and e is the corresponding vector of deviations, then

$$S_t = Z_t^\top e, \qquad V_{t,2} = I_d + Z_t^\top Z_t,$$

and hence

$$\|S_t\|_{V_{t,2}^{-1}}^2 = e^\top Z_t (I_d + Z_t^\top Z_t)^{-1} Z_t^\top e \leq \|e\|_2^2 \leq \varepsilon_2^2 n_t(2).$$

Since $a_t \leq 1$, equation 32 gives

$$\mathrm{UCB}_{t,2} \geq \mu_2 - \varepsilon_2 - \varepsilon_2 \sqrt{n_t(2)} \geq \mu_2 - \varepsilon_2(1 + \sqrt{T}) \geq \frac{\mu_2}{2} > 0,$$

where the last inequality follows from the choice of ε_2 in Example 15.

Under $\boldsymbol{\sigma}_{-2}^\star$, every other arm reports the zero vector and therefore has PALinUCB index zero at every round. Thus arm 2 is uniquely selected at every round under any admissible alternative strategy σ_2'. In particular,

$$n_{T+1}(2) = T$$

under $\boldsymbol{\sigma}^\star$, and every alternative strategy of arm 2 yields the same utility

$$U_2(\sigma_2', \boldsymbol{\sigma}_{-2}^\star) = T(\mu_2 - \eta) = U_2(\boldsymbol{\sigma}^\star).$$

Hence arm 2 has no profitable alternative strategy.

For every arm $i \neq 2$, its utility under NE is zero because it is never selected. Since $\eta > \mu_1 > \mu_i$, every selection gives arm i strictly negative utility. Consequently, under any alternative joint strategy profile $(\sigma_i', \boldsymbol{\sigma}_{-i}^*)$,

$$U_i(\sigma_i', \boldsymbol{\sigma}_{-i}^\star) \leq 0 = U_i(\boldsymbol{\sigma}^\star).$$

Thus no arm has a profitable alternative strategy, and $\boldsymbol{\sigma}^\star$ is an Nash equilibrium.

Finally, since arm 2 is selected in every round,

$$R_T(\boldsymbol{\sigma}^\star) = (\mu_1 - \mu_2)T = \underline{\Delta}_2 T = \Theta(T).$$

□

Example 15 shows that the strategic separation condition alone does not guarantee sublinear regret at a Nash Equilibrium. The failure is driven by the utility design: when all arms prefer to avoid selection, a suboptimal arm may be selected throughout the horizon and be unable to profitably escape selection. The formal proof is deferred to Appendix E. We summarize this failure through the following impossibility result.

Theorem 16. *Fix any $K \geq 2$ and $d \geq 2$. Let $\mathfrak{I}_-(K,d)$ be the class of strategic linear contextual bandit instances with a unique uniformly optimal arm $i^\star$ satisfying $\eta > \sup_{c\in\mathcal{C}} \mu_{i^\star}(c)$ and $0 < \varepsilon_i < \underline{\Delta}_i$ for every $i \neq i^\star$. Then there exists a constant $c > 0$, independent of T, such that for every horizon T there exists a noiseless instance $\mathcal{I}_T \in \mathfrak{I}_-(K,d)$ and a strategy profile $\boldsymbol{\sigma}_T^\star$ forming an exact Nash equilibrium satisfying*

$$R_T(\boldsymbol{\sigma}_T^\star) \geq cT.$$

Consequently, no $o(T)$ regret bound can hold uniformly over all exact Nash equilibria given PALinUCB on $\mathfrak{I}_-(K,d)$.

Proof. Fix any $K \geq 2$, $d \geq 2$, and horizon T. Consider the instance $\mathcal{I}_T$ constructed in Example 15, where the reward means $\{\mu_i\}_{i=1}^K$ are chosen independently of T, while ε_2 is chosen according to the example.

Arm 1 is the unique uniformly optimal arm, $\eta > \mu_1$, and, for every $i \neq 1$,

$$0 < \varepsilon_i < \underline{\Delta}_i.$$

Hence $\mathcal{I}_T \in \mathfrak{I}_-(K,d)$.

By Example 15, the T-round game induced by PALinUCB admits an Nash equilibrium $\boldsymbol{\sigma}_T^\star$ satisfying $n_{T+1}(2) = T$. Therefore,

$$R_T(\boldsymbol{\sigma}_T^\star) = \underline{\Delta}_2 T.$$

Since μ_1 and μ_2 are fixed independently of T, the constant

$$c := \underline{\Delta}_2 = \mu_1 - \mu_2 > 0$$

is independent of T. Thus, for every horizon T, there exist $\mathcal{I}_T \in \mathfrak{I}_-(K,d)$ and an Nash equilibrium $\boldsymbol{\sigma}_T^\star$ such that

$$R_T(\boldsymbol{\sigma}_T^\star) \geq cT.$$

It remains to rule out a uniform sublinear guarantee. Suppose there exists $f(T) = o(T)$ such that

$$R_T(\boldsymbol{\sigma}) \leq f(T)$$

for every instance in $\mathfrak{I}_-(K,d)$ and every Nash equilibrium of the corresponding T-round PALinUCB game. Since $f(T)/T \to 0$, for all sufficiently large T we have $f(T) < cT$. Applying the assumed bound to $(\mathcal{I}_T, \boldsymbol{\sigma}_T^\star)$ gives

$$cT \leq R_T(\boldsymbol{\sigma}_T^\star) \leq f(T),$$

a contradiction. Hence no uniform $o(T)$ regret guarantee exists. □

F Additional Experimental Details

F.1 Experimental Setup and Hyperparameters

Bandit instance. We consider a synthetic linear contextual bandit with $K = 7$ arms and dimension $d = 2$. Let $z_i \in \mathbb{R}^d$ denote the base feature vector of arm i. At the beginning of each run, we sample a length-T context sequence $\{c_t\}_{t=1}^T$, where each coordinate is drawn independently from $\mathrm{Unif}[0.5, 1]$. The true contextual feature of arm i at round t is constructed as

$$x_i^{c_t} = c_t \odot z_i,$$

where $\odot$ denotes element-wise multiplication. We set $\theta^\star = (1,0)^\top$ and use the following base arm features:

$$\begin{aligned} &z_1 = (2,0)^\top, \quad z_2 = (1.8,1)^\top, \quad z_3 = (1.7,0)^\top, \\ &z_4 = (1.6,0)^\top, \quad z_5 = (1.4,1)^\top, \quad z_6 = (1.2,0)^\top, \\ &z_7 = (1.0,1)^\top. \end{aligned}$$

The feature vectors in this experiment are not normalized to unit norm. As noted in Section 2, the unit-norm constraint is adopted only for notational convenience, and the analysis extends to any uniformly bounded feature radius. The reward is generated according to $r_{t,i} = \langle \theta^\star, x_i^{c_t} \rangle + \xi_{t,i}$, where $\xi_{t,i} \sim \mathcal{N}(0, 0.1^2)$.

Under this construction, arm 1 is uniformly optimal. In particular, $\mu_1(c) = 2c_1$, and hence

$$\inf_{c \in \mathcal{C}} \mu_1(c) = 1, \qquad \sup_{c \in \mathcal{C}} \mu_1(c) = 2.$$

Algorithms and strategic settings. We evaluate PALinUCB under strategic reporting and compare it with Vanilla LinUCB, truthful versions of both algorithms, and OptGTM. For PALinUCB and Vanilla LinUCB with strategic arms, each selected arm receives utility $r_{t,i} - \eta$. For OptGTM, we consider two strategic objectives. In the utility-maximizing variant, arms optimize the same cumulative utility as under PALinUCB.In the pull-count variant, following the original OptGTM setting, arms instead maximize their number of selections. The detailed strategic-behavior simulation used to update the reported features is described in Algorithm 2.

We consider the per-epoch horizons

$$T \in \{2000, 5000, 10000, 15000, 20000\},$$

and use $\eta \in \{0.5, 0.9, 1.2, 1.5, 2.2, 2.5\}$ for the horizon-scaling experiments. For the finer utility sweep, we fix $T = 10000$ and consider

$$\eta \in \{0, 0.5, 0.9, 1.0, 1.1, 1.2, 1.3, 1.5, 1.8, 2.0, 2.2, 2.5\}.$$

Unless otherwise specified, the main cross-algorithm comparison uses $\eta = 1.5$.

Evaluation protocol. Each run consists of $E = 20$ strategic-update epochs, with T learner–arm interactions in each epoch. Let $R_T^{(e)}$ denote the learner's regret incurred during epoch e. Since our goal is to evaluate learning performance throughout the strategic adaptation process, our primary regret metric is

$$R_{\text{sum}} := \sum_{e=1}^{E} R_T^{(e)}.$$

We additionally report the normalized regret $R_{\text{sum}}/(ET)$ and the optimal-arm selection rate

$$\frac{1}{ET} \sum_{e=1}^{E} n_{T+1}^{(e)}(1).$$

All reported curves are averaged over 300 independent runs, and the shaded regions represent one standard deviation across runs.

F.2 Strategic Behavior Simulation

Our theoretical results characterize strategic behavior through equilibrium concepts over a general strategy space. Numerically solving for an equilibrium over the full history-dependent strategy space is, however, computationally intractable. We therefore simulate adaptive strategic behavior through repeated utility-improving updates of a low-dimensional strategy parameterization.

Strategy parameterization. For each arm i, let $z_i^\star \in \mathbb{R}^d$ denote its base feature information. Given context c_t, its true contextual feature is

$$x_i^{c_t} = c_t \odot z_i^\star,$$

where $\odot$ denotes element-wise multiplication. At strategic-update epoch e, arm i maintains a reported base information feature $z_i^{(e)}$. Its report to the learner at round t is

$$x_{t,i}^{(e)} = c_t \odot z_i^{(e)}.$$

Thus, the reported feature is fixed as a function of the context within each epoch, while the base report $z_i^{(e)}$ evolves across epochs. We initialize the strategic interaction truthfully, $z_i^{(0)} = z_i^\star$ for every arm i.

Importantly, strategic reporting affects only the information observed by the learner. The reward is always generated from the true contextual feature,

$$r_{t,i} = \langle \theta^\star, x_i^{c_t} \rangle + \xi_{t,i},$$

whereas the learner constructs its estimator and confidence index using the reported feature $x_{t,i}^{(e)}$.

Epoch-wise strategic adaptation. Each experiment consists of E strategic-update epochs. During epoch e, the reported features $\{z_i^{(e)}\}_{i=1}^K$ are held fixed and the learner interacts with the arms for T rounds. For PALinUCB, the selected arm is

$$i_t \in \arg\max_{i\in[K]} \left\{ \langle \widehat{\theta}_{t,i}, x_{t,i}^{(e)} \rangle + \beta_{t,i} \|x_{t,i}^{(e)}\|_{V_{t,i}^{-1}} \right\}.$$

The learner observes the realized reward of the selected arm and updates only the corresponding per-arm estimator.

For an arm whose selection utility is $r_{t,i} - \eta$, its cumulative utility during an epoch is

$$U_i^{(e)} := \sum_{t=1}^{T} \mathbf{1}\{i_t = i\}(r_{t,i} - \eta).$$

After the T-round interaction, each arm updates its reported base feature in the direction of increasing cumulative utility. We approximate this direction using coordinate-wise finite differences.

Finite-difference utility ascent. Let $\delta_{\mathrm{fd}} > 0$ denote the nominal perturbation magnitude. For coordinate j of arm i, we choose a direction $s_{i,j}^{(e)} \in \{-1, +1\}$ such that the projected perturbation is nonzero, and define

$$\widetilde{z}_{i,j}^{(e)} = \Pi_{\mathcal{Z}_i}\left(z_i^{(e)} + s_{i,j}^{(e)} \delta_{\mathrm{fd}} e_j \right),$$

where e_j is the jth standard basis vector. Since the projection may shorten the perturbation near the boundary of $\mathcal{Z}_i$, we define the actual signed displacement by

$$h_{i,j}^{(e)} := \left[\widetilde{z}_{i,j}^{(e)}\right]_j - \left[z_i^{(e)}\right]_j.$$

All other arms keep their current reports unchanged. The corresponding utility is estimated by averaging M independent rollouts, denoted by $\widehat{U}_i(\widetilde{z}_{i,j}^{(e)}, z_{-i}^{(e)})$. Using the same learner state and matched rollouts for the unperturbed profile gives the finite-difference estimate

$$\widehat{g}_{i,j}^{(e)} = \frac{\widehat{U}_i(\widetilde{z}_{i,j}^{(e)}, z_{-i}^{(e)}) - \widehat{U}_i(z_i^{(e)}, z_{-i}^{(e)})}{h_{i,j}^{(e)}}.$$

We then update the reported feature according to

$$z_i^{(e+1)} = \Pi_{\mathcal{Z}_i}\left(z_i^{(e)} + \Delta_i^{(e)} \right),$$

where

$$\Delta_i^{(e)} := \mathrm{clip}\left(\alpha_e \widehat{g}_i^{(e)}, -0.1, 0.1 \right)$$

is applied coordinate-wise, and

$$\alpha_e = \frac{10^{-4}\alpha_0}{e+1}.$$

The projection ensures that every updated report remains in the admissible reporting domain $\mathcal{Z}_i$, while the clipping prevents unstable changes caused by noisy finite-difference estimates.

Our main experiments use $E = 20$, $M = 5$, $\delta_{\mathrm{fd}} = 0.2$, and base learning-rate parameter $\alpha_0 = 0.2$. The complete procedure is summarized in Algorithm 2.

Alternative strategic objective. For PALinUCB and Vanilla LinUCB, strategic arms optimize the cumulative utility induced by our utility design, namely $U_i^{(e)} = \sum_t \mathbf{1}\{i_t = i\}(r_{t,i} - \eta)$. For the OptGTM baseline, we follow its standard strategic objective and replace $U_i^{(e)}$ by the number of selections,

$$U_{i,\text{pull}}^{(e)} = \sum_{t=1}^{T} \mathbf{1}\{i_t = i\}.$$

The same finite-difference adaptation procedure is then applied with respect to this objective.

For each arm i, we set $\theta_i^\star = e_1$ and define its reported-base-feature domain as

$$\mathcal{Z}_i = [z_{i,1}^\star - b_i, z_{i,1}^\star + b_i] \times [-1, 1].$$

Since $x_{t,i} = c_t \odot z_i$ and $\mu_i(c_t) = c_{t,1} z_{i,1}^\star$, every $z_i \in \mathcal{Z}_i$ satisfies

$$|\langle \theta_i^\star, x_{t,i} \rangle - \mu_i(c_t)| = c_{t,1} |z_{i,1} - z_{i,1}^\star| \le b_i.$$

Hence the simulated reporting strategies are admissible with $\varepsilon_i = b_i$. For every suboptimal arm, we set $b_i = \Delta_i/2$, so that $\varepsilon_i < \Delta_i$.

Algorithm 2 Finite-Difference Simulation of Strategic Arm Adaptation

Require: Learner $\mathcal{L}$; horizon T; epochs E; arms $[K]$; contexts $\{c_t\}_{t=1}^T$; true base features $\{z_i^\star\}_{i=1}^K$; domains $\{\mathcal{Z}_i\}_{i=1}^K$; perturbation δ_{fd}; rollouts M; base step size α_0.

Initialize $z_i^{(0)} \leftarrow z_i^\star$ for all $i \in [K]$.
for $e = 0, \dots, E-1$ **do**
  Set $x_{t,i}^{(e)} \leftarrow c_t \odot z_i^{(e)}$ for all $t \in [T]$, $i \in [K]$, and save the initial learner state as $\mathcal{S}_e$.
  Run one T-round interaction from $\mathcal{S}_e$ under $\boldsymbol{z}^{(e)}$; record $R_T^{(e)}$ and advance the actual learner state.
  Estimate $\widehat{U}_i^{(e)}$ for all $i \in [K]$ by averaging M rollouts from $\mathcal{S}_e$, and set $\alpha_e \leftarrow 10^{-4}\alpha_0/(e+1)$.
  for $i = 1, \dots, K$ **do**
    for $j = 1, \dots, d$ **do**
      Choose $s_{i,j}^{(e)} \in \{-1, +1\}$ such that the projected perturbation is nonzero, and set

$$\widetilde{z}_{i,j}^{(e)} \leftarrow \Pi_{\mathcal{Z}_i}\left(z_i^{(e)} + s_{i,j}^{(e)} \delta_{\text{fd}} \mathbf{e}_j\right).$$

      Let

$$h_{i,j}^{(e)} \leftarrow [\widetilde{z}_{i,j}^{(e)}]_j - [z_i^{(e)}]_j.$$

      Estimate $\widehat{U}_{i,j}^{(e),\text{pert}}$ by averaging M rollouts from $\mathcal{S}_e$ under $(\widetilde{z}_{i,j}^{(e)}, \boldsymbol{z}_{-i}^{(e)})$, and set

$$\widehat{g}_{i,j}^{(e)} \leftarrow \frac{\widehat{U}_{i,j}^{(e),\text{pert}} - \widehat{U}_i^{(e)}}{h_{i,j}^{(e)}}.$$

    end for
    Set

$$\Delta_i^{(e)} \leftarrow \text{clip}\left(\alpha_e \widehat{g}_i^{(e)}, -0.1, 0.1\right), \qquad z_i^{(e+1)} \leftarrow \Pi_{\mathcal{Z}_i}\left(z_i^{(e)} + \Delta_i^{(e)}\right),$$

    where clipping is coordinate-wise.
  end for
end for
return $\{R_T^{(e)}\}_{e=0}^{E-1}$ and $\{\boldsymbol{z}^{(e)}\}_{e=0}^{E}$.

F.3 Additional Experiments

We compare PALinUCB against existing baseline algorithms. When LinUCB (Li et al., 2010) and LinTS (Agrawal & Goyal, 2013) were initally designed for contextual bandits when arms report true features to the learner, we evaluate them in both this setting and the strategic setting studied in this

work. We also evaluate the OptGTM (Kleine Buening et al., 2024) algorithm which was designed for strategic bandits where arms aim to maximize their pull counts. Although our theoretical analysis relies on Assumption 3, we evaluate all algorithms with different values of the contract parameter η that either satisfy or violate this assumption. We observe from Figure 5 that the performance of all algorithms deteriorates when the assumption is violated, highlighting the necessity of this assumption for the learner to achieve a low regret.

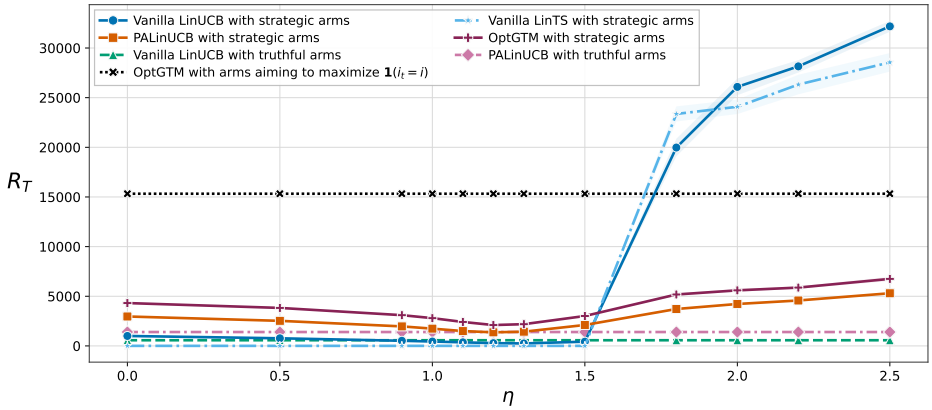


Figure 5: Effect of η on cumulative regret of PALinUCB and baseline algorithms.

Contract design and separate arms' estimators both help PALinUCB keep robust. In Figure 5, we vary the contract parameter η with $T = 10000$. First, Figure 5 indicates that OptGTM enjoys substantially lower regret when arms aim to maximize contract utility $r_{t,i} - \eta$ rather than pull-counts $\mathbf{1}(i_t = i)$, which suggests that our contract design shapes the arms' incentives in a way that can offer greater strategic advantage to the learner for regret minimization. Second, PALinUCB and OptGTM, both of which maintain a separate estimator for each arm, remain more robust across different values of η when strategic arms maximize the contract utility $r_{t,i} - \eta$. In contrast, LinUCB and LinTS maintain a global estimator across all arms and incur higher regret, suggesting that the separate estimator for each arm helps algorithms remain more robust against strategic arms. Third, PALinUCB consistently achieves lower regret than OptGTM across all values of η. In specific, this gap may be partly due to OptGTM's elimination mechanism, whereas PALinUCB does not eliminate arms. Specially, OptGTM employs a Grim Trigger Condition (GTC) to detect arms which have excessive strategic deviations and permanently removes an arm once the condition is triggered. The GTC in OptGTM may lead to a mistakenly elimination of a relatively good arm, which can incur a constant regret in every future rounds. In contrast, PALinUCB retains all arms throughout the interaction, which helps it remain robust as the contract parameter η varies.

G Comparison with Existing Strategic Contextual Bandit Algorithms

For completeness, we compare the regret guarantees of PALinUCB with existing algorithms for strategic linear contextual bandits. We use K, d, and T to denote the number of arms, the feature dimension and the horizon, respectively.

Table 1 summarizes regret guarantees of existing algorithms and their assumptions respectively. Existing theoretical guarantees for OptGTM and COBRA is $\widetilde{O}(\sqrt{T})$ under NE. In contrast, under our assumptions that $0 \leq \varepsilon_i < \underline{\Delta}_i$, PALinUCB achieves $\mathrm{poly}(\log T)$ regret at any Nash equilibrium. More generally, our guarantee does not condition on that the strategy profile forms a Nash equilibrium: for any admissible joint strategy profile of arms, the regret is largely determined by the strategy of the optimal arm i^*.

Table 1: Comparison of regret guarantees for strategic linear contextual bandits. We use K, d, and T uniformly for the number of arms (or strategic agents), feature dimension, and horizon. The notation $\widetilde{O}(\cdot)$ suppresses logarithmic factors in T.

Method	Strategy regime	Regret guarantee	Additional condition
OptGTM	Truthful approx. NE	$\widetilde{O}\big(d\sqrt{KT}\big)$	Truthful reporting
OptGTM	Conditional on existing NE	$\widetilde{O}\big(dK^2\sqrt{KT}\big)$	$\varepsilon_i = 0, \forall i \in [K]$
COBRA	Truthful approx. NE	$\widetilde{O}\big(d\sqrt{T}\big)$	Truthful reporting
COBRA	Conditional on existing NE	$\widetilde{O}\big(d\sqrt{T} + \sqrt{KT}\big)$	$\mathrm{UCB}_{t,i}(x_{t,i}) \geq \mu_{t,i}, \forall i \in [K]$
PALinUCB (ours)	Arbitrary admissible σ	$\widetilde{O}\left(\frac{g_2}{g_1} d^2 \sum_{i \neq i^*} \frac{\Delta_{\max}}{\kappa_i^3 \Delta_i^2} + \frac{\Delta_{\max}}{g_1} \mathrm{Expl}_{i^*}(\boldsymbol{\sigma})\right)$	$\varepsilon_i < \underline{\Delta}_i$, $g_1 > 0$, uniformly optimal arm
PALinUCB (ours)	Conditional on existing NE	$\widetilde{O}\left(\frac{g_2}{g_1} d^2 \sum_{i \neq i^*} \frac{\Delta_{\max}}{\kappa_i^3 \Delta_i^2}\right)$	$\varepsilon_i < \underline{\Delta}_i$, $g_1 > 0$, uniformly optimal arm

H A General Regret Guarantee under Incentive Design

In this section, we extend the regret guarantee of PALiUCB to a more general case and derive a more general bound that characterizes the relationship between the optimal arm's strategic behaviors and the learner's regret from selecting suboptimal arms.

H.1 A General Incentive–Regret Alignment Bound

For each context $c \in \mathcal{C}$, define the *selection incentive margin* of the uniformly optimal arm by

$$h_\eta(c) := \mu_1(c) - \eta, \tag{33}$$

and define the largest instantaneous regret gap at context c by

$$\Gamma(c) := \max_{i \neq 1} \big\{\mu_1(c) - \mu_i(c)\big\}. \tag{34}$$

We further let

$$\bar{g}_\eta := \sup_{c \in \mathcal{C}} \big[h_\eta(c)\big]_+, \tag{35}$$

where $[x]_+ := \max\{x, 0\}$.

Recall the quantity $L_i(T)$ from Lemma 13, and define

$$B_T := \sum_{i=2}^{K} \big(L_i(T) + 1\big). \tag{36}$$

By Lemma 13, when $0 \leq \varepsilon_i < \underline{\Delta}_i, \quad \forall i \neq i^*$,

$$B_T = \widetilde{O}\left(d^2 \sum_{i=2}^{K} \frac{1}{\kappa_i^3 \Delta_i^2}\right). \tag{37}$$

We now give a regret guarantee that does not require the uniformly optimal arm to benefit from selection under every context.

Theorem 17. *Suppose that $0 \leq \varepsilon_i < \underline{\Delta}_i, \forall i \neq i^*$. For any admissible strategy profile $\boldsymbol{\sigma}$ and any $\lambda \geq 0$,*

$$R_T(\boldsymbol{\sigma}, \mathrm{PALinUCB}) \leq \lambda \left(\bar{g}_\eta B_T + \mathrm{Expl}_1(\boldsymbol{\sigma})\right) + \mathcal{A}_T(\lambda, \eta), \tag{38}$$

where

$$\mathcal{A}_T(\lambda, \eta) := \mathbb{E}\left[\sum_{t=1}^{T} \left[\Gamma(c_t) - \lambda h_\eta(c_t)\right]_+\right]. \tag{39}$$

Consequently,

$$R_T(\sigma, \mathrm{PALinUCB}) \leq \inf_{\lambda \geq 0} \Big\{\lambda \left(\bar{g}_\eta B_T + \mathrm{Expl}_1(\sigma)\right) + \mathcal{A}_T(\lambda, \eta)\Big\}. \tag{40}$$

Proof. Fix an arbitrary calibrated strategy $\widetilde{\sigma}_1 \in \Sigma_1^{\mathrm{cal}}$ of arm 1, and let

$$\widetilde{\sigma} := (\widetilde{\sigma}_1, \sigma_{-1}).$$

Since every calibrated strategy is non-pessimistic, Lemma 13 yields

$$\mathbb{E}_{\widetilde{\sigma}}\left[T - n_{T+1}(1)\right] \le B_T. \tag{41}$$

Define the full-selection utility of arm 1 as

$$U_1^{\mathrm{full}} := \mathbb{E}\left[\sum_{t=1}^{T} h_\eta(c_t)\right]. \tag{42}$$

As the context process is exogenous, its distribution is independent of the arms' reporting strategies, and hence U_1^{full} is independent of the strategy profile.

By the definition of arm 1's utility,

$$\begin{aligned} U_1^{\mathrm{full}} - U_1(\widetilde{\boldsymbol{\sigma}}) &= \mathbb{E}_{\widetilde{\sigma}}\left[\sum_{t=1}^{T} \mathbf{1}\{i_t \neq 1\} h_\eta(c_t)\right] \\ &\le \bar{g}_\eta\, \mathbb{E}_{\widetilde{\boldsymbol{\sigma}}}\left[T - n_{T+1}(1)\right] \\ &\le \bar{g}_\eta B_T, \end{aligned} \tag{43}$$

where the last inequality follows from equation 41.

By the definition of the unilateral exploitability of arm 1,

$$U_1(\widetilde{\boldsymbol{\sigma}}) - U_1(\boldsymbol{\sigma}) \le \mathrm{Expl}_1(\boldsymbol{\sigma}). \tag{44}$$

Combining equation 43 and equation 44, we obtain

$$U_1^{\mathrm{full}} - U_1(\boldsymbol{\sigma}) \le \bar{g}_\eta B_T + \mathrm{Expl}_1(\boldsymbol{\sigma}). \tag{45}$$

We next relate this utility loss to the learner's regret. For any $\lambda \ge 0$ and any round t, if $i_t \neq 1$, then

$$\mu_1(c_t) - \mu_{i_t}(c_t) \le \Gamma(c_t).$$

Moreover,

$$\Gamma(c_t) \le \lambda h_\eta(c_t) + \left[\Gamma(c_t) - \lambda h_\eta(c_t)\right]_+.$$

Therefore,

$$\begin{aligned} &\mathbf{1}\{i_t \neq 1\}\big(\mu_1(c_t) - \mu_{i_t}(c_t)\big) \\ &\quad \le \lambda \mathbf{1}\{i_t \neq 1\} h_\eta(c_t) + \left[\Gamma(c_t) - \lambda h_\eta(c_t)\right]_+. \end{aligned} \tag{46}$$

Summing equation 46 over $t = 1, \ldots, T$ and taking expectations gives

$$\begin{aligned} R_T(\boldsymbol{\sigma}, \mathrm{PALinUCB}) &\le \lambda \mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T} \mathbf{1}\{i_t \neq 1\} h_\eta(c_t)\right] + \mathcal{A}_T(\lambda, \eta) \\ &= \lambda\left(U_1^{\mathrm{full}} - U_1(\boldsymbol{\sigma})\right) + \mathcal{A}_T(\lambda, \eta) \\ &\le \lambda\left(\bar{g}_\eta B_T + \mathrm{Expl}_1(\boldsymbol{\sigma})\right) + \mathcal{A}_T(\lambda, \eta), \end{aligned} \tag{47}$$

where the last inequality follows from equation 45. This proves equation 38. Since the bound holds for every $\lambda \ge 0$, taking the infimum over λ proves equation 40. □

H.2 Discussion

We first show that Theorem 17 recovers the regret guarantee in the uniformly positive utility regime.

Corollary 18. *Suppose Assumptions 3 and $0 \le \varepsilon_i < \underline{\Delta}_i$ hold for any arm $i \neq i^*$. Then, for any admissible strategy profile $\boldsymbol{\sigma}$,*

$$R_T(\boldsymbol{\sigma}, \mathrm{PALinUCB}) \le \frac{\Delta_{\max} g_2}{g_1} B_T + \frac{\Delta_{\max}}{g_1} \mathrm{Expl}_1(\boldsymbol{\sigma}). \tag{48}$$

Consequently,

$$R_T(\boldsymbol{\sigma}, \mathrm{PALinUCB}) = \widetilde{O}\left(\frac{\Delta_{\max} g_2}{g_1} d^2 \sum_{i=2}^{K} \frac{1}{\kappa_i^3 \Delta_i^2} \right) + \frac{\Delta_{\max}}{g_1} \mathrm{Expl}_1(\boldsymbol{\sigma}), \tag{49}$$

which recovers the result of Theorem 5.

Proof. Under Assumption 3,

$$0 < g_1 \le h_\eta(c) \le g_2, \qquad \forall c \in \mathcal{C}.$$

Moreover, by the definition of $\Delta_{\max}$,

$$\Gamma(c) \le \Delta_{\max}, \qquad \forall c \in \mathcal{C}.$$

Choose

$$\lambda = \frac{\Delta_{\max}}{g_1}.$$

Then, for every $c \in \mathcal{C}$,

$$\Gamma(c) - \lambda h_\eta(c) \le \Delta_{\max} - \frac{\Delta_{\max}}{g_1} g_1 = 0.$$

Hence,

$$\mathcal{A}_T(\lambda, \eta) = 0.$$

Also,

$$\bar{g}_\eta \le g_2.$$

Applying Theorem 17 with this choice of λ gives equation 48. Finally, substituting equation 37 yields equation 49. □

We next give a more interpretable specialization that quantifies deviations from a uniformly positive incentive margin. For any reference margin $g > 0$, define the cumulative incentive deficit

$$D_T(g, \eta) := \mathbb{E}\left[\sum_{t=1}^{T} [g - h_\eta(c_t)]_+ \right]. \tag{50}$$

Corollary 19. *Suppose that $0 \le \varepsilon_i < \underline{\Delta}_i, \forall i \neq i^*$ For any admissible strategy profile $\boldsymbol{\sigma}$ and any $g > 0$,*

$$R_T(\boldsymbol{\sigma}, \mathrm{PALinUCB}) \le \frac{\Delta_{\max}}{g} \left(\bar{g}_\eta B_T + \mathrm{Expl}_1(\boldsymbol{\sigma}) + D_T(g, \eta) \right). \tag{51}$$

Consequently,

$$R_T(\boldsymbol{\sigma}, \mathrm{PALinUCB}) \le \inf_{g > 0} \frac{\Delta_{\max}}{g} \left(\bar{g}_\eta B_T + \mathrm{Expl}_1(\boldsymbol{\sigma}) + D_T(g, \eta) \right). \tag{52}$$

Proof. Fix any $g > 0$ and choose

$$\lambda = \frac{\Delta_{\max}}{g}.$$

Since $\Gamma(c) \le \Delta_{\max}$ for every $c \in \mathcal{C}$,

$$\begin{aligned} \mathcal{A}_T(\lambda, \eta) &= \mathbb{E}\left[\sum_{t=1}^{T} \left[\Gamma(c_t) - \frac{\Delta_{\max}}{g} h_\eta(c_t) \right]_+ \right] \\ &\le \mathbb{E}\left[\sum_{t=1}^{T} \left[\Delta_{\max} - \frac{\Delta_{\max}}{g} h_\eta(c_t) \right]_+ \right] \\ &= \frac{\Delta_{\max}}{g} D_T(g, \eta). \end{aligned} \tag{53}$$

Applying Theorem 17 with $\lambda = \Delta_{\max}/g$ gives equation 51. Taking the infimum over $g > 0$ proves equation 52. □

Corollary 19 shows that a uniformly positive selection incentive is sufficient but not necessary for low regret. More generally, the relevant quantity is the context-dependent alignment between the optimal arm's selection incentive $h_\eta(c)$ and the learner's loss from selecting a suboptimal arm. The deficit $D_T(g, \eta)$ measures the cumulative shortfall of the optimal arm's incentive relative to a reference margin g. Thus, whenever this deficit is small, the regret guarantee can remain favorable even when $h_\eta(c)$ is not uniformly bounded away from zero.

In particular, if contexts with $h_\eta(c) < 0$ occur persistently and are associated with non-negligible regret gaps, then the alignment penalty in equation 39, and correspondingly the cumulative incentive deficit in equation 50, may scale linearly with T. Hence, the regret behavior depends on how the optimal arm's selection incentive $\mu_1(c) - \eta$ varies with the reward loss $\max_{i \neq 1}\{\mu_1(c) - \mu_i(c)\}$ across contexts.