Strategic Linear Contextual Bandits

Thomas Kleine Buening
The Alan Turing Institute

Aadirupa Saha
Apple ML Research

Christos Dimitrakakis
University of Neuchatel

Haifeng Xu
University of Chicago

Abstract

Motivated by the phenomenon of strategic agents gaming a recommender system to maximize the number of times they are recommended to users, we study a strategic variant of the linear contextual bandit problem, where the arms can strategically misreport privately observed contexts to the learner. We treat the algorithm design problem as one of *mechanism design* under uncertainty and propose the Optimistic Grim Trigger Mechanism (OptGTM) that incentivizes the agents (i.e., arms) to report their contexts truthfully while simultaneously minimizing regret. We also show that failing to account for the strategic nature of the agents results in linear regret. However, a trade-off between mechanism design and regret minimization appears to be unavoidable. More broadly, this work aims to provide insight into the intersection of online learning and mechanism design.

1 Introduction

Recommendation algorithms that select the most relevant item for sequentially arriving users or queries have become vital for navigating the internet and its many online platforms. However, recommender systems are susceptible to manipulation by strategic agents seeking to artificially increase their frequency of recommendation [31, 33, 38]. These agents, ranging from sellers on platforms like Amazon to websites aiming for higher visibility in search results, employ tactics such as altering product attributes or engage in aggressive search engine optimization [29, 32]. By gaming the algorithms, agents attempt to appear more relevant than they actually are, often compromising the integrity and intended functionality of the recommender system. Here, the key issue lies in the agents' incentive to manipulate the learning algorithm to maximize their utility (i.e., profit). To address this challenge, we study and design algorithms in a strategic variant of the linear contextual bandit, where the agents (i.e., arms) have the ability to misreport privately observed contexts to the learner. Our main contribution is connecting online learning with approximate mechanism design to minimize regret while, at the same time, discouraging the arms from gaming our learning algorithm.

The contextual bandit [2, 24] is a generalization of the multi-armed bandit problem to the case where the learner observes relevant contextual information before pulling an arm. It has found application in various domains including healthcare [37] and online recommendation [25]. We here focus on the linearly realizable setting [1, 6], where each arm's reward is a linear function of the arm's context in the given round. In the standard linear contextual bandit, at the beginning of round t, the learner observes the context $x^*_{t,i} \in \mathbb{R}^d$ of every arm $i \in [K]$, selects an arm i_t, and receives a reward drawn from a distribution with mean $\langle \theta^*, x^*_{t,i_t} \rangle$ where $\theta^* \in \mathbb{R}^d$ is an unknown parameter. In the *strategic linear contextual bandit*, we assume that each arm is a self-interested agent that wants to maximize the number of times it gets pulled by manipulating its contexts.

More precisely, we consider the situation where each arm i *privately* observes its true context $x^*_{t,i}$ every round, e.g., its relevance to the user arriving in round t, but reports a potentially gamed context vector $x_{t,i}$ to the learner. The learner does not observe the true contexts, but only the reported contexts

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3703

$\mathcal{X}_t = \{x_{t,1}, \ldots, x_{t,K}\}$ and chooses an action from this gamed action set $\mathcal{X}_t$. When the learner pulls arm i_t, the learner then observes a reward r_{t,i_t} drawn from a distribution with mean $\langle \theta^*, x^*_{t,i_t} \rangle$. In other words, the arms can manipulate the contexts the learner observes, but cannot influence the underlying reward. This is often the case as superficially changing attributes or meta data has no effect on an item's true relevance to a user.

In summary, our contributions are:

- We introduce a strategic variant of the linear contextual bandit problem, where each arm, in every round, can misreport its context to the learner to maximize its utility, defined as the total number of times the learner selects the arm over T rounds (Section 3). We demonstrate that incentive-unaware algorithms, which do not explicitly consider the incentives they (implicitly) create for the arms, suffer linear regret in this strategic setting when the arms respond in Nash Equilibrium (NE) (Proposition 3.3). This highlights the necessity of integrating mechanism design with online learning techniques to minimize regret in the presence of strategic arms.
- We begin with the case where θ^* is known to the learner in advance (Section 4). This simplifies the problem setup, allowing us to establish fundamental concepts while highlighting the challenges of designing a sequential mechanism *without* payments. For this scenario, we propose the Greedy Grim Trigger Mechanism (GGTM), which incentivizes the arms to be approximately truthful while minimizing regret. We show that **(a)** Truthful reporting is an $\tilde{\mathcal{O}}(\sqrt{T})$-NE for the arms (Theorem 4.1) and **(b)** GGTM has $\tilde{\mathcal{O}}(K^2\sqrt{KT})$ regret under *every* NE of the arms (Theorem 4.2).
- Next, we consider the case where θ^* is unknown to the learner in advance (Section 5). Without access to the true contexts, estimating θ^* accurately appears intractable, as the arms can manipulate the estimation process. Surprisingly, we show that learning θ^* is not necessary for minimizing regret in the strategic linear contextual bandit. We construct confidence sets (which may not contain θ^*) to derive pessimistic and optimistic estimates of our expected reward. These estimates are used to construct the Optimistic Grim Trigger Mechanism (OptGTM). Despite possibly incorrect estimates of θ^*, OptGTM bounds the impact of misreported contexts on both regret *and* the utility of all arms. We show that **(a)** Truthfulness is an $\tilde{\mathcal{O}}(d\sqrt{KT})$-NE for which OptGTM has regret $\tilde{\mathcal{O}}(d\sqrt{KT})$ (Theorem 5.1) and **(b)** OptGTM incurs at most $\tilde{\mathcal{O}}(dK^2\sqrt{KT})$ regret under *every* NE of the arms (Theorem 5.2).
- Finally, we support our theoretical findings with simulations of strategic gaming behavior in response to OptGTM and LinUCB (Section 6). We simulate how strategic arms adapt what contexts to report over time by equipping the arms with decentralized gradient ascent and letting the arms (e.g., vendors) and the learner (e.g., platform) repeatedly interact over several epochs. The experiments confirm the effectiveness of OptGTM and illustrate the shortcomings of incentive-unaware algorithms, such as LinUCB.

2 Related Work

Linear Contextual Bandits. In related work on linear contextual bandits with adversarial *reward* corruptions [3, 20, 39, 45], an adversary corrupts the reward observation in round t by some amount c_t but not the observed contexts. In this problem, the optimal regret is given by $\Theta(d\sqrt{T} + dC)$, where $C := \sum_t |c_t|$ is the adversary's budget. To the best of our knowledge, adversarial *context* corruptions have only been studied by [8], who achieve $\tilde{\mathcal{O}}(d\tilde{C}\sqrt{T})$ regret with $\tilde{C} := \sum_{t,i} \|x^*_{t,i} - x_{t,i}\|$, where $x^*_{t,i}$ and $x_{t,i}$ are the true and corrupted contexts, respectively. In contrast, we do not assume a bounded corruption budget so that these regret guarantees become vacuous (cf. Proposition 3.3). Moreover, instead of taking the worst-case perspective of purely adversarial manipulation, we assume that each arm is a self-interested agent maximizing their own utility.

Strategic Multi-Armed Bandits. Braverman et al. [4] were the first to study a strategic variant of the multi-armed bandit problem and considered the case where the pulled arm privately receives the reward and shares only a fraction of it with the learner. An extension of this setting has recently been studied in [12]. In other lines of work, [9, 13] study the robustness of bandit learning against strategic manipulation, however, simply assume a bounded manipulation budget instead of performing mechanism design. [11, 35] consider multi-armed bandits with replicas where strategic agents can submit replicas of the same arm to increase the number of times one of their arms is pulled. Buening et al. [5] combine multi-armed bandits with mechanism design to discourage clickbait in online

Interaction Protocol 1: Strategic Linear Contextual Bandits

Learner publicly commits to algorithm M
for $t = 1, \ldots, T$ **do**
  Every arm $i \in [K]$ privately observes its context $x^*_{t,i} \in \mathbb{R}^d$
  Every arm $i \in [K]$ reports a (potentially gamed) context $x_{t,i} \in \mathbb{R}^d$ to the learner
  Learner observes the gamed contexts $\mathcal{X}_t = \{x_{t,1}, \ldots, x_{t,K}\}$, selects arm $i_t \in [K]$, and receives reward

$$r_{t,i_t} := \langle \theta^*, x^*_{t,i_t} \rangle + \eta_t,$$

  where η_t is zero-mean sub-Gaussian noise. Note that the reward is generated with respect to the unknown parameter $\theta^* \in \mathbb{R}^d$ and the unobserved true context x^*_{t,i_t}.

recommendation. In their model, each arm maximizes its total number of clicks and is characterized by a strategically chosen click-rate and a fixed post-click reward. However, all of these works substantially differ from our work in problem setup and/or methodology.

Modeling Incentives in Recommender Systems. A complementary line of work studies content creator incentives in recommender systems [16, 21, 22, 23, 27, 40, 41, 42] and how algorithms shape the behavior of agents more generally [7]. These works primarily focus on modeling content creator behavior and studying content creator incentives under existing algorithms. Instead, our goal is the design of incentive-aware learning algorithms which incentivize content creators (arms) to act in a desirable fashion (truthfully) while maximizing the recommender system's performance.

Strategic Learning. We also want to mention the extensive literature on strategic learning [14, 15, 18, 19, 26, 43, 44] and strategic classification [10, 17, 34, 36]. Similarly to the model we study in this paper, the premise is that rational agents strategically respond to the learner's algorithm (e.g., classifier) to obtain a desired outcome. However, the learner interacts with the agents only once and the agents are assumed to be myopic and to suffer a cost for, e.g., altering their features. Moreover, there is no competition among the agents like in the strategic linear contextual bandit. In contrast to these works, we wish to design a sequential (online learning) mechanism to incentivize truthfulness, which is only possible because we repeatedly interact with the same set of agents (i.e., arms).

3 Strategic Linear Contextual Bandits

We study a strategic-variant of the linear contextual bandit problem, where K strategic agents (i.e., arms) aim to maximize their number of pulls by misreporting privately observed contexts to the learner. The learner follows the usual objective of minimizing regret, i.e., maximizing cumulative rewards, despite not observing the true contexts. We here focus on the case where the strategic arms respond in *Nash equilibrium* to the learning algorithm. The interaction between the environment, the learning algorithm, and the arms is specified in Interaction Protocol 1.

Notice that the arms can manipulate the contexts that the learner observes (and the learner only observes these gamed contexts and never the actual contexts), but the rewards are generated w.r.t. the true contexts. Moreover, if all arms are non-strategic and—irrespective of the learning algorithm—report their features truthfully every round, i.e., $x_{t,i} = x^*_{t,i}$ for all $(t, i) \in [T] \times [K]$, the problem reduces to the standard non-strategic linear contextual bandit.

3.1 Strategic Arms and Nash Equilibrium

We assume that each arm i reports a possibly gamed context $x_{t,i}$ to the learner after observing its true context $x^*_{t,i}$ and potentially other information. For example, the arms may have prior knowledge of θ^* and observe the identity of the selected arm at the end of each round. However, we do not use any specific assumptions about the observational model of the arms. Our results can be viewed as a worst-case analysis over all such models. For concreteness, consider the case where the arms have prior knowledge of θ^* and, at the end of every round, observe which arm was selected and the generated reward.[1]

[1]We naturally expect that the more information the arms observe each round, the more challenging the problem becomes for the learner, as the arms' ability to manipulate the learning algorithm increases.

 https://doi.org/10.52202/079017-3703

Let σ_i be a (mixed) strategy of arm i that is history-dependent and in every round t maps from observed true contexts $x^*_{t,i}$ to a distribution over reported contexts $x_{t,i}$ in $\mathbb{R}^d$. We define σ_{-i} as the strategies of all arms except i and define a strategy profile of the arms as $\boldsymbol{\sigma} \coloneqq (\sigma_1, \dots, \sigma_K)$. We call arm i *truthful* if it truthfully reports its privately context every round, i.e., $x_{t,i} = x^*_{t,i}$ for all $t \in [T]$. This truthful strategy is denoted σ^*_i and we let $\boldsymbol{\sigma}^* = (\sigma^*_1, \dots, \sigma^*_K)$.

We now formally define the objective of the arms. Let $n_T(i) \coloneqq \sum_{t=1}^{T} \mathbb{1}\{i_t = i\}$ be the number of times arm i is pulled by the learner's algorithm M. The objective of every arm is to maximize the expected number of times it is pulled by the algorithm given by

$$u_i(M, \boldsymbol{\sigma}) \coloneqq \mathbb{E}_M\big[n_T(i) \mid \boldsymbol{\sigma}\big],$$

where we condition on the arm strategies $\boldsymbol{\sigma}$ as these will (typically) impact the algorithm's decisions. We assume that the arms respond to the learning algorithm M in Nash Equilibrium (NE).

Definition 3.1 (Nash Equilibrium)**.** We say that $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_K)$ forms a NE under the learner's algorithm M if for all $i \in [K]$ and any deviating strategy σ'_i:

$$\mathbb{E}_M\big[n_T(i) \mid \sigma_i, \sigma_{-i}\big] \geq \mathbb{E}_M\big[n_T(i) \mid \sigma'_i, \sigma_{-i}\big].$$

Let $\mathrm{NE}(M) \coloneqq \{\boldsymbol{\sigma} \colon \boldsymbol{\sigma} \text{ is a NE under } M\}$ be the set of NE under algorithm M. We also consider ε-NE, which relax the requirement that no arm has an incentive to deviate.

Definition 3.2 (ε-Nash Equilibrium)**.** We say that $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_K)$ forms a ε-NE under algorithm M if for all $i \in [K]$ and any deviating strategy σ':

$$\mathbb{E}_M\big[n_T(i) \mid \sigma_i, \sigma_{-i}\big] \geq \mathbb{E}_M\big[n_T(i) \mid \sigma'_i, \sigma_{-i}\big] - \varepsilon.$$

3.2 Strategic Regret

In the strategic linear contextual bandit, the performance of an algorithm depends on the arm strategies that it incentivizes. Naturally, minimizing regret when the arms always report their context truthfully is easier than when contexts are manipulated adversarially. We are interested in the *strategic regret* of an algorithm M when the arms act according to a Nash equilibrium under M. Formally, for $\boldsymbol{\sigma} \in \mathrm{NE}(M)$ the strategic regret of M is defined as

$$R_T(M, \boldsymbol{\sigma}) = \mathbb{E}_{M, \boldsymbol{\sigma}}\left[\sum_{t=1}^{T} \langle \theta^*, x^*_{t,i^*_t} \rangle - \langle \theta^*, x^*_{t,i_t} \rangle\right],$$

where $i^*_t = \operatorname{argmax}_{i \in [K]} \langle \theta^*, x^*_{t,i} \rangle$ is the optimal arm in round t. The regret guarantees of our algorithms hold uniformly over all NE that they induce, i.e., for $\max_{\boldsymbol{\sigma} \in \mathrm{NE}(M)} R_T(M, \boldsymbol{\sigma})$.

Regularity Assumptions. We allow for the true context vectors $x^*_{t,i}$ to be chosen adversarially by nature, and make the following assumptions about the linear contextual bandit model. We assume that both the context vectors and the rewards are bounded, i.e., $\max_{i,j \in [K]} \langle \theta^*, x^*_{t,i} - x^*_{t,j} \rangle \leq 1$ and $\|x^*_{t,i}\|_2 \leq 1$ for all $t \in [T]$. Moreover, we assume a constant optimality gap. That is, letting $\Delta_{t,i} \coloneqq \langle \theta^*, x^*_{t,i^*_t} - x^*_{t,i} \rangle$, we assume that $\Delta \coloneqq \min_{t,i:\Delta_{t,i}>0} \Delta_{t,i}$ is constant.

3.3 The Necessity of Mechanism Design

The first question that arises in this strategic setup is whether mechanism design, i.e., actively aligning the arms' incentives, is necessary to minimize regret. As expected, we find that this is the case. Standard algorithms, which are oblivious to the incentives they create, implicitly incentivize the arms to heavily misreport their contexts which makes minimizing regret virtually impossible.

We call a problem instance *trivial* if the algorithm that selects an arm uniformly at random every round achieves sublinear regret. Conversely, we call a problem instance *non-trivial* if the uniform selection suffers linear expected regret. We show that being incentive-unaware generally leads to linear regret in non-trivial instances (even when the learner has prior knowledge of θ^*).

Proposition 3.3. *On any non-trivial problem instance, the incentive-unaware greedy algorithm that in round t plays $i_t = \operatorname{argmax}_{i \in [K]} \langle \theta^*, x_{t,i} \rangle$ (with ties broken uniformly) suffers linear regret $\Omega(T)$ when the arms act according to any Nash equilibrium under the incentive-unaware greedy algorithm. Note that the incentive-unaware greedy algorithm has knowledge of θ^*.*

Mechanism 1: The Greedy Grim Trigger Mechanism (GGTM)

initialize: $A_1 = [K]$
for $t = 1, \ldots, T$ **do**
  Observe reported contexts $x_{t,1}, \ldots, x_{t,K}$
  Play the (active) arm with largest reported reward: $i_t = \text{argmax}_{i \in A_t} \langle \theta^*, x_{t,i} \rangle$
  Observe reward r_{t,i_t} from playing arm i_t.
  if $\sum_{\ell \leq t:\ i_\ell = i_t} \langle \theta^*, x_{\ell,i_t} \rangle > \text{UCB}_t(\hat{r}_{t,i_t})$ **then**
    Eliminate arm i_t from the active set: $A_{t+1} \leftarrow A_t \setminus \{i_t\}$.
  if $A_{t+1} = \emptyset$ **then**
    Stop playing any arm and receive zero reward for all remaining rounds.

Similarly, algorithms for stochastic linear contextual bandits (LinUCB [1, 6]) and algorithms for linear contextual bandits with adversarial context corruptions (RobustBandit [8]) suffer linear regret when the arms act according to any Nash equilibrium that the algorithms incentivize.

Proof Sketch. We demonstrate that the only NE for the arms lies in strategies that myopically maximize the probability of being selected in every round, which results in linear regret for the learner, because all arms always appear similarly good. The proof can be found in Appendix B. □

Another natural question to ask is whether exact incentive-compatibility is possible in the strategic linear contextual bandit. A learning algorithm is called *incentive-compatible* if truthfulness is a NE, i.e., reporting the true context $x_{t,i} = x^*_{t,i}$ every round is maximizing each arm's utility [14, 30]. For the interested reader, in Appendix A, we provide an incentive-compatible algorithm with constant regret in the *fully deterministic case*, where θ^* is known a priori as well as the rewards of pulled arms directly observable. However, when θ^* is unknown and/or the reward observations are subject to noise, we conjecture that exact incentive-compatibility (i.e., truthfulness is an exact NE, not ε-NE) is irreconcilable with regret minimization (cf. Appendix A).

4 Warm-Up: θ^* is Known in Advance

There are a number of challenges in the strategic linear contextual bandit. The most significant one is the need to incentivize the arms to be (approximately) truthful while simultaneously minimizing regret by learning about θ^* and selecting the best arms, even when observing (potentially) manipulated contexts. Notably, Proposition 3.3 showed that if we fail to align the arms' incentives, minimizing regret becomes impossible. Therefore, in the strategic linear contextual bandit, we must combine mechanism design with online learning techniques.

The uncertainty about θ^* poses a serious difficulty when trying to design such incentive-aware learning algorithms. As we only observe $x_{t,i}$ and $r_{t,i} = \langle \theta^*, x^*_{t,i} \rangle + \eta_t$, but do not observe the true context $x^*_{t,i}$, accurate estimation of θ^* is extremely challenging (and arguably intractable). We go into more depth in Section 5 when we introduce the Optimistic Grim Trigger Mechanism. For now, we consider the special case when $\boldsymbol{\theta}^*$ is known to the learner in advance. This lets us highlight some of the challenges when connecting mechanism design with online learning in a less complex setting and introduce high-level ideas and concepts. When θ^* is known in advance, it can be instructive to consider what we refer to as the *reported (expected) reward* $\langle \theta^*, x_{t,i} \rangle$ instead of the *reported context vector* $x_{t,i}$ itself. Taking this perspective, when arm i reports a d-dimensional vector $x_{t,i}$, we simply think of arm i reporting a scalar reward $\langle \theta^*, x_{t,i} \rangle$. In what follows, it will prove useful to keep this abstraction in mind.[2]

4.1 The Greedy Grim Trigger Mechanism

One idea for a mechanism is to use a grim trigger. In repeated social dilemmas, the grim trigger strategy ensures cooperation among self-interested players by threatening with defection for all remaining rounds if the grim trigger condition is satisfied [28]. Typically, the grim trigger condition is defined so that it is immediately satisfied if a player defected at least once.

[2]We use the expressions 'reported reward' and 'expected reward' interchangeably to mean the reward we would expect to observe based on the context reported by the arm.

 https://doi.org/10.52202/079017-3703

In the strategic contextual bandit, from the perspective of the learner, an arm can be considered to 'cooperate' if it is reporting its context truthfully. In turn, an arm 'defects' when it is reporting a gamed context. However, when an arm is reporting some context $x_{t,i}$ we do not know whether this arm truthfully reported its context or not, because we do not have access to the true context $x^*_{t,i}$. For this reason, we instead compare the expected reward $\langle \theta^*, x_{t,i} \rangle$ and the true reward $\langle \theta^*, x^*_{t,i} \rangle$. While we also cannot observe $\langle \theta^*, x^*_{t,i} \rangle$ directly, we do observe $r_{t,i} \coloneqq \langle \theta^*, x^*_{t,i} \rangle + \eta_t$.

Grim Trigger Condition. Intuitively, if for any arm i the total expected reward $\sum_{\ell \leq t\colon i_\ell = i} \langle \theta^*, x_{\ell,i} \rangle$ is larger than the total observed reward $\hat{r}_{t,i} \coloneqq \sum_{\ell \leq t\colon i_\ell = i} r_{\ell,i}$, then arm i must have been misreporting its contexts. However, $r_{\ell,i} \coloneqq \langle \theta^*, x^*_{\ell,i} \rangle + \eta_\ell$ is random so that we instead use the *optimistic estimate* of the *observed reward* given by

$$\mathrm{UCB}_t(\hat{r}_{t,i}) \coloneqq \sum_{\ell \leq t\colon i_\ell = i} r_{\ell,i} + 2\sqrt{n_t(i) \log(T)} \tag{1}$$

where $2\sqrt{n_t(i)\log(T)}$ is the confidence width which can be derived from Hoeffding's inequality. To implement the grim trigger, we then eliminate arm i in round t if the *total expected reward* is larger than the *optimistic estimate of the total observed reward*, i.e.,

$$\sum\nolimits_{\ell \leq t\colon i_\ell = i} \langle \theta^*, x_{\ell,i} \rangle > \mathrm{UCB}_t(\hat{r}_{t,i}).$$

Note that using the optimistic estimate of the total observed reward ensures that elimination is justified with high probability. Conversely, we can guarantee with high probability that we do not erroneously eliminate a truthful arm.

Selection Rule. To complete the Greedy Grim Trigger Mechanism (GGTM, Mechanism 1), we then combine this with a greedy selection rule that pulls the arm with largest *reported reward* $\langle \theta^*, x_{t,i} \rangle$ in round t from the set of arms that we believe have been truthful so far. Interestingly, even though we here assumed θ^* to be known in advance, we see that GGTM still utilizes online learning techniques such as the optimistic estimate (1) to align the arms' incentives.

It is also worth noting that—similar to its use in repeated social dilemmas—our grim trigger mechanism is *mutually destructive* in the sense that eliminating an arm for all remaining rounds is inherently bad for the learner (and of course for the eliminated arm as well).[3] Here lies the main challenge of the mechanism design and we must ensure that the arms are incentivized to "cooperate" (i.e., remain active) for a sufficiently long time.

4.2 Regret Analysis of GGTM

In what follows, we assume that each arm's strategy is restricted to reporting their 'reward' $\langle \theta^*, x_{t,i} \rangle$ not strictly lower than their true (mean) reward $\langle \theta^*, x^*_{t,i} \rangle$. It seems intuitive that no rational arm would ever under-report its value to the learner and make itself seem worse than it actually is. However, there are special cases, where under-reporting allows an arm to arbitrarily manipulate without detection. We discuss this later in Remark 4.3 and, more extensively, in Appendix C.

Assumption 1. We assume that $\langle \theta^*, x_{t,i} \rangle \geq \langle \theta^*, x^*_{t,i} \rangle$ for all $(t, i) \in [T] \times [K]$.

We now demonstrate that GGTM approximately incentivizes the arms to be truthful in the sense that the truthful strategy profile $\boldsymbol{\sigma}^*$ such that $x_{t,i} = x^*_{t,i}$ for all $(t, i) \in [T] \times [K]$ is an $\tilde{\mathcal{O}}(\sqrt{T})$-NE under GGTM. When the arms always report truthfully and no arm is erroneously eliminated, the greedy selection rule naturally selects the best arm every round so that GGTM's regret is constant.

Theorem 4.1. *Under the Greedy Grim Trigger Mechanism, being truthful is a $\tilde{\mathcal{O}}(\sqrt{T})$-NE for the arms. The strategic regret of GGTM when the arms act according to this equilibrium is at most*

$$R_T(\mathrm{GGTM}, \boldsymbol{\sigma}^*) \leq 1/T.$$

Proof Sketch. By design of the grim trigger, it is straightforward to show that the probability that a truthful arm gets eliminated is at most $1/T^2$. Moreover, the grim trigger ensures that no arm can 'poach' selections from a truthful arm more than order $\sqrt{T}$ times by misreporting its contexts. This achieves two things: (a) it protects truthful arms and guarantees that truthfulness is a good strategy, and (b) limits an arm's profit from being untruthful. The proof can be found in Appendix D.2. □

[3] Note that in linear contextual bandits there is no single optimal arm, but the optimal arm changes per round.

Theorem 4.1 tells us that truthfulness is an approximate NE. We now also provide a more holistic strategic regret guarantee of $\tilde{\mathcal{O}}(K^2\sqrt{KT})$ in *every* Nash equilibrium under GGTM. Proving this is more complicated as the arms can profit from exploiting our uncertainty about their truthfulness (i.e., the looseness of the grim trigger).

Theorem 4.2. *The Greedy Grim Trigger Mechanism has strategic regret*

$$R_T(\text{GGTM}, \boldsymbol{\sigma}) = \mathcal{O}\left(\underbrace{\sqrt{KT\log(T)}}_{\textit{cost of manipulation}} + \underbrace{K^2\sqrt{KT\log(T)}}_{\textit{cost of mechanism design}}\right) \tag{2}$$

for every $\boldsymbol{\sigma} \in \text{NE}(\text{GGTM})$. *Hence,* $\max_{\boldsymbol{\sigma}\in\text{NE(GGTM)}} R_T(\text{GGTM}, \boldsymbol{\sigma}) = \tilde{\mathcal{O}}\big(K^2\sqrt{KT}\big)$.

Proof Sketch. The regret analysis is notably more complicated than the one in Theorem 4.1, as we must bound the regret due to the arms exploiting our uncertainty as well as the cost of committing to the grim trigger. Both of these quantities do not play a role when the arms always report truthfully (like in Theorem 4.1). A complete proof can be found in Appendix D. □

The regret bound (2) suggests that there are two sources of regret. The first term is due to our mechanism design being approximate (relying on estimates), which leaves room for the arms to exploit our uncertainty and misreport their contexts to obtain additional selections. The second part of (2) is the cost of the mechanism design, i.e., the cost of committing to the grim trigger. We suffer constant regret any round in which the round-optimal arm is no longer in the active set. In the worst-case, this quantity is of order $K^2\sqrt{KT}$.

Remark 4.3. *We want to briefly comment on Assumption 1. It appears intuitive that any rational arm would never under-report its value, i.e., make itself look worse than it actually is. However, in Appendix C, we provide a simple example where occasionally under-reporting its value allows an arm to simulate an environment where it is always optimal, even though it is in fact only optimal half of the time. We will explain in the example that without additional strong assumptions on the noise distribution the two environments are indistinguishable so that such manipulation by the arms appears unavoidable when trying to maximize rewards.*

5 The Optimistic Grim Trigger Mechanism

The problem of estimating the unknown parameter θ^* appears daunting given that the arms can strategically alter their contexts to manipulate our estimate of θ^* to their advantage. In fact, imagine an arm manipulating its contexts orthogonal to θ^* so that $\langle \theta^*, x_{t,i} - x^*_{t,i}\rangle = 0$ but $x_{t,i} \neq x^*_{t,i}$. Observing only $x_{t,i}$ and $r_{t,i} \coloneqq \langle \theta^*, x^*_{t,i}\rangle + \eta_t$, our estimate of θ^* becomes biased and could be arbitrarily far off the true parameter θ^* even though the gamed context and true context have the same reward w.r.t. θ^*. This is also the case more generally. Since we observe neither θ^* nor $x^*_{t,i}$, any observed combination of $x_{t,i}$ and $r_{t,i}$ will "make sense" to us. *But, how can we incentivize the arms to report truthfully and minimize regret despite incorrect estimates of* θ^**?*

Our key observation is that learning θ^* is not necessary to incentivize the arms or minimize regret; it appears to be a hopeless endeavour after all. The idea of the Optimistic Grim Trigger Mechanism (OptGTM, Mechanism 2) is to construct pessimistic estimates of the total reward we expected from pulling an arm. Importantly, we can construct such pessimistic estimates of the expected (i.e., "reported") reward even when the contexts are manipulated. OptGTM then threatens arms with elimination if our *pessimistic* estimate of the expected reward exceeds the *optimistic* estimate of the observed reward. Interestingly, this does not relate to the amount of corruption in the feature space and, in fact, $\sum_{t,i}\|x_{t,i} - x^*_{t,i}\|_2$ could become arbitrarily large. However, it does bound the effect of each arm's strategic manipulation on the decisions we make and thereby allows for effective incentive design and regret minimization.

To construct pessimistic (and optimistic) estimates of the expected reward, we use independent estimators $\hat{\theta}_{t,i}$ and confidence sets $C_{t,i}$ around $\hat{\theta}_{t,i}$, which do not take into account that the contexts are potentially manipulated. That is, we have a separate estimator and confidence set for each arm $i \in [K]$. This will prevent one arm influencing the elimination of another. It also stops collusive arm behavior, where a majority group of the arms could dominate and steer our estimation process.

https://doi.org/10.52202/079017-3703

Mechanism 2: The Optimistic Grim Trigger Mechanism (OptGTM)

initialize: $A_1 = [K]$
for $t = 1, \dots, T$ **do**
  Observe reported contexts $\mathcal{X}_t = \{x_{t,1}, \dots, x_{t,K}\}$.
  Play the active arm with largest reported optimistic reward

$$i_t = \operatorname*{argmax}_{i \in A_t} \mathrm{UCB}_{t,i}\big(x_{t,i}\big).$$

  Receive reward r_{t,i_t} from playing arm i_t.
  if $\sum_{\ell \le t\colon i_\ell = i_t} \mathrm{LCB}_{\ell,i_t}(x_{\ell,i_t}) > \mathrm{UCB}_t(\hat{r}_{t,i_t})$ **then**
    Eliminate arm i_t from the active set: $A_{t+1} \leftarrow A_t \setminus \{i_t\}$.
  if $A_{t+1} = \emptyset$ **then**
    Stop playing any arm and receive zero reward for all remaining rounds.

Confidence Sets. For every arm $i \in [K]$ we define the least-squares estimator $\hat{\theta}_{t,i}$ w.r.t. its *reported* contexts and observed rewards before round t as

$$\hat{\theta}_{t,i} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left(\sum\nolimits_{\ell < t\colon i_\ell = i} \big(\langle \theta, x_{\ell,i} \rangle - r_{\ell,i}\big)^2 + \lambda \|\theta\|_2^2 \right), \tag{3}$$

where $\lambda > 0$. We then define the confidence set $C_{t,i} \coloneqq \{\theta \in \mathbb{R}^d\colon \|\hat{\theta}_{t,i} - \theta\|_{V_{t,i}}^2 \le \beta_{t,i}\}$ where $\beta_{t,i} \coloneqq \mathcal{O}(d \log(n_t(i)))$ is the confidence size. Here, $V_{t,i}$ is the covariance matrix of reported contexts of arm i given by $V_{1,i} \coloneqq \lambda I$ and $V_{t,i} \coloneqq \lambda I + \sum_{\ell < t\colon i_\ell = i} x_{\ell,i} x_{\ell,i}^\top$.[4]

It is well-known that if the contexts were always reported truthfully, i.e., $x_{t,i} = x_{t,i}^*$, then with high probability $\theta^* \in C_{t,i}$. But, if the sequence of reported contexts $x_{t,i}$ substantially differs from the true contexts $x_{t,i}^*$ there is no (high probability) guarantee that the true parameter θ^* is element in $C_{t,i}$. In the literature on learning with adversarial corruptions (in linear contextual bandits), the standard approach to deal with this is to widen the confidence set. However, for this to be effective we would need to assume a sufficiently small corruption budget for the arms and prior knowledge of the total amount of corruption, both of which we explicitly do not assume here.

Slightly overloading notation, we instead define the optimistic and pessimistic estimate of the expected reward of a context vector x w.r.t. arm i as

$$\mathrm{UCB}_{t,i}(x) \coloneqq \langle \hat{\theta}_{t,i}, x \rangle + \sqrt{\beta_{t,i}} \|x\|_{V_{t,i}^{-1}} \quad \text{and} \quad \mathrm{LCB}_{t,i}(x) \coloneqq \langle \hat{\theta}_{t,i}, x \rangle - \sqrt{\beta_{t,i}} \|x\|_{V_{t,i}^{-1}}.$$

We chose to state these estimates using additive bonuses. However, it can be convenient to think of them as $\mathrm{UCB}_{t,i}(x) \approx \max_{\theta \in C_{t,i}} \langle \theta, x \rangle$ and $\mathrm{LCB}_{t,i}(x) \approx \min_{\theta \in C_{t,i}} \langle \theta, x \rangle$.

Grim Trigger Condition. In round $t \in [T]$, we eliminate arm i from A_t if the pessimistic estimate using the reports is larger than the optimistic estimate using the total observed reward, i.e.,

$$\sum_{\ell \le t\colon i_\ell = i} \left(\langle \hat{\theta}_{\ell,i}, x_{\ell,i} \rangle - \sqrt{\beta_\ell} \|x_{\ell,i}\|_{V_{\ell,i}^{-1}} \right) > \sum_{\ell \le t\colon i_\ell = i} r_{\ell,i} + 2\sqrt{n_t(i) \log(T)}. \tag{4}$$

In other words, $\sum_{\ell \le t\colon i_\ell = i} \mathrm{LCB}_{\ell,i}(x_{\ell,i}) > \mathrm{UCB}_t(\hat{r}_{t,i})$.

Examining the left side of (4), the careful reader may wonder why we do not simply use our latest and arguably best estimate $\hat{\theta}_{t,i}$, but instead the whole sequence of "out-dated" estimators $\hat{\theta}_{\ell,i}$ from previous rounds. In fact, this is crucial for the grim trigger. Using $\hat{\theta}_{t,i}$ renders the grim trigger condition ineffective, because, by definition, $\hat{\theta}_{t,i}$ is the (least-squares) minimizer (3) of the difference between $\sum_{\ell \le t\colon i_\ell = i} \langle \hat{\theta}_{t,i}, x_{\ell,i} \rangle$ and $\sum_{\ell \le t\colon i_\ell = i} r_{\ell,i}$. Hence, when using $\hat{\theta}_{t,i}$ the grim trigger condition may not be satisfied even when the arms significantly and repeatedly misreport their contexts.

Selection Rule. We complete the OptGTM algorithm by selecting arms optimistically with respect to each arm's own estimator and confidence set. That is, OptGTM selects the active arm with maximal optimistic (expected) reward $\mathrm{UCB}_{t,i}(x_{t,i}) \coloneqq \langle \hat{\theta}_{t,i}, x_{t,i} \rangle + \sqrt{\beta_{t,i}} \|x_{t,i}\|_{V_{t,i}^{-1}}$ in round t. We see that

[4] For more details on the design of least-squares estimators and the confidence sets, we refer to Abbasi-Yadkori et al. [1] and Lattimore and Szepesvári [24] (Chapter 20).

the grim trigger (4) incentivizes arms to ensure that over the course of all rounds $\mathrm{LCB}_{t,i}(x_{t,i})$ is not much smaller than $r_{t,i} \coloneqq \langle \theta^*, x^*_{t,i} \rangle + \eta_t$. Hence, $\mathrm{UCB}_{t,i}(x_{t,i})$ is also not substantially smaller than the true mean reward $\langle \theta^*, x^*_{t,i} \rangle$. This suggests that playing optimistically is a good strategy for the learner as long as the selected arm does not satisfy (4).

5.1 Regret Analysis of OptGTM

When θ^* was known to the learner in advance, we assumed that the arms never report a value smaller than their true value, i.e., $\langle \theta^*, x_{t,i} \rangle \geq \langle \theta^*, x^*_{t,i} \rangle$ for all $(t, i) \in [T] \times [K]$. Now, when θ^* is unknown to the learner, we similarly assume that the arms do not report their optimistic value less than their true value. Again, it seems intuitive that in any given round, no arm would under-report its worth.

Assumption 2. We assume that $\max_{\theta \in C_{t,i}} \langle \theta, x_{t,i} \rangle \geq \langle \theta^*, x^*_{t,i} \rangle$ for all $(t, i) \in [T] \times [K]$.

We find that OptGTM approximately incentivizes the arms to be truthful and, when the arms report truthfully, OptGTM suffers at most $\tilde{\mathcal{O}}(d\sqrt{KT})$ regret.

Theorem 5.1. *Under the Optimistic Grim Trigger Mechanism, being truthful is a $\tilde{\mathcal{O}}(d\sqrt{KT})$-NE. When the arms report truthfully, the strategic regret of OptGTM under this approximate NE is at most*

$$R_T(\text{OptGTM}, \boldsymbol{\sigma}^*) = \tilde{\mathcal{O}}(d\sqrt{KT}).$$

The optimal regret in standard non-strategic linear contextual bandits is $\Theta(d\sqrt{T})$ so that OptGTM is optimal up to a factor of $\sqrt{K}$ (and logarithmic factors) when the arms report truthfully. The additional factor of $\sqrt{K}$ is caused by the fact that OptGTM maintains independent estimates for each arm. We now also provide a strategic regret bound for every NE of the arms under OptGTM.

Theorem 5.2. *The Optimistic Grim Trigger Mechanism has strategic regret*

$$R_T(\text{OptGTM}, \boldsymbol{\sigma}) = \mathcal{O}\left(d \log(T)\sqrt{KT} + d \log(T) K^2 \sqrt{KT}\right).$$

for every $\boldsymbol{\sigma} \in \mathrm{NE}(\text{OptGTM})$. Hence, $\max_{\boldsymbol{\sigma} \in \mathrm{NE}(\text{OptGTM})} R_T(\text{OptGTM}, \boldsymbol{\sigma}) = \tilde{\mathcal{O}}\left(dK^2\sqrt{KT}\right)$.

The proof ideas of Theorem 5.1 and Theorem 5.2 are similar to their counterparts in Section 4. The main difference lies in a more technically challenging analysis of the grim trigger condition (4). We also see that in contrast to non-strategic linear contextual bandits, where the regret typically does not depend on the number of arms K, Theorem 5.1 and Theorem 5.2 include a factor of $\sqrt{K}$ and $K^2\sqrt{K}$, respectively. A dependence on K is expected due to the strategic nature of the arms which forces us to explicitly incentivize each arm to be truthful. However, we conjecture that the regret bound in Theorem 5.2 is not tight in K and expect the optimal dependence on the number of arms to be $\sqrt{K}$. The proofs of Theorem 5.1 and Theorem 5.2 can be found in Appendix E.

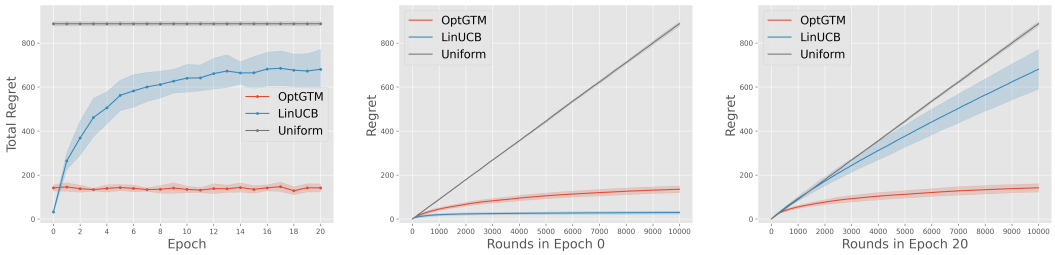
6 Experiments: Simulating Strategic Context Manipulation

We here experimentally analyze the efficacy of OptGTM when the arms strategically manipulate their contexts in response to our learning algorithm. We compare the performance of OptGTM with the traditional LinUCB algorithm [1, 6], which—as shown in Proposition 3.3—implicitly incentivizes the arms to manipulate their contexts and suffers large regret when the arms are strategic.

Contrary to the assumption of arms playing NE, we here model strategic arm behavior by letting the arms update their strategy (i.e., what contexts to report) based on past interactions with the algorithms. More precisely, we assume that the strategic arms interact with the deployed algorithm (i.e., OptGTM or LinUCB) over the course of 20 epochs, with each epoch consisting of $T = 10k$ rounds. At the end of each epoch, every arm then updates its strategy using gradient ascent w.r.t. its utility. Importantly, this approach requires no prior knowledge from the arms, as they learn entirely through sequential interaction. This does not necessarily lead to equilibrium strategies, but serves as a natural model of strategic gaming behavior under which to study the algorithms.

Experimental Setup. We associate each arm with a true feature vector $y^*_i \in \mathbb{R}^{d_1}$ (e.g., product features) and randomly sample a sequence of user vectors $c_t \in \mathbb{R}^{d_2}$ (i.e., customer features). We assume that every arm can alter its feature vector y^*_i by reporting some other vector y_i, but cannot alter the user contexts c_t. We use a feature mapping $\varphi(c_t, y_i) = x_{t,i}$ to map $y_i \in \mathbb{R}^{d_1}$ and $c_t \in \mathbb{R}^{d_2}$ to an arm-specific context $x_{t,i} \in \mathbb{R}^d$ that the algorithm observes. At the end of every epoch, each arm then performs an approximated gradient step on y_i w.r.t. its utility, i.e., the number of times it is selected. We let $K = 5$ and $d = d_1 = d_2 = 5$ and average the results over 10 runs.

 https://doi.org/10.52202/079017-3703



(a) Total strategic regret R_T as the arms adapt their strategies to the deployed algorithm over the course of 20 epochs.

(b) Epoch 0 (Truthful Arms): Regret as a function of t *before* the arms have interacted with the deployed algorithm.

(c) Epoch 20 (Strategic Arms): Regret as a function of t *after* the arms have interacted with the deployed algorithm.

Figure 1: Comparison of the strategic regret of OptGTM and LinUCB. The strategic arms adapt their strategies gradually over the course of 20 epochs. OptGTM performs similarly across all epochs, whereas LinUCB performs increasingly worse as the arms adapt to the algorithm (Figure 1a). Figure 1b and 1c provide a closer look at the regret of the algorithms across the T rounds in the initial epoch, where the arms are truthful, and the final epoch after the arms have adapted to the algorithms.

Results. In Figure 1a, we observe that OptGTM performs similarly well across all epochs, which suggests that OptGTM successfully discourages the emergence of harmful gaming behavior. In contrast, as the arms adapt their strategies (i.e., what features to report), LinUCB suffers increasingly more regret and almost performs as badly as uniform sampling in the final epoch. In epoch 0, when the all arms are truthful, i.e., are non-strategic, LinUCB performs better than OptGTM (Figure 1b). This is expected as OptGTM suffers additional regret due to maintaining independent estimates of θ^* for each arm (as a mechanism to incentivize truthfulness). However, OptGTM significantly outperforms LinUCB as the arms strategically adapt, which is most evident in the final epoch (Figure 1c). Interestingly, as already suggested in Section 5, OptGTM cannot prevent manipulation in the feature space (see Figure 2). However, OptGTM does manage to bound the effect of the manipulation on the regret (Figure 1a) and, most importantly, the effect on the utility of the arms as well (Figure 3). As a result, the arms are discouraged from heavily gaming their contexts and the context manipulation has only a minor effect on the actions taken by OptGTM.

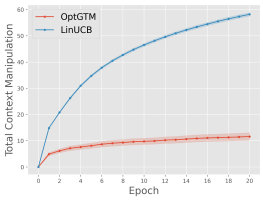


Figure 2: Context manipulation $\sum_{t,i} \|x^*_{t,i} - x_{t,i}\|_2$.

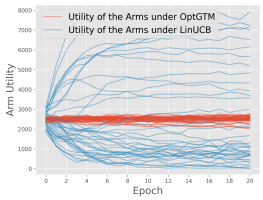


Figure 3: Utility of the arms for each of the 10 runs.

7 Discussion

We study a strategic variant of the linear contextual bandit problem, where the arms strategically misreport privately observed contexts to maximize their selection frequency. To address this, we design two online learning mechanisms: the Greedy and the Optimistic Grim Trigger Mechanism, for the scenario where θ^* is known and unknown, respectively. We demonstrate that our mechanisms incentivize the arms to be approximately truthful and, in doing so, effectively minimize regret. More generally, with this work, we aim to advance our understanding of problems at the intersection of online learning and mechanism design. As the digital landscape, including online platforms and marketplaces, becomes increasingly agentic and dominated by self-interested agents, it will be crucial to understand the incentives created by learning algorithms and to align these incentives while optimizing for performance.

Limitations. One limitation is the otherwise intuitive assumption that the arms do not under-report their value to the learner (Assumption 1 and Assumption 2). Secondly, we believe that the factor of K^2 in the universal regret guarantees of Theorem 4.2 and Theorem 5.2 is suboptimal and we conjecture that the optimal worst-case strategic regret is given by $\mathcal{O}(d\sqrt{KT})$. We leave this investigation for future work.

Acknowledgements

Thomas Kleine Buening is supported by the UKRI Prosperity Partnership Scheme (Project FAIR). Haifeng Xu is supported in part by the Army Research Office Award W911NF-23-1-0030, the ONR Award N00014-23-1-2802 and the NSF Award CCF-2303372.

References

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

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

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

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

[5] Thomas Kleine Buening, Aadirupa Saha, Christos Dimitrakakis, and Haifeng Xu. Bandits meet mechanism design to combat clickbait in online recommendation. In *The Twelfth International Conference on Learning Representations*, 2023.

[6] 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*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[7] Sarah Dean, Evan Dong, Meena Jagadeesan, and Liu Leqi. Accounting for ai and users shaping one another: The role of mathematical models. *arXiv preprint arXiv:2404.12366*, 2024.

[8] Qin Ding, Cho-Jui Hsieh, and James Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 7111–7123. PMLR, 2022.

[9] Jing Dong, Ke Li, Shuai Li, and Baoxiang Wang. Combinatorial bandits under strategic manipulations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 219–229, 2022.

[10] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

[11] Seyed Esmaeili, MohammadTaghi Hajiaghayi, and Suho Shin. Replication-proof bandit mechanism design. *arXiv preprint arXiv:2312.16896*, 2023.

[12] Seyed A Esmaeili, Suho Shin, and Aleksandrs Slivkins. Robust and performance incentivizing algorithms for multi-armed bandits with strategic agents. *arXiv preprint arXiv:2312.07929*, 2023.

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

[14] Rupert Freeman, David Pennock, Chara Podimata, and Jennifer Wortman Vaughan. No-regret and incentive-compatible online learning. In *International Conference on Machine Learning*, pages 3270–3279. PMLR, 2020.

[15] Nicolas Gast, Stratis Ioannidis, Patrick Loiseau, and Benjamin Roussillon. Linear regression from strategic data sources. *ACM Transactions on Economics and Computation (TEAC)*, 8(2): 1–24, 2020.

 https://doi.org/10.52202/079017-3703

[16] Arpita Ghosh and Patrick Hummel. Learning and incentives in user-generated content: Multi-armed bandits with endogenous arms. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 233–246, 2013.

[17] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

[18] Keegan Harris, Dung Daniel T Ngo, Logan Stapleton, Hoda Heidari, and Steven Wu. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *International Conference on Machine Learning*, pages 8502–8522. PMLR, 2022.

[19] Keegan Harris, Chara Podimata, and Steven Z Wu. Strategic apple tasting. *Advances in Neural Information Processing Systems*, 36:79918–79945, 2023.

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

[21] Jiri Hron, Karl Krauth, Michael I Jordan, Niki Kilbertus, and Sarah Dean. Modeling content creator incentives on algorithm-curated platforms. *arXiv preprint arXiv:2206.13102*, 2022.

[22] Xinyan Hu, Meena Jagadeesan, Michael I Jordan, and Jacob Steinhard. Incentivizing high-quality content in online recommender systems. *arXiv preprint arXiv:2306.07479*, 2023.

[23] Meena Jagadeesan, Nikhil Garg, and Jacob Steinhardt. Supply-side equilibria in recommender systems. *Advances in Neural Information Processing Systems*, 36, 2024.

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

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

[26] Yang Liu and Yiling Chen. A bandit framework for strategic regression. *Advances in Neural Information Processing Systems*, 29, 2016.

[27] Yang Liu and Chien-Ju Ho. Incentivizing high quality user contributions: New arm generation in bandit learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[28] Michael W Macy and Andreas Flache. Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences*, 99(suppl_3):7229–7236, 2002.

[29] Ross A Malaga. Worst practices in search engine optimization. *Communications of the ACM*, 51(12):147–150, 2008.

[30] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582, 2015.

[31] Nicolò Pagan, Joachim Baumann, Ezzat Elokda, Giulia De Pasquale, Saverio Bolognani, and Anikó Hannák. A classification of feedback loops and their relation to biases in automated decision-making systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–14, 2023.

[32] Shankar Prawesh and Balaji Padmanabhan. The "most popular news" recommender: Count amplification and manipulation resistance. *Information Systems Research*, 25(3):569–589, 2014.

[33] Paul Resnick and Rahul Sami. The influence limiter: provably manipulation-resistant recommender systems. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 25–32, 2007.

[34] Elan Rosenfeld and Nir Rosenfeld. One-shot strategic classification under unknown costs. *arXiv preprint arXiv:2311.02761*, 2023.

[35] Suho Shin, Seungjoon Lee, and Jungseul Ok. Multi-armed bandit algorithm against strategic replication. In *International Conference on Artificial Intelligence and Statistics*, pages 403–431. PMLR, 2022.

[36] Ravi Sundaram, Anil Vullikanti, Haifeng Xu, and Fan Yao. Pac-learning for strategic classification. *Journal of Machine Learning Research*, 24(192):1–38, 2023.

[37] Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. *Mobile health: sensors, analytic methods, and applications*, pages 495–517, 2017.

[38] Zongwei Wang, Min Gao, Junliang Yu, Hao Ma, Hongzhi Yin, and Shazia Sadiq. Poisoning attacks against recommender systems: A survey. *arXiv preprint arXiv:2401.01527*, 2024.

[39] Chen-Yu Wei, Christoph Dann, and Julian Zimmert. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pages 1043–1096. PMLR, 2022.

[40] Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. How bad is top-k recommendation under competing content creators? In *International Conference on Machine Learning*, pages 39674–39701. PMLR, 2023.

[41] Fan Yao, Chuanhao Li, Karthik Abinav Sankararaman, Yiming Liao, Yan Zhu, Qifan Wang, Hongning Wang, and Haifeng Xu. Rethinking incentives in recommender systems: Are monotone rewards always beneficial? *Advances in Neural Information Processing Systems*, 36, 2024.

[42] Fan Yao, Yiming Liao, Mingzhe Wu, Chuanhao Li, Yan Zhu, James Yang, Qifan Wang, Haifeng Xu, and Hongning Wang. User welfare optimization in recommender systems with competing content creators. *arXiv preprint arXiv:2404.18319*, 2024.

[43] Mengxin Yu, Zhuoran Yang, and Jianqing Fan. Strategic decision-making in the presence of information asymmetry: Provably efficient rl with algorithmic instruments. *arXiv preprint arXiv:2208.11040*, 2022.

[44] Hanrui Zhang and Vincent Conitzer. Incentive-aware pac learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5797–5804, 2021.

[45] Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2110.12615*, 2021.

 https://doi.org/10.52202/079017-3703

A Remarks on Incentive-Compatible No-Regret Algorithms

In Section 3, we conjectured that there exists no incentive-compatible no-regret algorithm when the reward observations are subject to noise and θ^* unknown. For the interested reader, we here consider the *fully deterministic case* where θ^* is known a priori and reward observations are directly observable, i.e., subject to no noise so that $\eta_t \equiv 0$. We can design the following provably incentive-compatible no-regret algorithm. In fact, we show that this mechanism is strategyproof, i.e., incentive-compatible in weakly dominant strategies.

Mechanism 3: Incentive-Compatible No-Regret Algorithm in the Fully Deterministic Case

initialize: $A_1 = [K]$
for $t < T - (K+1)$ **do**
  Play $i_t \in \operatorname{argmax}_{i \in A_t} \langle \theta^*, x_{t,i} \rangle$
  Observe reward $r^*_{t,i_t} := \langle \theta^*, x^*_{t,i_t} \rangle$ (i.e., rewards of chosen arms are directly observable)
  if $\langle \theta^*, x_{t,i_t} \rangle \neq r^*_{t,i_t}$ **then**
    Eliminate arm i_t: $A_{t+1} \leftarrow A_t \setminus \{i_t\}$.
for $t \geq T - (K+1)$ **do**
  Play $i_t \sim \text{Uniform}(A_t)$
  Observe reward $r^*_{t,i_t} := \langle \theta^*, x^*_{t,i_t} \rangle$
  if $\langle \theta^*, x_{t,i_t} \rangle \neq r^*_{t,i_t}$ **then**
    Eliminate arm i_t: $A_{t+1} \leftarrow A_t \setminus \{i_t\}$.

Lemma A.1. *Mechanism 3 is strategyproof, i.e., being truthful is a weakly dominant strategy for every arm. Moreover, Mechanism 3 suffers at most $K+1$ strategic regret in every Nash equilibrium of the arms.*[5]

Proof. **Incentive-Compatibility in Weakly Dominant Strategies.** It is easy to see that for the last $K+1$ rounds, reporting truthfully, i.e., reporting $x^*_{t,i}$, is a weakly dominant strategy, since the the set of active arms is played uniformly and nothing can be gained from misreporting (an arm can only miss out on being selected by misreporting in the last $K+1$ steps). Hence, conditional on any history, reporting truthfully is the best continuation for any arm. In particular, when an arm plays truthfully in these rounds the obtained utility in the last $K+1$ steps is at least $\frac{K+1}{K}$, since $|A_t| \leq K$.

Now, for the time steps $t < T - (K+1)$ note that any untruthful strategy can obtain at most one more allocation than the truthful strategy, because if $i_t = i$ and $\langle \theta^*, x_{t,i} \rangle > \langle \theta^*, x^*_{t,i} \rangle$, then arm i is eliminated immediately. Hence, at most utility 1 can be gained from receiving an allocation by misreporting. However, in this case the arm gets eliminated and receives utility 0 in the last $K+1$ rounds. As seen before the minimum utility the truthful strategy receives in the last $K+1$ receives is $\frac{K+1}{K} > 1$. Consequently, irrespective of the other arms strategies, the truthful strategy is (weakly) optimal for arm i.

One may wonder why the truthful strategy is not *strictly* dominant. To see this note that reporting any $x_{t,i} \neq x^*_{t,i}$ such that the difference $x_{t,i} - x^*_{t,i}$ is orthogonal to θ^*, i.e., $\langle \theta^*, x_{t,i} - x^*_{t,i} \rangle = 0$, does not cause elimination and is equivalent under Mechanism 3. In other words, such untruthful strategies, which however have no effect on the selection, are equally good.

Regret. The regret in the last $K+1$ rounds is trivially bounded by $K+1$. When showing that the algorithm is strategyproof we showed that any untruthful strategy such that there exists $i_t = i$ with $\langle \theta^*, x_{t,i} \rangle > \langle \theta^*, x^*_{t,i} \rangle$ is worse than the truthful strategy independently from what the other arms are playing. Hence, in any Nash equilibrium arm i chooses strategies such that if $i_t = i$ then $\langle \theta^*, x_{t,i} \rangle = \langle \theta^*, x^*_{t,i} \rangle$. In other words, since the selection is greedy, Mechanism 3 selected the best arm in the given round. Mechanism 3 therefore suffers zero regret in the first $T - (K+1)$ regret in any Nash equilibrium of the arms. □

As discussed in Section 3, we conjecture that there exists no incentive-compatible no-regret algorithm for the strategic linear contextual bandits when the reward observations are subject to noise. The

[5] Note that since truthfulness is only weakly dominant there could be other Nash equilibria.

intuition for this conjecture is as follows. Suppose there exists a learning algorithm M that is incentive-compatible and no-regret, that is, the strategy profile where every arm is *always* truthful is a NE. Since M is also no-regret, the selection policy of M must depend on the reported contexts in some way. In particular, in some round t in which M does not select arm i—but M maps from reported contexts to an action in $[K]$—there must exist a context $\tilde{x}_t$ that arm i could report that increases its probability of being selected.

Suppose arm i changes its strategy from σ_i^* (i.e., being truthful) to the strategy that is always truthful except for round t where it reports $\tilde{x}_t$ instead of $x_{t,i}^*$. The algorithm M then observes a reward drawn from a distribution with mean $\langle \theta^*, x_{t,i}^* \rangle$, but might have expected a reward drawn from a distribution with mean $\langle \theta^*, \tilde{x}_t \rangle$. We believe that the difference in observed and expected reward is statistically insignificant when arm i only misreports a single or constant number of times. However, due to the intricate relationship between the learning algorithm and the induced NE strategies for the K arms, providing a rigorous argument for this is challenging.

B Proof of Proposition 3.3

Proof of Proposition 3.3. We begin with the incentive-unaware greedy algorithm that in round t pulls arm $i_t = \operatorname{argmax}_{i \in [K]} \langle \theta^*, x_{t,i} \rangle$. Let $\tilde{x} := \operatorname{argmax}_{\|x\| \leq 1} \langle \theta^*, x \rangle$ and w.l.o.g. we assume that $\tilde{x}$ is unique. We show that the strategy profile, where every arm always reports $\tilde{x}$ is the only Nash equilibrium under the incentive-unaware greedy algorithm. Let $\boldsymbol{\sigma}$ be any strategy profile which is such that there exists a round t and arm i such that $x_{t,i} \neq \tilde{x}$. We distinguish between two cases.

Case 1: There exists a round t and arm i such that $\langle \theta^*, x_{t,i} \rangle < \max_{j \in [K]} \langle \theta^*, x_{t,j} \rangle$.

Note that this implies that arm i is not selected by the learner. However, by reporting $\tilde{x}$ instead of $x_{t,i}$, arm i is guaranteed to be selected with probability at least $1/K$. Hence, reporting $x_{t,i}$ is strictly worse than reporting $\tilde{x}$ so that $\boldsymbol{\sigma}$ cannot be a NE.

Case 2: $\langle \theta^*, x_{t,i} \rangle = \max_{j \in [K]} \langle \theta^*, x_{t,j} \rangle$ for all rounds t and arms i.

Note that this implies that each arm i is selected with probability $1/K$ every round.[6] Suppose that for any of these rounds t we have $\max_{j \in [K]} \langle \theta^*, x_{t,j} \rangle < \langle \theta^*, \tilde{x} \rangle$. Then, by reporting $\tilde{x}$ instead of $x_{t,i}$ arm i could ensure to be selected with probability one. Hence, the strategy where arm i in round t reports $\tilde{x}$ instead of $x_{t,i}$ is a strictly better response. Therefore, $\boldsymbol{\sigma}$ cannot be a NE. The other case is when $\max_{j \in [K]} \langle \theta^*, x_{t,j} \rangle = \langle \theta^*, \tilde{x} \rangle$, but this cannot be because $\boldsymbol{\sigma}$ is supposed to be different to always reporting $\tilde{x}$.

Consequently, the strategy profile where every arm always reports $\tilde{x}$ is the only NE under the incentive-unaware greedy algorithm. Under this strategy profile, the incentive-unaware greedy algorithm will play uniformly and therefore suffer linear regret.

Insufficiency of Non-Strategic Linear Contextual Bandit Algorithms. It is not really surprising that algorithms for non-strategic linear contextual bandits fail in the strategic linear contextual bandit, since such algorithms implicitly incentivize the arms to "compete" in every round by misreporting their context as the best possible one. Nothing prevents the arms to not myopically optimize their probability of being selected every round. As an example of a standard algorithm for non-strategic linear contextual bandits we consider LinUCB that in the non-strategic problem setup enjoys a regret guarantee of $\tilde{\mathcal{O}}(d\sqrt{T})$.

The reasons for LinUCB's failure in this strategic problem are the same as for the incentive-unaware greedy algorithm from before. It will be the strictly dominant strategy for the arms to maximize their selection probability in the given round by misreporting their context. Recall that LinUCB maintains a least-squares estimator given by

$$\hat{\theta}_t = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{\ell=1}^{t-1} (\langle \theta, x_{\ell,i_\ell} \rangle - r_{\ell,i_\ell})^2 + \lambda \|\theta\|_2^2$$

[6] This already implies linear regret, but it will be instructive to still show that the only NE is in maximally gaming strategies.

 https://doi.org/10.52202/079017-3703

and in round t selects arm (ties broken uniformly at random)

$$i_t = \operatorname*{argmax}_{i \in [K]} \langle \hat{\theta}_t, x_{t,i} \rangle + \sqrt{\beta_t} \|x_{t,i}\|_{V_t^{-1}},$$

where $\beta_t \approx d \log(T)$ and $V_t = \lambda I + \sum_{i=1}^{t-1} x_{\ell,i_\ell} x_{\ell,i}^\top$. Let $\mathrm{UCB}_t(x) \coloneqq \langle \hat{\theta}_t, x \rangle + \sqrt{\beta_t} \|x\|_{V_t^{-1}}$.

The argument for LinUCB will be the same as for the incentive-unaware greedy algorithm. Let $\tilde{x}_t \coloneqq \operatorname{argmax}_{\|x\|_2 \leq 1} \mathrm{UCB}_t(x)$ and w.l.o.g. assume that $\tilde{x}_t$ is unique.

Importantly, in what follows, keep in mind that it will not matter how the reports of arm i influenced $\tilde{\theta}_t$ or V_t in previous rounds. Once again, suppose $\boldsymbol{\sigma}$ is a strategy profile such that there exists a round t and arm i such that conditioned on the history $x_{t,i} \neq \tilde{x}_t$. Once again we distinguish between the following two cases:

Case 1: There exists a round t and arm i such that $\mathrm{UCB}_t(x_{t,i}) < \max_{j \in [K]} \mathrm{UCB}_t(x_{t,j})$.

This implies that arm i was not selected by the learner in round t. However, by reporting $\tilde{x}_t$ and in all future rounds report $\tilde{x}_\ell$ for $\ell > t$, arm i can guarantee to be selected with probability at least $1/K$ in round t and at least as many selections as under σ_i. Hence, σ_i cannot be a best response to σ_{-i}.

Case 2: $\mathrm{UCB}_t(x_{t,i}) = \max_{j \in [K]} \mathrm{UCB}_t(x_{t,j})$ for all rounds t and arms i.

Note that this implies that arm i is selected with probability $1/K$ every round. Suppose that for any round t it is the case that $\max_{j \in [K]} \mathrm{UCB}_t(x_{t,j}) < \mathrm{UCB}_t(\tilde{x})$. Then, by choosing strategy $\tilde{x}_t$ in round t and $\tilde{x}_\ell$ adaptively for all future rounds $\ell > t$, arm i obtains more selections than when reporting $x_{t,i}$. Hence, $\boldsymbol{\sigma}$ cannot be a NE.

As a result, the strategy profile where all arms report $\tilde{x}_t$ in round t is the only NE and LinUCB suffers linear regret, as it pulls arms uniformly at random. In exactly the same way, we can also show that the algorithms for linear contextual bandits with adversarial context corruptions in [8] suffer linear regret. □

C Assumption 1 and Remark 4.3

Example 1. We here give a simple example where a strategic arm can simulate a situation where it is always optimal even though it is only optimal half of the time.

Let $\theta^* = 1$ and consider the following problem instance with two arms 1 and 2, where

$$x_{t,1}^* = \begin{cases} 0, & t \text{ is even} \\ 1, & t \text{ is odd} \end{cases} \qquad \text{and} \qquad x_{t,2}^* = 1/4.$$

Now, suppose that arm 1 always reports $x_{t,1} = 1/2$ and arm 2 reports truthfully (or approximately so). Then, arm 1 appears optimal every round t. In particular, on average arm when we pull arm 1 it has reward $1/2$, which is consistent with its report of $x_{t,1} = 1/2$.

Now, consider a second problem instance, where

$$x_{t,1}^* = 1/2 \quad \text{and} \quad x_{t,2}^* = 1/4.$$

Recall that we assume that, like almost always in the literature, the noise is sub-Gaussian. As an example, let's consider Bernoulli-noise such that $\mathbb{P}(r_{t,i} = 1) = x_{t,i}^* = 1 - \mathbb{P}(r_{t,i} = 0)$. Then, the first environment when arm 1 manipulates as suggested is identical to the second environment when the arms are truthful. In the second environment, to suffer sublinear regret we must select arm 1 order T many times. However, in the first environment, we must select arm 1 only order $o(T)$ many times.

Discussion. Based on these observations, we expect that we would have to make additional (strong) assumptions about the distribution of the noise and the prior knowledge of the learner in order to drop Assumption 1. As an example, let's assume standard normal noise $\mathcal{N}(0, 1)$ and that the learner knows that the variance is always 1 a priori. Then, one potentially effective approach would be to extend our current grim trigger to additionally threaten arms that misreport their variance. Of course, the variance is unknown, however, we could estimate the variance of each arm's reports separately and use confidence intervals around the estimated variance. We could then threaten an arm with elimination if the arm's estimated variance falls out of the confidence interval. However, we expect there to be several technical subtleties in analyzing such variance-aware mechanisms.

D Proof of Theorem 4.1 and Theorem 4.2

D.1 Preliminaries

We begin with some preliminaries that we use in the proofs of both Theorem 4.1 and Theorem 4.2. Note that $\mathbb{E}[r_{t,i}] = \langle \theta^*, x^*_{t,i} \rangle$ and we denote the total true mean reward by

$$\hat{r}^*_{t,i} \coloneqq \sum_{\ell \le t \colon i_\ell = i} \langle \theta^*, x^*_{\ell,i} \rangle.$$

Let us also recall the definition of the total observed reward as $\hat{r}_{t,i} \coloneqq \sum_{\ell \le t \colon i_\ell = i} r_{\ell,i}$ and recall its upper confidence bound $\mathrm{UCB}_t(\hat{r}_{t,i}) \coloneqq \hat{r}_{t,i} + 2\sqrt{n_t(i)\log(T)}$ and define the lower confidence bound $\mathrm{LCB}_t(\hat{r}_{t,i}) \coloneqq \hat{r}_{t,i} - 2\sqrt{n_t(i)\log(T)}$.

We now analyze the basic properties of the grim trigger condition of GGTM, which eliminates arm i in round t if

$$\sum_{\ell \le t \colon i_\ell = i} \langle \theta^*, x_{\ell,i} \rangle > \mathrm{UCB}_t(\hat{r}_{t,i}),$$

or equivalently

$$\sum_{\ell \le t \colon i_\ell = i} \big(\langle \theta^*, x_{\ell,i} \rangle - r_{\ell,i} \big) > 2\sqrt{n_t(i)\log(T)}.$$

Define the *good event* $\mathcal{G}$ as the event that

$$\mathcal{G} \coloneqq \left\{ \mathrm{LCB}_t(\hat{r}_{t,i}) \le \hat{r}^*_{t,i} \le \mathrm{UCB}_t(\hat{r}_{t,i}) \ \forall t \in [T], i \in [K] \right\}.$$

By Hoeffding's inequality, we know that the good event occurs with probability at least $\mathbb{P}(\mathcal{G}) \ge 1 - \frac{1}{T^2}$. Next, let

$$\tau_i \coloneqq \min\{t \in [T] \colon i \notin A_t\}$$

denote the first round in which i is no longer active and, by convention, let $\tau_i = T$ if $i \in A_T$. By design of the grim trigger condition, note that $\tau_i = T$ for all $i \in [K]$ on the good event $\mathcal{G}$ if all arms always report truthfully.

We now provide a general result bounding the maximal amount of manipulation any arm can exercise before being eliminated by GGTM.

Lemma D.1. *On the good event $\mathcal{G}$, for any round $t \in [T]$ and any arm $i \in A_t$ it holds that*

$$\sum_{\ell \le t \colon i_\ell = i} \big(\langle \theta^*, x_{\ell,i} \rangle - x^*_{\ell,i} \big) \le 4\sqrt{n_t(i)\log(T)}.$$

From the definition of τ_i this entails that

$$\sum_{\ell \le \tau_i \colon i_\ell = i} \big(\langle \theta^*, x_{\ell,i} \rangle - x^*_{\ell,i} \big) \le 4\sqrt{n_{\tau_i}(i)\log(T)}.$$

Proof. On the good event $\mathcal{G}$, it holds that

$$\sum_{\ell \le t \colon i_\ell = i} \big(\langle \theta^*, x^*_{\ell,i} \rangle - r_{\ell,i} \big) \in \big[-2\sqrt{n_t(i)\log(T)}, +2\sqrt{n_t(i)\log(T)} \big],$$

which implies that

$$\sum_{\ell \le t \colon i_\ell = i} \big(\langle \theta^*, x_{\ell,i} \rangle - r_{\ell,i} \big) \ge \sum_{\ell \le t \colon i_\ell = i} \langle \theta^*, x_{\ell,i} - x^*_{\ell,i} \rangle - 2\sqrt{n_t(i)\log(T)}.$$

Hence, if $\sum_{\ell \le t \colon i_\ell = i} \langle \theta^*, x_{\ell,i} - x^*_{\ell,i} \rangle > 4\sqrt{n_t(i)\log(T)}$, then

$$\sum_{\ell \le t \colon i_\ell = i} \big(\langle \theta^*, x_{\ell,i} \rangle - r_{\ell,i} \big) > 2\sqrt{n_t(i)\log(T)},$$

which means that arm i is eliminated from A_t. Finally, τ_i is defined as the first round such that $i \notin A_t$ so that

$$\sum_{\ell \le t \colon i_\ell = i} \big(\langle \theta^*, x_{\ell,i} \rangle - x^*_{\ell,i} \big) \le 4\sqrt{n_t(i)\log(T)}$$

for all $t \le \tau_i$ and $\sum_{\ell \le \tau_i + 1 \colon i_\ell = i} \big(\langle \theta^*, x_{\ell,i} \rangle - x^*_{\ell,i} \big) > 4\sqrt{n_{\tau_i}(i)\log(T)}$. □

https://doi.org/10.52202/079017-3703

For completeness, we also formally state the fact that on the good event $\mathcal{G}$ any truthful arm is not eliminated with high probability.

Lemma D.2. *If arm i reports truthfully every round, i.e., plays strategy σ_i^* with $x_{t,i} = x_{t,i}^*$ for all round $t \in [T]$, then on the good event $\mathcal{G}$ arm i stays active for all rounds.*

Proof. When arm i is truthful, then $\sum_{\ell \le t:\, i_\ell = i} \langle \theta^*, x_{\ell,i} \rangle = \hat{r}_{t,i}^*$. On the good event, $\hat{r}_{t,i}^* \le \mathrm{UCB}_t(\hat{r}_{t,i})$ for all $t \in [T]$. Hence, the grim trigger condition is never satisfied and arm i remains active throughout all T rounds. □

D.2 Proof of Theorem 4.1

Proof of Theorem 4.1. We have to show that the strategy profile $\boldsymbol{\sigma}^*$, where every arm always truthfully reports their context, i.e., $x_{t,i} = x_{t,i}^*$ for all $(t,i) \in [T] \times [K]$, forms a $\tilde{\mathcal{O}}(\sqrt{T})$-Nash equilibrium for the arms under GGTM. We do this by showing that any deviating strategy σ_i for arm i cannot gain more than this $\sqrt{T}$ clicks. Recall that i_t^* is the optimal arm in round t and i_t the arm the learner selects.

We begin by deriving the minimum utility of every arm when everyone is truthful. To this end, let $n_T^*(i) := \sum_{t=1}^T \mathbb{1}\{i_t^* = i\}$ be the number of times arm i is the optimal arm. If every arm i is truthful, then on the good event $\mathcal{G}$ no arm gets eliminated (Lemma D.2) and $\langle \theta^*, x_{t,i} \rangle = \langle \theta^*, x_{t,i}^* \rangle$ for all $(t,i) \in [T] \times [K]$. As a result, GGTM pulls the optimal arm i_t^* in every round t. First, note that:

$$\mathbb{E}_{\boldsymbol{\sigma}^*}[n_T(i)] \ge n_T^*(i) - \frac{1}{T},$$

because on the good event $\mathcal{G}$ (when everyone is truthful), we have $n_T(i) \ge n_T^*(i)$. Since by construction $\mathbb{P}(\mathcal{G}) \ge 1 - 1/T^2$, the lower bound follows.

Next, we bound the utility of a deviating strategy σ_i in response to GGTM and the other arms' truthful strategies σ_{-i}^*. On the good event $\mathcal{G}$, when the arms play strategies $(\sigma_i, \sigma_{-i}^*)$, we have

$$\begin{aligned} n_T(i) &= \sum_{t=1}^T \mathbb{1}\{i_t = i, i_t^* = i\} + \sum_{t=1}^T \mathbb{1}\{i_t = i, i_t^* \ne i\} \\ &\le \sum_{t=1}^T \mathbb{1}\{i_t^* = i\} + \sum_{t=1}^T \mathbb{1}\{i_t = i, i_t^* \ne i\} \\ &= n_T^*(i) + \sum_{t=1}^T \mathbb{1}\{i_t = i, i_t^* \ne i\}. \end{aligned}$$

We will now bound the sum on the right hand side from above.

Every arm $j \ne i$ is truthful and therefore, on the good event, $j \in A_t$ for all t. If the optimal arm is not i, i.e., $i_t^* \ne i$, it means that $\langle \theta^*, x_{t,i_t^*}^* - x_{t,i}^* \rangle > 0$. Next, since GGTM selects the arms greedily according to the reported reward, the event $i_t = i$ implies that

$$\langle \theta^*, x_{t,i} \rangle \ge \langle \theta^*, x_{t,i_t^*}^* \rangle,$$

where we used that any arm $i_t^* \ne i$ is truthful so that $\langle \theta^*, x_{t,i_t^*} \rangle = \langle \theta^*, x_{t,i_t^*}^* \rangle$. As a result, we can apply Lemma D.1 to obtain

$$\begin{aligned} \sum_{t=1}^T \mathbb{1}\{i_t^* \ne i, i_t = i\} \langle \theta^*, x_{t,i_t^*}^* - x_{t,i}^* \rangle &\le \sum_{t=1}^T \mathbb{1}\{i_t = i, i_t^* \ne i\} \langle \theta^*, x_{t,i} - x_{t,i}^* \rangle \\ &\le \sum_{t=1}^{\tau_i} \mathbb{1}\{i_t = i\} \langle \theta^*, x_{t,i} - x_{t,i}^* \rangle \\ &\le 4\sqrt{n_{\tau_i}(i) \log(T)}. \end{aligned}$$

Since for $i_t^* \neq i$ the gap $\langle \theta^*, x^*_{t,i_t^*} - x^*_{t,i} \rangle$ is positive and assumed to be constant, we get that $\sum_{t=1}^{T} \mathbb{1}\{i_t^* \neq i, i_t = i\} \leq \mathcal{O}(\sqrt{n_{\tau_i}(i)\log(T)})$. We coarsely upper bound $n_{\tau_i}(i)$ by T and using that the good event $\mathcal{G}$ has probability at least $1 - 1/T^2$, we obtain

$$\mathbb{E}_{\sigma_i, \sigma^*_{-i}}[n_T(i)] \leq n_T^*(i) + \mathcal{O}\left(\sqrt{T\log(T)}\right).$$

We have thus shown that

$$\mathbb{E}_{\boldsymbol{\sigma}^*}[n_T(i)] \geq \mathbb{E}_{\sigma_i, \sigma^*_{-i}}[n_T(i)] + \mathcal{O}\left(\sqrt{T\log(T)}\right)$$

for any deviating (dishonest) strategy σ_i. This means that $\boldsymbol{\sigma}^*$ is a $\tilde{\mathcal{O}}(\sqrt{T})$-Nash equilibrium for the arms.

Finally, the regret of GGTM when the arms are truthful is quickly bounded by $1/T$ by using the fact that on the good event no arm gets eliminated and, therefore, GGTM picks the round-optimal arm every round. The event that $\mathcal{G}$ does not hold has probability at most $1/T^2$ which implies expected regret $1/T$, i.e., $R_T(\text{GGTM}, \boldsymbol{\sigma}^*) \leq 1/T$.

□

D.3 Proof of Theorem 4.2

Proof of Theorem 4.2. The proof of Theorem 4.2 is notably more involved than that of Theorem 4.1, even though the general proof idea remains similar.

We begin by decomposing of GGTM into the rounds where the optimal arm is active and the rounds in which it is being ignored. To this end, recall the definition of the arm that is optimal in round t as $i_t^* := \text{argmax}_{i\in[K]}\langle \theta^*, x^*_{t,i}\rangle$. We have

$$R_T = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* \in A_t\}\langle \theta^*, x^*_{t,i_t^*} - x^*_{t,i_t}\rangle\right]}_{I_1} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* \notin A_t\}\langle \theta^*, x^*_{t,i_t^*} - x^*_{t,i_t}\rangle\right]}_{I_2}.$$

We now bound I_1 and I_2 separately as follows.

Lemma D.3 (Bounding I_1)**.**

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* \in A_t\}\langle \theta^*, x^*_{t,i_t^*} - x^*_{t,i_t}\rangle\right] \leq \mathcal{O}\left(\sqrt{KT\log(T)}\right).$$

Proof. Let i_t denote the selection of GGTM in round t. Recall that GGTM greedily selects the arm in A_t with highest reported value in round t, that is, $\langle \theta^*, x_{t,i_t}\rangle = \max_{i\in A_t}\langle \theta^*, x_{t,i}\rangle$. Consequently, on event $\{i_t^* \in A_t\}$, we have

$$\langle \theta^*, x^*_{t,i_t^*}\rangle \leq \langle \theta^*, x_{t,i_t^*}\rangle \leq \max_{i\in A_t}\langle \theta^*, x_{t,i}\rangle = \langle \theta^*, x_{t,i_t}\rangle,$$

where the first inequality holds by the assumption the optimal arm i_t^* reports their value at least as high as their true value. As a consequence, it holds that $\langle \theta^*, x^*_{t,i_t^*} - x^*_{t,i_t}\rangle \leq \langle \theta^*, x_{t,i_t} - x^*_{t,i_t}\rangle$ which implies:

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* \in A_t\}\langle \theta^*, x^*_{t,i_t^*} - x^*_{t,i_t}\rangle\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* \in A_t\}\langle \theta^*, x_{t,i_t} - x^*_{t,i_t}\rangle\right]. \tag{5}$$

When $i_t^* \in A_t$, a necessary condition for arm i to be selected in round t (i.e., $i_t = i$) is that $t \leq \tau_i$. Finally, we split the sum into each arm's contribution and apply Lemma D.1 to obtain

$$(5) \leq \mathbb{E}\left[\sum_{i=1}^{K}\sum_{t=1}^{\tau_i} \mathbb{1}\{i_t = i\}\langle \theta^*, x_{t,i_t} - x^*_{t,i_t}\rangle\right] \leq \mathbb{E}\left[\sum_{i=1}^{K} 4\sqrt{n_{\tau_i}(i)\log(T)}\right] \leq 4\sqrt{KT\log(T)},$$

where the last step follows from Jensen's inequality by bounding $n_{\tau_i}(i)$ by $n_T(i)$ and using that $\sum_{i=1}^{K} n_T(i) \leq T$ by definition of $n_T(i)$. □

https://doi.org/10.52202/079017-3703

While bounding I_1 is fairly straightforward and we did not have to rely on the fact that the arms respond in Nash equilibrium, bounding I_2 becomes more challenging as we must argue that it is in each arms' interest to maintain active for a sufficiently long time.

Lemma D.4 (Bounding I_2)**.**

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* \notin A_t\}\big(\mu_{t,i_t^*}^* - \mu_{t,i_t}^*\big)\right] \leq 5K^2\sqrt{KT\log(T)} \tag{6}$$

Proof. To bound I_2 we argue via the best response property of the Nash equilibrium. This requires some intermediate steps. We begin with a lower bound on the expected number of selections any arm must receive when the arms act according to a Nash equilibrium under GGTM.

Recall the definition $n_\tau^*(i) \coloneqq \sum_{t=1}^{\tau} \mathbb{1}\{i_t^* = i\}$ and that the indicator variables $\mathbb{1}\{i_t^* = i\}$ are not random, since we work under an adversarially chosen sequence of true contexts. In contrast, the indicator $\mathbb{1}\{i_t = i\}$ is a random variable as it generally depends on the random reward observations and any randomization of the algorithm.

The following lemma provides a lower bound on the number of allocations any arm must receive in equilibrium. To prove the lemma, we show that we are able to protect any truthful arm from losing more than order $\sqrt{KT}$ allocations to manipulating arms. This is crucial as it would be impossible to incentivize approximately truthful arm behavior if an arm would lose too many allocations, e.g., order T many, by doing so.

A key challenge here is that under two different strategies σ_i and σ_i', the set of active arms can be quite different. This is the case since even though we estimate each arm's expected reward independently, arm i can still slightly influence the elimination of some other arm j by poaching selections from them. As a result, we must content ourselves with a more conservative bound than one may originally expect.

Lemma D.5. *Let* $\boldsymbol{\sigma} \in \mathrm{NE}(\mathrm{GGTM})$. *Then,*

$$\mathbb{E}_{\boldsymbol{\sigma}}[n_T(i)] \geq n_T^*(i) - \mathcal{O}\big(\sqrt{KT\log(T)}\big).$$

In particular, it holds that $\mathbb{E}_{\boldsymbol{\sigma}}[n_t(i)] \geq n_t^*(i) - \mathcal{O}\big(\sqrt{KT\log(T)}\big)$ *for any* $t \in [T]$.

Proof. We use the fact that if $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_K)$ is a NE under GGTM, then σ_i must be a best response to σ_{-i}, i.e., $\mathbb{E}_{\sigma_i,\sigma_{-i}}[n_T(i)] \geq \mathbb{E}_{\sigma_i',\sigma_{-i}}[n_T(i)]$ for all strategies σ_i'. In particular, it must hold for the truthful strategy σ_i^* that

$$\mathbb{E}_{\sigma_i,\sigma_{-i}}[n_T(i)] \geq \mathbb{E}_{\sigma_i^*,\sigma_{-i}}[n_T(i)].$$

We focus on the good event $\mathcal{G}$ so that $i \in A_t$ for all t given that arm i is truthful. We are interested in the number of rounds such that $i_t^* = i$ and $i_t \neq i$. Given strategies $(\sigma_i^*, \sigma_{-i})$ so that $i \in A_T$ on the good event, we have

$$\sum_{t=1}^{T} \mathbb{1}\{i_t^* = i, i_t \neq i\} = \sum_{j\neq i}\sum_{t=1}^{\tau_j} \mathbb{1}\{i_t^* = i, i_t = j\}.$$

Note that $i_t = j$ with $i_t^* = i \in A_t$ implies that $\langle\theta^*, x_{t,j}\rangle \geq \langle\theta^*, x_{t,i}\rangle$. Moreover, because i is truthful and $i_t^* = i$, we have $\langle\theta^*, x_{t,i}\rangle = \langle\theta^*, x_{t,i_t^*}^*\rangle$ so that $\langle\theta^*, x_{t,j}\rangle > \langle\theta^*, x_{t,i_t^*}^*\rangle$. As a result,

$$\langle\theta^*, x_{t,i_t^*}^* - x_{t,j}^*\rangle < \langle\theta^*, x_{t,j} - x_{t,j}^*\rangle.$$

It then follows from Lemma D.1 that

$$\begin{aligned}
\sum_{t=1}^{\tau_j} \mathbb{1}\{i_t^* = i, i_t = j\}\langle\theta^*, x_{t,i_t^*}^* - x_{t,j}^*\rangle &< \sum_{t=1}^{\tau_j} \mathbb{1}\{i_t^* = i, i_t = j\}\langle\theta^*, x_{t,j} - x_{t,j}^*\rangle &(7)\\
&\leq \sum_{t=1}^{\tau_j} \mathbb{1}\{i_t = j\}\langle\theta^*, x_{t,j} - x_{t,j}^*\rangle &(8)\\
&\leq 4\sqrt{n_{\tau_j}(j)\log(T)}. &(9)
\end{aligned}$$

Since $\langle \theta^*, x^*_{t,i^*_t} - x^*_{t,j} \rangle$ is constant for $i^*_t = i, i_t = j$, we obtain

$$\sum_{j \neq i} \sum_{t=1}^{T} \mathbb{1}\{i^*_t = i, i_t \neq i\} \leq \sum_{j \neq i} \mathcal{O}\big(\sqrt{n_{\tau_j}(j) \log(T)}\big) \leq \mathcal{O}\big(\sqrt{KT \log(T)}\big),$$

where the last inequality follows from Jensen's inequality. Recalling that $\mathbb{P}(\mathcal{G}) \geq 1 - 1/T^2$, this provides us with the following lower bound on the utility of the truthful strategy

$$\begin{aligned}
\mathbb{E}_{\sigma^*_i, \sigma_{-i}}[n_T(i)] &= \mathbb{E}_{\sigma^*_i, \sigma_{-i}} \left[\sum_{t=1}^{T} \mathbb{1}\{i_t = i\} \right] \\
&\geq \mathbb{E}_{\sigma^*_i, \sigma_{-i}} \left[\sum_{t=1}^{T} \mathbb{1}\{i^*_t = i\} - \sum_{t=1}^{T} \mathbb{1}\{i^*_t = i, i_t \neq i\} \right] \\
&\geq n^*_T(i) - \mathcal{O}\big(\sqrt{KT \log(T)}\big)
\end{aligned}$$

Note that we, like before, we account for the event $\mathcal{G}^c$ by increasing the constant factor by one, since $(1 - 1/T^2) n_T(i) \geq n_T(i) - 1/T$ as $n_T(i) \leq T$.

Since σ_i has to be a best response to σ_{-i}, it must be as least as good as σ^*_i sot that

$$\mathbb{E}_{\boldsymbol{\sigma}}[n_T(i)] \geq \mathbb{E}_{\sigma^*_i, \sigma_{-i}}[n_T(i)] \geq n^*_T(i) - \mathcal{O}\big(\sqrt{KT \log(T)}\big).$$

To get the result for any $t \in [T]$, suppose that on the good event $\mathcal{G}$ it holds that $n_t(i) < n^*_t(i) - \omega(\sqrt{KT \log(T)})$. Now, recall from equation (7) that the number of rounds such that $i^*_t = i$ and $i_t \neq i$ is bounded by $\mathcal{O}(\sqrt{Kt \log(T)})$ on event $\mathcal{G}$. Hence, since we assumed that $n_t(i) < n^*_t(i) - \omega(\sqrt{KT \log(T)})$ and $\langle \theta^*, x_{t,i} - x^*_{t,i} \rangle \geq 0$, it must hold that $\tau_i < t$. Consequently, on the good event $\mathcal{G}$, we obtain

$$n_{\tau_i}(i) \leq n_t(i) < n^*_t(i) - \omega\big(\sqrt{KT \log(T)}\big).$$

This implies that

$$\mathbb{E}_{\boldsymbol{\sigma}}[n_{\tau_i}(i)] < \big(1 - 1/T^2\big) \left(n^*_t(i) - \omega\big(\sqrt{KT \log(T)}\big) \right) + 1/T \leq n^*_t(i) - \omega\big(\sqrt{KT \log(T)}\big).$$

This stands in contradiction to the earlier lower bound of $\mathbb{E}_{\boldsymbol{\sigma}}[n_T(i)] = \mathbb{E}_{\boldsymbol{\sigma}}[n_{\tau_i}(i)] \geq n^*_T(i) - \mathcal{O}\big(\sqrt{KT \log(T)}\big)$.

□

Next, we provide an upper bound on the number of times an arm is pulled in any Nash equilibrium. In other words, we bound the profit any arm can make under GGTM from misreporting contexts.

Lemma D.6. *Let $\boldsymbol{\sigma}$ be any NE under* GGTM. *Then,*

$$\mathbb{E}_{\boldsymbol{\sigma}}[n_T(i)] \leq \mathbb{E}_{\boldsymbol{\sigma}}[n^*_{\tau_i}(i)] + \mathcal{O}\big((K-1)\sqrt{KT \log(T)}\big)$$

Proof. Note that $\sum_{i=1}^{K} n^*_\tau(i) = \sum_{i=1}^{K} \sum_{t=1}^{\tau} \mathbb{1}\{i^*_t = i\} = \tau$ for any $\tau \in [T]$. Using Lemma D.5, we then obtain

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}}[n_T(i)] &= \mathbb{E}_{\boldsymbol{\sigma}}[n_{\tau_i}(i)] \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \left[\sum_{t=1}^{\tau_i} \mathbb{1}\{i_t = i\} \right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \left[\sum_{t=1}^{\tau_i} (1 - \mathbb{1}\{i_t \neq i\}) \right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}[\tau_i] - \sum_{j \neq i} \mathbb{E}_{\boldsymbol{\sigma}}[n_{\tau_i}(j)] \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}}[\tau_i] - \sum_{j \neq i} \mathbb{E}_{\boldsymbol{\sigma}}[n^*_{\tau_i}(j)] + \mathcal{O}\big((K-1)\sqrt{KT \log(T)}\big) \\
&= \mathbb{E}_{\boldsymbol{\sigma}}[n^*_{\tau_i}(i)] + \mathcal{O}\big((K-1)\sqrt{KT \log(T)}\big)
\end{aligned}$$

https://doi.org/10.52202/079017-3703

□

Combining Lemma D.5 and Lemma D.6 we get for any Nash equilibrium $\boldsymbol{\sigma} \in \mathrm{NE}(\mathrm{GGTM})$ that

$$\mathbb{E}_{\boldsymbol{\sigma}}[n_T^*(i)] - \mathcal{O}\big(\sqrt{KT\log(T)}\big) \leq \mathbb{E}_{\boldsymbol{\sigma}}[n_T(i)] \leq \mathbb{E}_{\boldsymbol{\sigma}}[n_{\tau_i}^*(i)] - \mathcal{O}\big((K-1)\sqrt{KT\log(T)}\big),$$

which implies that

$$\mathbb{E}_{\boldsymbol{\sigma}}[n_T^*(i) - n_{\tau_i}^*(i)] \leq \mathcal{O}\big(K\sqrt{KT\log(T)}\big). \tag{10}$$

The expression $n_T^*(i) - n_{\tau_i}^*(i)$ is the number of rounds where arm i was optimal but already eliminated by the grim trigger. As a result, we can express the total number of rounds where the round-optimal arm i_t^* was no longer active as follows.

Lemma D.7. *For any $\boldsymbol{\sigma}$, we have*

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* \notin A_t\}\right] = \sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\sigma}}[n_T^*(i) - n_{\tau_i}^*(i)].$$

Proof. Rewriting $\{i_t^* \notin A_t\}$ yields

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* \notin A_t\}\right] &= \sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* = i, i \notin A_t\}\right] \\
&= \sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* = i\} - \sum_{t=1}^{T} \mathbb{1}\{i_t^* = i, i \in A_t\}\right] \\
&= \sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\sigma}}\left[n_T^*(i) - n_{\tau_i}^*(i)\right].
\end{aligned}$$

□

Finally, note that $\langle \theta^*, x_{t,i_t^*}^* - x_{t,i_t}^* \rangle \leq 1$ so that from equation (10) it follows that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* \notin A_t\}\langle \theta^*, x_{t,i_t^*}^* - x_{t,i_t}^* \rangle\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* \notin A_t\}\right] \\
&= \sum_{i=1}^{K} \mathbb{E}\left[n_T^*(i) - n_{\tau_i}^*(i)\right] \\
&\leq \sum_{i=1}^{K} \mathcal{O}\big(K\sqrt{KT\log(T)}\big) = \mathcal{O}\big(K^2\sqrt{KT\log(T)}\big).
\end{aligned}$$

□

Connecting the bound on I_1 and I_2, we then obtain the final regret bound of Theorem 4.2

$$R_T(\mathrm{GGTM}, \boldsymbol{\sigma}) \leq \mathcal{O}\left(\underbrace{\sqrt{KT\log(T)}}_{\text{Lemma D.3}} + \underbrace{K^2\sqrt{KT\log(T)}}_{\text{Lemma D.4}}\right) \leq \tilde{\mathcal{O}}\left(K^2\sqrt{KT}\right).$$

□

Remark D.8. *We want to briefly comment on the existence of a Nash equilibrium. Since each arm's strategy space, given by $\{x \in \mathbb{R}^d \colon \|x\|_2 \leq 1\}$ in every round, is continuous, it is not obvious that a Nash equilibrium for the arms exists under every algorithm. However, Glickberg's theorem shows that the continuity of the arms' utility in the arms' strategies is a sufficient condition for the existence of a NE, since the strategy space is compact. We can then ensure the continuity by, e.g., choosing arms proportionally to $\exp(T\langle \theta^*, x_{t,i}\rangle)$ in* GGTM *and $\exp(T\mathrm{UCB}_t(x_{t,i}))$ in* OptGTM*, and remarking that the probability of eliminating arm i in round t is continuous in $x_{t,i}$ conditional on any history. Due to the exponential scaling with T the effect of such slight randomization is negligible in the regret analysis.*

E Proof of Theorem 5.1 and Theorem 5.2

The following preliminaries are fundamental to the proofs of Theorem 5.1 and Theorem 5.2 so that we derive them jointly here.

E.1 Preliminaries

We begin by recalling the definition of the least-squares estimator w.r.t. arm i's reported contexts and the corresponding confidence ellipsoid $C_{t,i}$. Note that since the arms are manipulating their contexts, the least-squares estimator may not accurately estimate θ^* and θ^* may not be contained in $C_{t,i}$ w.h.p. As discussed in the main text, since accurate estimation of θ^* appears hopeless, the main idea idea is to incentivize arms to report contexts such that our expected reward does not differ substantially from the observed reward.

The least-squares estimator w.r.t. arm i is given by

$$\hat{\theta}_{t,i} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left(\sum\nolimits_{\ell < t\colon i_\ell = i} \left(\langle \theta, x_{\ell,i} \rangle - r_{\ell,i} \right)^2 + \lambda \|\theta\|_2^2 \right),$$

where $\lambda > 0$. In the algorithm, we set the penalty factor to $\lambda = 1$. The closed form solution is then given by

$$\hat{\theta}_{t,i} = V_{t,i}^{-1} \sum_{\ell < t\colon i_\ell = i} x_{\ell,i} r_{\ell,i} \quad \text{with} \quad V_{t,i} = \lambda I + \sum_{\ell < t\colon i_\ell = i} x_{\ell,i} x_{\ell,i}^\top.$$

The confidence set $C_{t,i}$ is defined as

$$C_{t,i} = \left\{ \theta \in \mathbb{R}^d \colon \|\hat{\theta}_{t,i} - \theta\|_{V_t}^2 \leq \beta_{t,i} \right\},$$

where we let $\beta_{t,i} = \sqrt{d \log\left(\frac{1 + n_t(i)/\lambda}{\delta} \right)} + \sqrt{\lambda} S$ with $\|\theta^*\|_2 \leq S$ and $\delta = 1/T^2$.

We now translate the standard result used to assert the validity of the confidence set to our situation. Clearly, when the sequence of $x_{t,i}$ differs significantly from the true contexts $x_{t,i}^*$, the true parameter θ^* will not be contained in $C_{t,i}$. Instead, we will formulate the concentration result as follows.

Lemma E.1. *Suppose there exists $\theta_i^* \in \mathbb{R}^d$ such that for all t with $i_t = i$:*

$$\langle \theta^*, x_{t,i}^* \rangle = \langle \theta_i^*, x_{t,i} \rangle. \tag{11}$$

In other words, the reported features $x_{t,i}$ are linearly realizable by some parameter θ_i^*.

For any $\delta \in (0,1)$ let the confidence size be

$$\beta_{t,i} = \sqrt{d \log\left(\frac{1 + n_t(i)/\lambda}{\delta} \right)} + \sqrt{\lambda} S,$$

where $\|\theta_i^*\|_2 \leq S$. Note that the typical expression also includes some constant L such that $\|x_{t,i}\|_2 \leq L$, which we here simply set to 1. With probability at least $1 - \delta$ it then holds that $\theta_i^* \in C_{t,i}$. In what follows, we choose $\delta = 1/T^2$.

As a special case, when arm i is always truthful so that $x_{t,i} = x_{t,i}^*$, the true parameter θ^* trivially satisfies (11) *and the result reduces to the standard confidence bound statement [1, 24] restricted to observations from arm i.*

Proof. Let θ_i^* satisfy (11). Then, note that $r_{t,i} \coloneqq \langle \theta^*, x_{t,i}^* \rangle + \eta_t = \langle \theta_i^*, x_{t,i} \rangle + \eta_t$. Hence, the sequence of reported features $x_{t,i}$ and observed reward $r_{t,i}$ yield a standard linear contextual bandit structure with unknown parameter θ_i^* (instead of θ^*). Then, to obtain the confidence bound follow the arguments from [1, 24], where we remark that we choose the confidence radius $\beta_{t,i}$ arm specific. However, we could also choose a larger confidence radius such as $\beta_t \approx d \log(t)$ or even constant $\beta \approx d \log(T)$. This will only have a negligible effect on the final regret. □

The grim trigger condition (4) of OptGTM stated that arm i gets eliminated in round t if

$$\sum_{\ell \leq t\colon i_\ell = i} \left(\langle \hat{\theta}_{\ell,i}, x_{\ell,i} \rangle - \sqrt{\beta_\ell} \|x_{\ell,i}\|_{V_{\ell,i}^{-1}} \right) > \sum_{\ell \leq t\colon i_\ell = i} r_{\ell,i} + 2\sqrt{n_t(i) \log(T)}. \tag{12}$$

https://doi.org/10.52202/079017-3703

Equivalently, $\sum_{\ell \le t:\ i_\ell = i} \mathrm{LCB}_{\ell,i}(x_{\ell,i}) > \mathrm{UCB}_t(\hat{r}_{t,i})$.

As a sanity check, we show that when an arm always reports truthfully, i.e., $x_{t,i} = x^*_{t,i}$ for all t, it doesn't get eliminated with probability at least $1 - 1/T^2$.

Lemma E.2. *When arm i always reports truthfully it does not get eliminated with high probability, that is, $i \in A_T$ with probability at least $1 - 1/T^2$.*

Proof. We consider the event that the true parameter θ^* is contained in $C_{t,i}$, i.e., $\mathcal{G}'_i \coloneqq \{\theta^* \in C_{t,i}\ \forall t \in [T]\}$. The event $\mathcal{G}'_i$ has probability at least $1 - 1/T^2$ according to Lemma E.1 when arm i is truthful. Moreover, suppose that the reward observations concentrate as well, i.e., we assume the good event $\mathcal{G} \coloneqq \left\{\mathrm{LCB}_t(\hat{r}_{t,i}) \le \hat{r}^*_{t,i} \le \mathrm{UCB}_t(\hat{r}_{t,i})\ \forall t \in [T], i \in [K]\right\}$. Recall that $\mathcal{G}$ has probability at least $1 - 1/T^2$ according to Hoeffding's inequality. A union bound then shows that the intersection of the two events has probability at least $1 - 2/T^2$.

Now, since arm i is truthful, we have $x_{t,i} = x^*_{t,i}$ and $\langle \theta, x_{t,i} \rangle = \langle \theta, x^*_{t,i} \rangle$ for all $\theta \in \mathbb{R}^d$. Using Cauchy-Schwarz inequality and the fact that $\theta^* \in C_{t,i}$, we get that

$$\langle \hat{\theta}_{t,i}, x_{t,i} \rangle - \langle \theta^*, x^*_{t,i} \rangle = \langle \hat{\theta}_{t,i} - \theta^*, x^*_{t,i} \rangle \le \|\hat{\theta}_{t,i} - \theta^*\|_{V_{t,i}} \|x^*_{t,i}\|_{V^{-1}_{t,i}} \le \sqrt{\beta_{t,i}} \|x_{t,i}\|_{V^{-1}_{t,i}}$$

Moreover, as we work on the good event $\mathcal{G}$, we have

$$\sum_{\ell \le t:\ i_\ell = i} \left(\langle \theta^*, x^*_{\ell,i} \rangle - r_{\ell,i}\right) \in [-2\sqrt{n_t(i)\log(T)}, +2\sqrt{n_t(i)\log(T)}].$$

Combining these two statements yields

$$\sum_{\ell \le t:\ i_\ell = i} \left(\langle \hat{\theta}_{\ell,i}, x_{t,i} \rangle - r_{\ell,i}\right) \le \sum_{\ell \le t:\ i_\ell = i} \sqrt{\beta_{\ell,i}} \|x_{\ell,i}\|_{V^{-1}_{\ell,i}} + 2\sqrt{n_t(i)\log(T)}$$

for all $t \in [T]$. In other words, the grim trigger condition is never satisfied so that $i \in A_T$ on event $\mathcal{G} \cap \mathcal{G}'_i$, which, as we saw, occurs with probability at least $1 - 1/T^2$. □

We now analyze the grim trigger of the OptGTM algorithm. As before, let $\tau_i \coloneqq \min\{t\colon i \notin A_t\}$ with the convention that $\tau_i = T$ if $i \in A_T$. The following lemma upper bounds the total amount of manipulation that an arm can exert before being eliminated by OptGTM's grim trigger elimination rule.

Lemma E.3. *On the good event $\mathcal{G}$:*

$$\sum_{t \le \tau_i:\ i_t = i} \left(\langle \hat{\theta}_{t,i}, x_{t,i} \rangle - \langle \theta^*, x^*_{t,i} \rangle\right) \le \sum_{t \le \tau_i:\ i_t = i} \sqrt{\beta_{t,i}} \|x_{t,i}\|_{V^{-1}_{t,i}} + 4\sqrt{n_{\tau_i}(i)\log(T)}.$$

Or equivalently, since $\mathrm{UCB}_{t,i}(x_{t,i}) = \langle \hat{\theta}_{t,i}, x_{t,i} \rangle + \sqrt{\beta_{t,i}} \|x_{t,i}\|_{V^{-1}_{t,i}}$, it holds that

$$\sum_{t \le \tau_i:\ i_t = i} \left(\mathrm{UCB}_{t,i}\left(x_{t,i}\right) - \langle \theta^*, x^*_{t,i} \rangle\right) \le \sum_{t \le \tau_i:\ i_t = i} 2\sqrt{\beta_{t,i}} \|x_{t,i}\|_{V^{-1}_{t,i}} + 4\sqrt{n_{\tau_i}(i)\log(T)}.$$

Proof. Let $t \in [T]$. On the good event $\mathcal{G} \coloneqq \left\{\mathrm{LCB}_t(\hat{r}_{t,i}) \le \hat{r}^*_{t,i} \le \mathrm{UCB}_t(\hat{r}_{t,i})\ \forall t \in [T], i \in [K]\right\}$ and by definition of $\mathrm{UCB}_{\ell,i}(x_{\ell,i}$, it follows that

$$\begin{aligned}
&\sum_{\ell \le t:\ i_\ell = i} \left(\langle \hat{\theta}_{\ell,i}, x_{\ell,i} \rangle - r_{\ell,i}\right) \\
&\qquad \ge \sum_{\ell \le t:\ i_\ell = i} \left(\mathrm{UCB}_{\ell,i}(x_{\ell,i}) - \langle \theta^*, x^*_{\ell,i} \rangle\right) - \underbrace{\sum_{\ell \le t:\ i_\ell = i} \sqrt{\beta_{\ell,i}} \|x_{\ell,i}\|_{V^{-1}_{\ell-1,i}} - 2\sqrt{n_t(i)\log(T)}}_{R \coloneqq}.
\end{aligned}$$

Hence, if $\sum_{\ell \le t:\ i_\ell = i} \left(\mathrm{UCB}_{\ell,i}(x_{\ell,i}) - \langle \theta^*, x^*_{\ell,i} \rangle\right) > 2R$, then

$$\sum_{\ell \le t:\ i_\ell = i} \left(\langle \hat{\theta}_{\ell,i}, x_{\ell,i} \rangle - r_{\ell,i}\right) > \sum_{\ell \le t:\ i_t = i} \sqrt{\beta_{\ell,i}} \|x_{\ell,i}\|_{V^{-1}_{\ell,i}} + 2\sqrt{n_t(i)\log(T)},$$

which means that arm i must have been eliminated in a previous round or in round t, i.e., $\tau_i > t$. Hence, for any $t \leq \tau_i$, the the left hand side must be smaller or equal to the right hand side.

Interestingly, notice that we worked on the good event $\mathcal{G}$ that only concerns the realization of the rewards and not the validity of the confidence set. This is important, since it is generally not true that the true parameter θ^* is contained in the confidence set $C_{t,i}$.

□

Lastly, before we begin with the proof Theorem 5.1 and Theorem 5.2, we establish a bound on the total exploration bonus, which after some additional work follows from the well-known elliptical potential lemma [1, 24].

Lemma E.4. *It holds that*

$$\sum_{t \leq \tau_i \colon i_t = i} \sqrt{\beta_t} \|x_{t,i}\|_{V_{t,i}^{-1}} \leq \mathcal{O}\left(d \log(T) \sqrt{n_{\tau_i}(i)}\right).$$

The constant on the right hand side can be derived from the choice of $\beta_{t,i}$.

Proof. Before we can apply the elliptical potential lemma[1], we need to make sure that the exploration bonus does not blow up in early rounds. To this end, recall the definition of

$$V_{t,i} \coloneqq \lambda I + \sum_{\ell < t \colon i_\ell = i} x_{\ell,i} x_{\ell,i}^\top.$$

Let $A = \sum_{\ell \leq t \colon i_\ell = i} x_{\ell,i} x_{\ell,i}^\top$. Note that A is positive semi-definite so that $\lambda I + A$ is positive definite and its inverse $(\lambda I + A)^{-1}$ as well. The matrix inversion lemma let's us express this inverse as

$$(\lambda I + A)^{-1} = \lambda I - (\lambda I + A)^{-1} A.$$

Now, the eigenvalues of $B = (\lambda I + A)^{-1} A$ are given by $\lambda_i / (1 + \lambda_i)$, where $\lambda_i \geq 0$ are the eigenvalues of A, which means that B is positive semi-definite. Consequently,

$$\|x_{t,i}\|_{V_{t,i}^{-1}}^2 = x_{t,i}^\top (\lambda I + A)^{-1} x_{t,i} = x_{t,i}^\top \lambda x_{t,i} - x_{t,i}^\top B x_{t,i} \leq \lambda \|x_{t,i}\|_2^2.$$

We assumed that $\|x_{t,i}\|_2^2 \leq 1$ (similarly we could assume an upper bound L) so that $\|x_{t,i}\|_{V_{t,i}^{-1}} \leq \sqrt{\lambda}$. For convenience, we set the penalty factor to $\lambda = 1$. We then apply Cauchy-Schwarz to get

$$\sum_{t \leq \tau_i \colon i_t = i} \sqrt{\beta_{t,i}} \|x_{t,i}\|_{V_{t,i}^{-1}} \leq \sqrt{n_{\tau_i}(i) \beta_T \sum_{t \leq \tau_i \colon i_t = i} \min\{1, \|x_{t,i}\|_{V_{t,i}^{-1}}^2\}}.$$

The elliptical potential lemma [1, 24] bounds the sum on the right hand side as

$$\sum_{t \leq \tau_i \colon i_t = i} \min\{1, \|x_{t,i}\|_{V_{t,i}^{-1}}^2\} \leq 2d \log\left(\frac{d + n_{\tau_i}(i)}{d}\right)$$

Finally, recall that we chose $\beta_{t,i} = \mathcal{O}\left(d \log\left(n_t(i)\right)\right)$, which then yields the claimed bound. □

E.2 Proof of Theorem 5.1

Proof of Theorem 5.1. We begin by proving that being truthful is an approximate Nash equilibrium under OptGTM.

Truthfulness is a $\tilde{\mathcal{O}}(d\sqrt{KT})$-NE. In a fist step, we show that if every arm is truthful, every arm is guaranteed at least $n_T^*(i) - \tilde{\mathcal{O}}(d\sqrt{T})$ utility, where $n_T^*(i) \coloneqq \sum_{t=1}^T \mathbb{1}\{i_t^* = i\}$. To this end, we write

$$\begin{aligned} n_T(i) &= \sum_{t=1}^T \mathbb{1}\{i_t = i, i_t^* = i\} + \sum_{t=1}^T \mathbb{1}\{i_t = i, i_t^* \neq i\} \\ &\geq \sum_{t=1}^T \mathbb{1}\{i_t^* = i\} - \sum_{t=1}^T \mathbb{1}\{i_t^* = i, i_t \neq i\}, \end{aligned}$$

https://doi.org/10.52202/079017-3703

and our task will be bounding the sum on the right hand side. We focus on the event that $\theta^* \in C_{t,i}$ for all $(t,i) \in [T] \times [K]$, which according to Lemma E.1 occurs with probability at least $1 - 1/T^2$.

Since $\theta^* \in C_{t,i}$, we have $\mathrm{UCB}_{t,i}(x^*_{t,i_t}) \geq \langle \theta^*, x^*_{t,i^*_t} \rangle$ for every $i \in [K]$. Let $i^*_t = i$ but $i_t = j$ with $j \neq i$. Keeping in mind that $x_{t,i} = x^*_{t,i}$ for all $(t,i) \in [T] \times [K]$, since all arms are truthful, this implies that

$$\mathrm{UCB}_{t,i}(x^*_{t,j}) \geq \mathrm{UCB}_{t,i}(x^*_{t,i}) \geq \langle \theta^*, x^*_{t,i^*_t} \rangle.$$

As a result, it holds that

$$\mathrm{UCB}_{t,i}(x^*_{t,j}) - \langle \theta^*, x^*_{t,j} \rangle \geq \langle \theta^*, x^*_{t,i^*_t} - x^*_{t,j} \rangle.$$

Next, since $\theta^* \in C_{t,i}$ and $\hat{\theta}_{t,i} \in C_{t,i}$, we find that

$$\begin{aligned}
\mathrm{UCB}_{t,j}(x^*_{t,j}) - \langle \theta^*, x^*_{t,j} \rangle &\leq \langle \hat{\theta}_{t,j}, x^*_{t,j} \rangle - \langle \theta^*, x^*_{t,j} \rangle + \sqrt{\beta_{t,j}} \|x_{t,j}\|_{V^{-1}_{t,j}} \\
&\leq \|\hat{\theta}_{t,j} - \theta^*\|_{V_{t,j}} \|x_{t,j}\|_{V^{-1}_{t,j}} \\
&\leq \sqrt{\beta_{t,j}} \|x_{t,j}\|_{V^{-1}_{t,j}},
\end{aligned}$$

where the second line follows from Cauchy-Schwarz inequality. Using Lemma E.4, this implies

$$\begin{aligned}
\sum_{t \leq T\colon i_t = j} \mathrm{UCB}_{t,j}(x^*_{t,j}) - \langle \theta^*, x^*_{t,j} \rangle &\leq \sum_{t \leq T\colon i_t = j} \sqrt{\beta_{t,j}} \|x_{t,j}\|_{V^{-1}_{t,j}} \\
&\leq \mathcal{O}\left(d \log(T) \sqrt{n_{\tau_j}(j)} \right).
\end{aligned}$$

Since $\langle \theta^*, x^*_{t,i^*_t} - x^*_{t,j} \rangle$ is constant for $i^*_t \neq j$, this means that

$$\sum_{t=1}^{T} \mathbb{1}\{i^*_t = i, i_t = j\} \leq \mathcal{O}\left(d \log(T) \sqrt{n_{\tau_j}(j)} \right)$$

so that by Jensen's inequality

$$\sum_{t=1}^{T} \mathbb{1}\{i^*_t = i, i_t \neq i\} = \sum_{j \neq i} \sum_{t=1}^{T} \mathbb{1}\{i^*_t = i, i_t = j\} \leq \mathcal{O}\left(d \log(T) \sqrt{KT} \right).$$

In a second step, we show that for any deviating strategy σ_i that is not truthful, the utility of arm i is upper bounded by $n^*_T(i) + \tilde{\mathcal{O}}(d\sqrt{T})$. In what follows, we work on the event that $j \in A_t$ and $\theta^* \in C_{t,j}$ for all $t \in [T]$ and $j \neq i$ and recall that this event has probability at least $1 - 1/T^2$ since the arms are reporting truthfully (see Lemma E.2). Since $j \in A_T$ for all $j \neq i$, we have

$$n_T(i) \leq \sum_{t=1}^{T} \mathbb{1}\{i^*_t = i\} + \sum_{t=1}^{T} \mathbb{1}\{i_t = i, i^*_t \neq i\},$$

and we are tasked with bounding the sum on the right hand side appropriately. Now, similarly to before if $i_t = i$ and $i^*_t \neq i$ (given $i^*_t \in A_t$), then

$$\mathrm{UCB}_{t,i}(x_{t,i}) \geq \mathrm{UCB}_{t,i^*_t}(x_{t,i^*_t}) \geq \langle \theta^*, x^*_{t,i^*_t} \rangle,$$

where we used that $\theta^* \in C_{t,i^*_t}$ since the arm $i^*_t \neq i$ is truthful. Consequently,

$$\mathrm{UCB}_{t,i}(x_{t,i}) - \langle \theta^*, x^*_{t,i} \rangle \geq \langle \theta^*, x^*_{t,i^*_t} - x^*_{t,i} \rangle.$$

Combining Lemma E.3 and Lemma E.4 tells us that

$$\sum_{t \leq \tau_i\colon i_t = i} \mathrm{UCB}_{t,i}(x_{t,i}) - \langle \theta^*, x^*_{t,i} \rangle \leq \mathcal{O}\left(d \log(T) \sqrt{T} \right),$$

where we coarsely upper bounded $n_{\tau_i}(i)$ by T. Since $\langle \theta^*, x^*_{t,i^*_t} - x^*_{t,i} \rangle$ for $i^*_t \neq i$ is positive and constant, it follows that

$$\sum_{t=1}^{T} \mathbb{1}\{i_t = i, i^*_t \neq i\} \leq \mathcal{O}\left(d \log(T) \sqrt{T} \right).$$

In summary, we have shown that

$$\mathbb{E}_{\boldsymbol{\sigma}^*}[n_T(i)] \geq n^*_T(i) - \tilde{\mathcal{O}}\left(d\sqrt{KT} \right) \quad \text{and} \quad \mathbb{E}_{\sigma_i, \boldsymbol{\sigma}^*_{-i}}[n_T(i)] \leq n^*_T(i) + \tilde{\mathcal{O}}\left(d\sqrt{T} \right)$$

for any deviating strategy σ_i. Hence, $\boldsymbol{\sigma}^*$ is a $\tilde{\mathcal{O}}(d\sqrt{KT})$-Nash equilibrium under OptGTM.

Regret analysis. Since we maintain estimates and confidence sets for each arm independently, it is natural to decompose the regret as

$$\sum_{t=1}^{T}\langle\theta^*, x^*_{t,i^*_t} - x^*_{t,i_t}\rangle = \sum_{i=1}^{K}\sum_{t\le T:\ i_t=i}\langle\theta^*, x^*_{t,i^*_t} - x^*_{t,i}\rangle. \tag{13}$$

Note that w.h.p. no truthful arm gets eliminated and $\theta^* \in C_{t,i}$ for all $(t,i) \in [T]\times[K]$. The regret analysis then proceeds similarly to that of LinUCB.

Since $\theta^* \in C_{t,i}$, we know that for any round such that $i_t = i$ that

$$\langle\theta^*, x^*_{t,i^*_t}\rangle \le \mathrm{UCB}_{t,i^*_t}(x^*_{t,i^*_t}) = \mathrm{UCB}_{t,i^*_t}(x_{t,i^*_t}) \le \mathrm{UCB}_{t,i}(x_{t,i}) = \mathrm{UCB}_{t,i}(x^*_{t,i}).$$

Then, again for any round with $i_t = i$, applying Cauchy-Schwarz inequality yields

$$\langle\theta^*, x^*_{t,i^*_t} - x^*_{t,i}\rangle \le \mathrm{UCB}_{t,i}(x^*_{t,i}) - \langle\theta^*, x^*_{t,i}\rangle \le 2\sqrt{\beta_{t,i}}\|x_{t,i}\|_{V^{-1}_{t,i}}. \tag{14}$$

Next, first using Cauchy-Schwarz inequality and the elliptical potential lemma (Lemma E.4), and then Jensen's inequality, it follows that

$$2\sum_{i=1}^{K}\sum_{t\le T:\ i_t=i}\sqrt{\beta_{t,i}}\|x_{t,i}\|_{V^{-1}_{t,i}} \le \mathcal{O}\left(d\log(T)\sqrt{KT}\right). \tag{15}$$

Hence, connecting equations (13)-(15), we obtain $R_T(\text{OptGTM}, \boldsymbol{\sigma}^*) \le \tilde{\mathcal{O}}(d\sqrt{KT})$. We see that the additional $\sqrt{K}$ factor emerges due to OptGTM maintaining independent estimates for each arm. Usually a dependence on the action set size can be prevented since observations from one arm can be used for another arm as well. However, to prevent collusion in the strategic linear contextual bandit it is important to limit the influence an arm has on the selection (and elimination) of other arms. □

E.3 Proof of Theorem 5.2

Proof of Theorem 5.2. We begin the proof of Theorem 5.2 by decomposing the regret into two expression, which we then separately bound.

Decomposing strategic regret. We now decompose the regret of OptGTM into the rounds t where the optimal arm in round t is still active and the rounds where it is not. Like before, let $i^*_t := \mathrm{argmax}_{i\in[K]}\langle\theta^*, x^*_{t,i}\rangle$ be the optimal arm in round t. For any $\boldsymbol{\sigma} \in \mathrm{NE}(\text{OptGTM})$, we have

$$R_T(\boldsymbol{\sigma}) = \underbrace{\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T}\mathbb{1}\{i^*_t \in A_t\}\langle\theta^*, x^*_{t,i^*_t} - x^*_{t,i_t}\rangle\right]}_{J_1} + \underbrace{\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T}\mathbb{1}\{i^*_t \notin A_t\}\langle\theta^*, x^*_{t,i^*_t} - x^*_{t,i_t}\rangle\right]}_{J_2}. \tag{16}$$

With the help of Lemma E.3 and Lemma E.4 we now bound J_1.

Lemma E.5 (Bounding J_1)**.**

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T}\mathbb{1}\{i^*_t \in A_t\}\langle\theta^*, x^*_{t,i^*_t} - x^*_{t,i_t}\rangle\right] \le \mathcal{O}\left(d\sqrt{KT}\log(T)\right).$$

Proof. By design of OptGTM, we we have$\langle\theta^*, x^*_{t,i^*_t}\rangle \le \mathrm{UCB}_{t,i^*_t}(x_{t,i^*_t}) \le \mathrm{UCB}_{t,i_t}(x_{t,i_t})$. Then, on the good event $\mathcal{G}$, Lemma E.3 yields

$$\begin{aligned}\sum_{t\le\tau_i:\ i_t=i}\langle\theta^*, x^*_{t,i^*_t} - x^*_{t,i}\rangle &\le \sum_{t\le\tau_i:\ i_t=i}\left(\mathrm{UCB}_{t,i_t}(x_{t,i_t}) - \langle\theta^*, x^*_{t,i}\rangle\right)\\ &\le 2\left(2\sqrt{n_{\tau_i}(i)\log(T)} + \sum_{t\le\tau_i:\ i_t=i}\sqrt{\beta_{t,i}}\|x_{t,i}\|_{V^{-1}_{t,i}}\right).\end{aligned}$$

 https://doi.org/10.52202/079017-3703

Then, on the good event, we have

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{1}\{i_t^* \in A_t\}\langle \theta^*, x_{t,i_t^*}^* - x_{t,i_t}^*\rangle &= \sum_{i=1}^{K}\sum_{t=1}^{\tau_i} \mathbb{1}\{i_t = i\}\langle \theta^*, x_{t,i_t^*}^* - x_{t,i}^*\rangle \\ &\le \sum_{i=1}^{K} \sum_{t\le\tau_i:\ i_t=i} 2\sqrt{\beta_{t,i}}\|x_{t,i}\|_{V_{t,i}^{-1}} + \sum_{i=1}^{K} 2\sqrt{n_{\tau_i}(i)\log(T)} \\ &\le \sum_{i=1}^{K} \mathcal{O}\left(d\log(T)\sqrt{n_{\tau_i}(i)}\right) + \sum_{i=1}^{K} 2\sqrt{n_{\tau_i}(i)\log(T)} \\ &\le \mathcal{O}\left(d\log(T)\sqrt{KT}\right)\end{aligned}$$

where we applied Jensen's inequality in the last inequality and used that $\sum_{i=1}^{K} n_{\tau_i}(i) \le T$.

□

Lemma E.6 (Bounding J_2)**.**

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* \notin A_t\}\langle \theta^*, x_{t,i_t^*}^* - x_{t,i_t}^*\rangle\right] \le \mathcal{O}\left(dK^2\sqrt{KT}\log(T)\right).$$

Proof. The proof idea is the same as the one for Lemma D.4, which was used to show the regret upper bound of the Greedy Grim Trigger Mechanism (Theorem 4.2, Appendix D). Recall that by assumption $\langle \theta^*, x_{t,i_t^*}^* - x_{t,i_t}^*\rangle \le 1$. We reuse Lemma D.7 from the proof of Theorem 4.2 to get

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* \notin A_t\}\langle \theta^*, x_{t,i_t^*}^* - x_{t,i_t}^*\rangle\right] \le \mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{T} \mathbb{1}\{i_t^* \notin A_t\}\right] = \sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\sigma}}\left[n_T^*(i) - n_{\tau_i}^*(i)\right],$$

where $n_t^*(i) \coloneqq \sum_{s=1}^{t} \mathbb{1}\{i_s^* = i\}$ is the number of rounds up to round t that i is the optimal. To bound the right hand side, we first prove a lower bound on $\mathbb{E}_{\boldsymbol{\sigma}}[n_T(i)]$ for any NE $\boldsymbol{\sigma} \in \mathrm{NE}(\text{OptGTM})$.

Lemma E.7. *Let $\boldsymbol{\sigma} \in \mathrm{NE}(\text{OptGTM})$. It holds that*

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[n_T(i)\right] \ge n_T^*(i) - \mathcal{O}\left(d\log(T)\sqrt{KT}\right).$$

In particular, it holds that $\mathbb{E}_{\boldsymbol{\sigma}}\left[n_t(i)\right] \ge n_t^*(i) - \mathcal{O}\left(d\log(T)\sqrt{KT}\right)$ for $t \in [T]$.

Since $\mathcal{G}$ occurs with probability $1 - 1/T^2$ and $n_T(i) \le T$ by definition, on event $\mathcal{G}$, we have

$$n_T(i) \ge n_T^*(i) - \mathcal{O}\left(d\log(T)\sqrt{KT}\right) \qquad n_t(i) \ge n_t^*(i) - \mathcal{O}\left(d\log(T)\sqrt{KT}\right).$$

Proof. Given that arm i always reports truthfully, i.e., $x_{t,i} = x_{t,i}^*$ for all t, consider the event that $\theta^* \in C_{t,i}$ for all t. Recall that this event has probability at least $1 - 1/T^2$ according to Lemma E.1.

We use the best response property of the Nash equilibrium by comparing against the truthful strategy. To this end, consider the strategy profile $(\sigma_i^*, \sigma_{-i})$ and the event that $\theta^* \in C_{t,i}$ for all t as well as $\mathcal{G}$. We then have that

$$\begin{aligned}n_T(i) &\ge \sum_{t=1}^{T} \mathbb{1}\{i_t^* = i\} - \sum_{t=1}^{T} \mathbb{1}\{i_t^* = i, i_t \ne i\} \\ &= n_T^*(i) - \sum_{j\ne i}\sum_{t=1}^{\tau_i} \mathbb{1}\{i_t^* = i, i_t = j\}, \end{aligned} \tag{17}$$

where the sum on the right hand side is the number of rounds where i is optimal but OptGTM pulls another arm (because it has reported a larger optimistic value).

Next, recall that $\theta^* \in C_{t,i}$ so that for $i_t^* = i$ it follows that

$$\mathrm{UCB}_{t,i}(x_{t,i}) = \mathrm{UCB}_{t,i}(x_{t,i}^*) \geq \langle \theta^*, x_{t,i}^* \rangle = \langle \theta^*, x_{t,i_t^*}^* \rangle.$$

Now, $i_t = j$ implies $\mathrm{UCB}_{t,j}(x_{t,j}) \geq \mathrm{UCB}_{t,i}(x_{t,i}) \geq \langle \theta^*, x_{t,i_t^*}^* \rangle$, since $i \in A_t$ for all t and OptGTM selects the arm with maximal optimistic value. As a result, when $i_t^* = i$ and $i_t = j$, we obtain that

$$\mathrm{UCB}_{t,j}(x_{t,j}) - \langle \theta^*, x_{t,j}^* \rangle \geq \langle \theta^*, x_{t,i_t^*}^* - x_{t,j}^* \rangle.$$

Importantly, we have shown in Lemma E.3 that the total difference on the left hand side is bounded before elimination, i.e., before round τ_j. As a consequence, we get

$$\begin{aligned}
\sum_{t=1}^{\tau_i} \mathbb{1}\{i_t^* = i, i_t = j\} \langle \theta^*, x_{t,i_t^*}^* - x_{t,j}^* \rangle &\leq \sum_{t=1}^{\tau_i} \mathbb{1}\{i_t^* = i, i_t = j\} \left(\mathrm{UCB}_{t,j}(x_{t,j}) - \langle \theta^*, x_{t,j}^* \rangle \right) \\
&\leq \sum_{t=1}^{\tau_i} \mathbb{1}\{i_t = j\} \left(\mathrm{UCB}_{t,j}(x_{t,j}) - \langle \theta^*, x_{t,j}^* \rangle \right) \\
&\leq 2 \sum_{t=1}^{\tau_j} \mathbb{1}\{i_t = j\} \sqrt{\beta_{t,j}} \|x_{t,j}\|_{V_{t,j}^{-1}} + 4\sqrt{n_{\tau_j}(i) \log(T)} \\
&\leq \mathcal{O}\left(d \log(T) \sqrt{n_{\tau_j}(j)} \right) + 4\sqrt{n_{\tau_j}(i) \log(T)} \\
&\leq \mathcal{O}\left(d \log(T) \sqrt{n_{\tau_j}(j)} \right),
\end{aligned}$$

where we first used Lemma E.3 and then Lemma E.4. Recalling that $\langle \theta^*, x_{t,i_t^*}^* - x_{t,j}^* \rangle > 0$ is constant for $j \neq i_t^*$ by assumption of a constant optimality gap, it then follows from Jensen's inequality that

$$\sum_{j \neq i} \sum_{t=1}^{\tau_i} \mathbb{1}\{i_t^* = i, i_t = j\} \leq \sum_{j \neq i} \mathcal{O}\left(d \log(T) \sqrt{n_{\tau_j}(j)} \right) \leq \mathcal{O}\left(d \log(T) \sqrt{KT} \right),$$

where we used that $\sum_{j \neq i} n_{\tau_j}(j) \leq T$. Hence, equation (17) yields

$$n_T(i) \geq n_T^*(i) - \mathcal{O}\left(d \log(T) \sqrt{KT} \right).$$

Since σ_i must be a best response to σ_{-i}, we obtain

$$\mathbb{E}_{\boldsymbol{\sigma}}[n_T(i)] \geq \mathbb{E}_{\sigma_i^*, \sigma_{-i}}[n_T(i)] \geq n_T^*(i) - \mathcal{O}\left(d \log(T) \sqrt{KT} \right).$$

To obtain the result for $t \in [T]$l, suppose that on the good event $\mathcal{G}$ the contrary is true so that $n_t(i) < n_t^*(i) - \omega(d \log(T) \sqrt{KT})$. However, similarly to before, Lemma E.3 tells us that the number of rounds that are "poached" from arm i, i.e., $i_t^* = i$ and $i_t \neq i$, is upper bounded from above by $\mathcal{O}(d \log(T) \sqrt{Kt})$. Hence, since $\mathrm{UCB}_{t,i}(x_{t,i}) \geq \langle \theta^*, x_{t,i_t}^* \rangle$ and $n_t(i) \leq n_t^*(i) - \omega(d \log(T) \sqrt{KT})$, it must hold that $\tau_i < t$. Then, on the good event $\mathcal{G}$, it follows that $n_{\tau_i}(i) \leq n_t(i) < n_t^*(i) - \omega(d \log(T) \sqrt{KT})$. Since event $\mathcal{G}$ has probability at least $1 - 1/T^2$, this implies

$$\mathbb{E}_{\boldsymbol{\sigma}}[n_T(i)] = \mathbb{E}_{\boldsymbol{\sigma}}[n_{\tau_i}(i)] \leq n_t^*(i) - \omega\big(d \log(T) \sqrt{KT}\big).$$

This contradicts the lower bound of $\mathbb{E}_{\boldsymbol{\sigma}}[n_T(i)] \geq n_T^*(i) - \mathcal{O}(d \log(T) \sqrt{KT})$.

□

Lemma E.8. *Let* $\boldsymbol{\sigma} \in \mathrm{NE}(\mathrm{OptGTM})$. *Then,*

$$\mathbb{E}_{\boldsymbol{\sigma}}[n_T(i)] \leq \mathbb{E}_{\boldsymbol{\sigma}}[n_{\tau_i}^*(i)] - \mathcal{O}\left(d \log(T) K \sqrt{KT} \right),$$

where the expectation on the right hand side is taken w.r.t. τ_i.

https://doi.org/10.52202/079017-3703

Proof. We have $\sum_{i=1}^{K} n_\tau^*(i) = \sum_{i=1}^{K} \sum_{t=1}^{\tau} \mathbb{1}\{i_t^* = i\} = \tau$ for any $\tau \in [T]$. Using Lemma E.7, we obtain

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}}[n_T(i)] &= \mathbb{E}_{\boldsymbol{\sigma}}[n_{\tau_i}(i)] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{\tau_i} \mathbb{1}\{i_t = i\}\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{t=1}^{\tau_i} (1 - \mathbb{1}\{i_t \neq i\})\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\tau_i\right] - \sum_{j \neq i} \mathbb{E}_{\boldsymbol{\sigma}}\left[n_{\tau_i}(j)\right] \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}}[\tau_i] - \sum_{j \neq i} \mathbb{E}_{\boldsymbol{\sigma}}[n_{\tau_i}^*(j)] + \mathcal{O}\left(d \log(T) K \sqrt{KT}\right) \\
&= \mathbb{E}_{\boldsymbol{\sigma}}[n_{\tau_i}^*(i)] + \mathcal{O}\left(d \log(T) K \sqrt{KT}\right).
\end{aligned}$$

□

Combing the lower and upper bounds on each arm's utility of Lemma E.7 and Lemma E.8, we get

$$n_T^*(i) - \mathcal{O}\left(d \log(T) \sqrt{KT}\right) \leq \mathbb{E}_{\boldsymbol{\sigma}}[n_T(i)] \leq \mathbb{E}_{\boldsymbol{\sigma}}[n_{\tau_i}^*(i)] - \mathcal{O}\left(d \log(T) K \sqrt{KT}\right). \tag{18}$$

It then follows that

$$\mathbb{E}_{\boldsymbol{\sigma}}[n_T^*(i) - n_{\tau_i}^*(i)] \leq \mathcal{O}\left(d \log(T) K \sqrt{KT}\right)$$

so that

$$\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\sigma}}\left[n_T^*(i) - n_{\tau_i}^*(i)\right] \leq \mathcal{O}\left(d \log(T) K^2 \sqrt{KT}\right).$$

This concludes the proof of Lemma E.6. □

Finally, recalling the regret decomposition from the beginning of the proof and using Lemma E.5 and Lemma E.6, we obtain for any $\boldsymbol{\sigma} \in \mathrm{NE}(\boldsymbol{\sigma})$ that

$$R_T(\mathrm{GGTM}, \boldsymbol{\sigma}) \leq \mathcal{O}\left(\underbrace{d \log(T) \sqrt{KT}}_{\text{Lemma E.5}} + \underbrace{d \log(T) K^2 \sqrt{KT}}_{\text{Lemma E.6}}\right) \leq \tilde{\mathcal{O}}\left(dK^2 \sqrt{KT}\right).$$

□

E.4 Linear Realizability of Reported Contexts

In the following, we comment on an interesting observation in the strategic linear contextual bandit that may also provide some insight into the effectiveness of OptGTM. Suppose that each arm reports its contexts in a linearly realizable fashion (without us explicitly incentivizing them to do so). Formally, we can express this as the following assumption.

Assumption 3 (Linear Realizability of Reported Contexts)**.** Every arm reports so that its reports follow some linear reward model. That is, for every arm $i \in [K]$, there exists a vector $\theta_i^* \in \mathbb{R}^d$ such that for all $t \in [T]$

$$\langle \theta^*, x_{t,i}^* \rangle = \langle \theta_i^*, x_{t,i} \rangle. \tag{19}$$

Perhaps surprisingly, the regret analysis of OptGTM becomes straightforward when the arms' strategies satisfy Assumption 3. Moreover, we can prove that OptGTM suffers $\tilde{\mathcal{O}}\big(d\sqrt{KT}\big)$ strategic regret in every Nash equilibrium of the arms. That is, the regret guarantee is better than that of Theorem 5.2.

A quick regret analysis. Let $\boldsymbol{\sigma}$ be any NE under OptGTM. When we observe a reward $r_{t,i}$ after pulling arm i in round t, we can interpret the reward as $r_{t,i} \coloneqq \langle \theta^*, x^*_{t,i} \rangle + \eta_t = \langle \theta^*_i, x_{t,i} \rangle + \eta_t$. Hence, isolating arm i, the learner is essentially playing a linear contextual bandit with true unknown parameter θ^*_i, contexts $x_{t,i}$, and rewards $r_{t,i} = \langle \theta^*_i, x_{t,i} \rangle + \eta_t$. As a result, the independent estimators $\hat{\theta}_{t,i}$ for every arm i, are in fact estimating θ^*_i and, according to Lemma E.1, with high probability $\theta^*_i \in C_{t,i}$. It is then also easy to see that OptGTM will never eliminate any of the arms with high probability. Now, since $\theta^*_i \in C_{t,i}$,

$$\langle \theta^*, x^*_{t,i} \rangle = \langle \theta^*_i, x_{t,i} \rangle \leq \mathrm{UCB}_{t,i}(x_{t,i}).$$

As a result, using Cauchy Schwarz inequality, we obtain

$$\begin{aligned}
\mathrm{UCB}_{t,i}(x_{t,i}) - \langle \theta^*, x^*_{t,i} \rangle &\leq \langle \hat{\theta}_{t,i} - \theta^*_i, x_{t,i} \rangle + \sqrt{\beta_{t,i}} \|x_{t,i}\|_{V^{-1}_{t,i}} \\
&\leq \|\hat{\theta}_{t,i} - \theta^*_i\|_{V_{t,i}} \|x_{t,i}\|_{V^{-1}_{t,i}} + \sqrt{\beta_{t,i}} \|x_{t,i}\|_{V^{-1}_{t,i}} \leq 2\sqrt{\beta_{t,i}} \|x_{t,i}\|_{V^{-1}_{t,i}}.
\end{aligned}$$

In every round t, OptGTM selects the arm with maximal optimistic reported value $\mathrm{UCB}_{t,i}(x_{t,i})$ so that $\mathrm{UCB}_{t,i^*_t}(x_{t,i^*_t}) - \mathrm{UCB}_{t,i_t}(x_{t,i_t}) \leq 0$. We can then bound the instantaneous regret in round t as

$$\begin{aligned}
\langle \theta^*, x^*_{t,i^*_t} - x^*_{t,i_t} \rangle &\leq \mathrm{UCB}_{t,i^*_t}(x_{t,i^*_t}) - \langle \theta^*, x^*_{t,i_t} \rangle \\
&\leq \mathrm{UCB}_{t,i^*_t}(x_{t,i^*_t}) - \mathrm{UCB}_{t,i_t}(x_{t,i_t}) + 2\sqrt{\beta_{t,i_t}} \|x_{t,i_t}\|_{V^{-1}_{t,i_t}} \\
&\leq 2\sqrt{\beta_{t,i_t}} \|x_{t,i_t}\|_{V^{-1}_{t,i_t}}.
\end{aligned}$$

Using the elliptical potential lemma and Jensen's inequality (Lemma E.4, [1, 24]), the total regret of OptGTM is given by

$$R_T(\text{OptGTM}, \boldsymbol{\sigma}) \leq \sum_{i=1}^{K} \sum_{t \leq T:\ i_t = i} 2\sqrt{\beta_{t,i}} \|x_{t,i}\|_{V^{-1}_{t,i}} \leq \mathcal{O}\left(d \log(T) \sqrt{KT}\right).$$

We have thus shown the following guarantee.

Corollary E.9. *Suppose that Assumption 3 holds. Then,*

$$R_T(\text{OptGTM}, \boldsymbol{\sigma}) = \tilde{\mathcal{O}}\left(d\sqrt{KT}\right)$$

for every Nash equilibrium $\boldsymbol{\sigma} \in \mathrm{NE}(\text{OptGTM})$.

https://doi.org/10.52202/079017-3703

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract, we list the main claims of this paper in a general fashion. Then, in the introduction we state them in more detail. They accurately reflect the paper's contribution and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations and assumptions of our work throughout the paper. Additional limitations are highlighted in the discussion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions can be found in Section 3 and in the paragraphs before the theorems. We provide proof sketches in the main text and complete proofs in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

https://doi.org/10.52202/079017-3703

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

https://doi.org/10.52202/079017-3703

Answer: [Yes]

Justification: We do not see any potential negative social impact of this work. On the contrary, this work aims to improve the fairness and equity of online platforms by discouraging harmful gaming behavior in response to recommendation algorithms.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This work addresses the problem of aligning agent incentives with that of learning algorithms. While there are some potential positive societal impacts of this theoretical work, we do not believe that it is necessary to highlight them here.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

 https://doi.org/10.52202/079017-3703

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.