The Intrinsic Robustness of Stochastic Bandits to Strategic Manipulation

Zhe Feng [1] **David C. Parkes** [1] **Haifeng Xu** [2]

Abstract

Motivated by economic applications such as recommender systems, we study the behavior of stochastic bandits algorithms under *strategic behavior* conducted by rational actors, i.e., the arms. Each arm is a *self-interested* strategic player who can modify its own reward whenever pulled, subject to a cross-period budget constraint, in order to maximize its own expected number of times of being pulled. We analyze the robustness of three popular bandit algorithms: UCB, ε-Greedy, and Thompson Sampling. We prove that all three algorithms achieve a regret upper bound $\mathcal{O}(\max\{B, K\ln T\})$ where B is the total budget across arms, K is the total number of arms and T is length of the time horizon. This regret guarantee holds under *arbitrary adaptive* manipulation strategy of arms. Our second set of main results shows that this regret bound is *tight*— in fact for UCB it is tight even when we restrict the arms' manipulation strategies to form a *Nash equilibrium*. The lower bound makes use of a simple manipulation strategy, the same for all three algorithms, yielding a bound of $\Omega(\max\{B, K\ln T\})$. Our results illustrate the robustness of classic bandits algorithms against strategic manipulations as long as $B = o(T)$.

1. Introduction

Multi-armed bandits (MAB) algorithms play a significant role in learning to make decisions across the digital economy, for example in online advertising (Chapelle et al., 2014; Feng et al., 2019), search engines (Kveton et al., 2015), and recommender systems (Li et al., 2010). Classical stochastic MAB models assume that the reward feedback of each arm is drawn from a fixed distribution. However, in many economic applications, an arm may be *strategic* and able to modulate its own reward feedback in order to further its own objective, e.g., increasing the number of times it is selected. For instance, restaurants may offer discounts or free dishes in order to entice customers to return, and sellers on Amazon may offer discounts or coupons in order to receive higher ratings and thus increase their ranking.

We distinguish two different kinds of actors in our strategic setting: the *principal* and the *arms*. The principal represents a multi-armed bandit algorithm, corresponding to a system, such as the Amazon marketplace platform. The arms represent the parties who generate reward feedback to the principal, for example the sellers on Amazon. We assume that the *true reward* of each arm is drawn from an underlying distribution. Further, we model each arm i as a strategic agent, able to manipulate its own reward, but subject to a total budget B_i across all time periods. The objective of an arm is to maximize its expected number of times being pulled. Arms can only modify their own reward feedback, and have no control over the rewards of the other arms. An arm's strategy can be adaptive— that is, the amount by which an arm modulates the current reward can depend on his own history of realized rewards and manipulations. Since arms' strategies affect each other, through the MAB algorithm, this dynamic interaction forms a situation of strategic interdependence among arms, more precisely, a *stochastic game*.

This study is motivated by various economic applications of MAB, where strategic manipulations appear more realistic than the more conservative consideration of *adversarial attacks* (Jun et al., 2018; Lykouris et al., 2018). The central question that we study in this paper is the following:

> *Are existing stochastic bandit algorithms robust to strategic manipulation by arms? Quantitatively, can we characterize their regret bounds?*

For a motivating example, suppose that a recommender system such as Yelp runs a stochastic bandit algorithm to recommend a single restaurant to each user. The arms correspond to restaurants to be recommended and each user access to the system corresponds to a pull of the arms. The true service quality of each restaurant follows some underlying distribution. However, restaurants are strategic, and a natural objective is to maximize the expected number of times a restaurant is recommended to users. To do so, it

[1]Harvard University, Cambridge, Massachusetts, USA [2]University of Virginia, Charlottesville, Virginia, USA. Correspondence to: Zhe Feng <zhe_feng@harvard.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

is common to provide discounts to some user (modified rewards in our model), subject to budget constraints because the restaurants cannot provide arbitrarily many discounts. In this context, our goal is to understand how the strategic behavior of restaurants can affect the platform's regret.

1.1. Our Results and Implications

Results. Our main results illustrate that the three popular stochastic bandits algorithms of Upper Confidence Bound (UCB), ϵ-Greedy, and Thompson Sampling, are robust to strategic manipulations. Specifically, we show that the regret of all three algorithms is upper bounded by $\mathcal{O}\big(\sum_{i\neq i^*} \max\{B_i, \frac{\ln T}{\Delta_i}\}\big)$, where i^* indexes the optimal arm w.r.t. the true rewards, and Δ_i is the difference in the mean of the true reward between arm i and i^*. For convenience, we assume throughout the paper that $B_{i^*} = 0$, since any $B_{i^*} > 0$ would only help i^* to be pulled more, and thus benefit the principal. Interestingly, the regret bound holds for arbitrary adaptive arm strategies.

One natural question is whether it is possible to achieve smaller regret bounds if we restrict strategies to form a *Nash equilibrium*, which is the standard solution concept in game theory. We answer this question in the negative, at least for UCB. We characterize the dominant-strategy equilibrium of the game induced by the UCB algorithm, and prove a lower bound on regret of $\Omega(\max\{B, K \ln T\})$ for equilibrium arm manipulations, where K is the number of arms and B is the total budget across arms. This shows that the upper bound is essentially tight, even under equilibrium behaviors. All our bounds hold for both bounded and unbounded rewards. We also provide a matching lower bound for ε-Greedy and TS under a natural, lump sum investing strategy, in which an arm spends all of its budget the first time it is pulled. We have not been able to show whether or not this strategy forms a Nash equilibrium in the induced stochastic game, and leave open the question of whether the regret bound is also tight for ε-Greedy and Thompson sampling (TS) under equilibrium behavior.

Implications. These results show that the performances of all three MAB algorithms deteriorates linearly in the total budget $B = \sum_{i\neq i^*} B_i$. As long as $B = o(T)$, the optimal arm will be pulled $T - o(T)$ times. The simulation results also validate this linear dependence on B.

Since our upper bounds on regret hold for arbitrary arm behaviors, even allowing for reducing the reward on arms, they can also correspond to the choices of a single adversary, and the results also shed light on *adversarial attacks* on stochastic bandit algorithms. In contrast to existing adversarial models, the key difference is that the reward of the optimal arm, i^*, cannot be modified. With rational behavior, this is without loss; if the optimal arm had an associated budget then this can only lead to more pulls of this arm and lower regret. Our results show that if a single adversary cannot contaminate the optimal arm, then standard bandits algorithms are already robust. The bound would also hold in a more general setting in which the optimal arm's reward can only be increased.

Concretely, the results can be alternatively interpreted as follows: for an adversarial corruption model that is modified to prevent contamination of the optimal arm, then UCB, ε-Greedy, and TS all have regret $O(\max\{B, K \ln T\})$, and are robust as long as $B = o(T)$. This is in sharp contrast to the situation of *unrestricted adversarial attacks*, where an attack budget of $\mathcal{O}(\ln T)$ can lead algorithms such as UCB and ε-Greedy to suffer regret $\Omega(T)$ (Jun et al., 2018; Lykouris et al., 2018). Even for state-of-the-art, robust bandits algorithms (Gupta et al., 2019), the regret bound $O(KB + \sum_{i\neq i^*} \frac{\log T}{\Delta_i} \log(\frac{K}{\delta} \log T))$ is worse than the bound in the present paper by a factor of K (when $B = \Omega(\log T)$).

Another implication of the present work is to the problem of *incentivizing exploration*, where the principal relies on users to pull arms (Frazier et al., 2014; Wang & Huang, 2018), and users are modeled as myopic and only care about their immediate reward. The idea is that the principal can provide rewards to encourage more exploration. At the same time, it has been observed in field experiments that users are generally biased towards reporting a higher evaluation when provided with these kinds of incentives, i.e., an upwards-biased reward. Our results have been applied by Liu et al. (2020) to show that bandit algorithms are robust to this kind of bias: if reported rewards can only be upwards-biased (a special case of our model), then the bandit algorithm will be robust, also allowing for the reward feedback on the optimal arm to be affected.

1.2. Additional Related Work

In this work, we study strategic manipulation in the context of classical stochastic bandit algorithms. This is similar in spirit to Jun et al. (2018), who study adversarial attacks to UCB and ϵ-Greedy. The relation and differences between their results and ours are elaborated above. Another related, and complementary, line of research is on designing new algorithms for stochastic bandits that are robust to adversarial corruptions (Lykouris et al., 2018; Gupta et al., 2019). In principle, we could have also studied these algorithms in the present context. However, we believe that it remains important to understand the conditions under which classical, simple bandit algorithms work well, because they are likely to be used in real-world applications. Moreover, the regret guarantees of these classical algorithms, in our strategic setup, is better than the bounds available for these robust algorithms under adversarial corruptions. It is an interesting open question to understand whether these robust algorithms

can achieve the same or even better regret bound when restricted to our strategic setup. Another further work is to understand strategic behavios in the recent line of works in non-stationary bandits, e.g., Besbes et al. (2019); Cheung et al. (2019).

This work belongs to the general field of *no-regret learning with strategic agents*. Much of this literature is focused on designing no-regret learning algorithms under strategic behavior, and has studied problems arising from concrete applications such as auctions, e.g., (Blum et al., 2004; Weed et al., 2016; Feldman et al., 2016; Feng et al., 2018) and recommender systems (Mansour et al., 2015; Immorlica et al., 2019). However, the strategic behavior in these models do not correspond to arm manipulation, but rather correspond to bidding strategies or auction mechanisms. To our knowledge, Braverman et al. (2019) are the first to consider strategic behaviors of arms in stochastic bandit settings. In their model, when an arm is pulled, it receives a private reward v and strategically chooses an amount x to pass to the principal, leaving the remaining amount of $v - x$ to the arm itself. Motivated by a different application context, our model considers strategic arms that seek to maximize their expected number of plays by manipulating their reward feedback under a budget.

2. The Model: Strategic Manipulations in Stochastic Bandits

We consider a strategic variant of the stochastic multi-armed bandit problem. There are K arms, denoted by $[K] = \{1, 2, ..., K\}$. The reward of each arm $i \in [K]$ follows a σ-sub-Gaussian distribution (see Definition A.1 in Appendix) with mean μ_i, where parameter σ is publicly known. The σ-sub-Gaussian assumption is widely used in MAB literature (Bubeck & Cesa-Bianchi, 2012). Let $i^* = \arg\max_{i\in[K]} \mu_i$ denote the unique arm (WLOG) with maximum mean, $\Delta_i = \mu_{i^*} - \mu_i > 0$ denote the difference of the reward mean between the optimal arm i^* and arm i ($\neq i^*$), and $\underline{\Delta} := \min_{i\neq i^*} \Delta_i$.

There are two different parties: the principal and the arms. The principal represents a bandit algorithm, in particular, UCB, ϵ-Greedy, or TS. At each time $t = 1, \cdots, T$, the principal pulls arm I_t, which generates a reward r_t. Here T is some fixed time horizon. Let $n_i(t) = \sum_{\tau=1}^{t} \mathbb{I}(I_t = i)$ denote the number of times that arm i has been pulled up to and including time t, and $\widehat{\mu}_i(t) = \frac{1}{n_i(t)} \sum_{\tau=1}^{t} r_\tau \cdot \mathbb{I}(I_\tau = i)$ denote the average rewards obtained from pulling arm i up to and including time t.

Each arm $i \in [K]$ is a strategic actor, equipped with the objective of maximizing $\mathbb{E}[n_i(T)]$, i.e., the expected total number of times it is pulled. This is a natural objective in systems such as recommender systems.

The actions available to arm i is to modify its reward feedback when pulled, subject to a total budget B_i across rounds. Concretely, when $I_t = i$, arm i can add an additional reward amount $\alpha_t^{(i)}$ to the realized reward r_t,[1] subject to budget constraint $\sum_{t=1}^{T} |\alpha_t^{(i)}| \leq B_i$, so that the revealed reward to the principal is $\widetilde{r}_t = r_t + \alpha_t^{(i)}$. We refer to r_t as the *true reward* and $\widetilde{r}_t$ the *manipulated reward*. The *adaptive* manipulation strategy of arm i is a function $S^{(i)} : \mathcal{H}_{t-1}^{(i)} \times [K] \to \mathbb{R}$, mapping its own up-to-$t$ history $h_{t-1}^{(i)} \in \mathcal{H}_{t-1}^{(i)}$ and I_t to a manipulation $\alpha_t^{(i)}$. The history $h_t^{(i)} = \{I_\tau, r_\tau, \alpha_\tau^{(i)}\}_{\tau: I_\tau = i, \tau \leq t}$ is the information that arm i observed up to time t, which includes the pulling history, realized rewards, and manipulations of arm i at past t rounds. Let $h_t = \{h_t^{(i)}\}_{i\in[n]}$ denote the histories of all arms until time t. Arm i has no access to the information of the other arms, hence the strategy only takes his own historical information as input. We use $S^{(-i)}$ to define the strategies of the other arms. Given a history $h_t^{(i)}$, the remaining budget and $n_i(t)$ are determined.

Arm i has no control over other arms' rewards. Therefore, $\alpha_t^{(i)}$ must equal 0 for $I_t \neq i$ and any history $h_{t-1}^{(i)}$. For convenience of the analysis, we assume $B_{i^*} = 0$ throughout the paper and thus $\alpha_t^{(i^*)} = 0$ for any t, since any reasonable $\mathcal{S}^{(i^*)}$ with $B_{i^*} > 0$ would only lead to more pulls of i^* and thus benefit the principal. Let

$$\beta_t^{(i)}(h_{t-1}^{(i)}, I_t) = \sum_{\tau \leq t} S^{(i)}(h_{\tau-1}^{(i)}, I_\tau)$$

denote the total manipulation by arm i until time t with manipulation strategy $S^{(i)}$ and a realized history $h^{(i)}$, which satisfies $\beta_T^{(i)}(h, I_T) \leq B_i, \forall h \in \mathcal{H}_{T-1}^{(i)}$ and $I_T \in [K]$. When the history $h_{t-1}^{(i)}$ and selected arm I_t are clear from the context, we sometimes omit this and write $\beta_t^{(i)}$ for notational convenience.

The objective of arm i is to find a strategy $S^{(i)}$ to maximize $\mathbb{E}[\sum_{t=1}^{T} \mathbb{I}\{I_t = i\}]$,[2] by manipulating its reward to trick the principal to pull arm i more. The principal observes only $\widetilde{r}_t$ and not true reward r_t. The goal of the principal is to minimize regret with respect to the true reward r_t. This is without loss of generality since the aggregated reward with respect to $\widetilde{r}_t$ differs from the true reward by at most the total manipulation budget $B = \sum_i B_i$, which is the same order as our regret bounds.

`LSI` manipulation. A particular manipulation strategy that will be of interest is the *Lump Sum Investing* (`LSI`) strategy, in which an arm simply spends all of its remaining budget whenever first pulled. For arm i, the `LSI` is a strat-

[1] In this paper, $\alpha_t^{(i)}$ can be negative, if that helps i. None of our results rely on the positivity of $\alpha_t^{(i)}$'s.

[2] Throughout the paper, the expectation is over all the randomness in algorithms and the rewards.

egy $S^{(i)}$ that at any time t and any history $h_{t-1}^{(i)} \in \mathcal{H}_{t-1}^{(i)}$, $S^{(i)}(h_{t-1}^{(i)}, I_t) = B_i - \sum_{\tau=1}^{t-1} \alpha_\tau^{(i)}$ when $I_t = i$.

2.1. Solution Concepts

This is a situation of strategic interaction, where the MAB algorithms induce a stochastic game. Our main goal is to quantify the principal's regret in this game, as measured with respect to the true reward. Despite the widely-known intractability in characterizing Nash Equilibria for general stochastic games (Ben-Porath, 1990; Conitzer & Sandholm, 2003), we show that when the principal runs UCB, there is a *subgame perfect Nash equilibrium* (SPE) in our game, where each arm simply plays the LSI strategy. A strategy profile $S^* = (S^{*(1)}, \cdots, S^{*(K)})$ is a SPE if $S^{*(i)}$ is an optimal strategy for any arm i, given any history h_{t-1}, and given the strategies $S^{(-i)}$ of the other arms, for any t. In fact, we show that LSI is *dominant strategy* when the principal runs the UCB algorithm, that is LSI is an optimal strategy for arm i for any t, given any history h_{t-1}, and whatever the strategies $S^{(-i)}$ of the other arms. This provides a very strong suggestion as to the kind of behavior we should expect from arms. The upper bounds on regret hold for arbitrary adaptive manipulations, regardless whether they form a SPE or not. The matching lower bounds on regret for UCB are proved under the dominant-strategy SPE. Not only does this show that the upper bounds are tight, but it highlights the special role of the SPE in this UCB setting.

3. UCB is Robust to Strategic Manipulations

In this section, we provide a regret analysis for the Upper Confidence Bound (UCB) principal in our strategic setup. We first show an upper bound on the regret for arbitrary arm strategies. Next, we prove that this regret bound is tight even under equilibrium arm behaviors. Finally, we discuss how to generalize the results to the bounded reward setting. The formal proofs can be found in Appendix B.

3.1. Regret Upper Bound for UCB Principal

We consider a standard $(\alpha, \psi)-$UCB with $\alpha = 4.5$, $\psi : \lambda \to \frac{\sigma^2\lambda^2}{2}$ and thus $(\psi^*)^{-1}(\epsilon) = \sqrt{2\sigma^2\epsilon}$ (Bubeck & Cesa-Bianchi, 2012) . Concretely, the algorithm selects each arm once in the first K rounds, i.e. $I_t = t, \forall t < K$. For $t \geq K$,

$$I_t = \arg\max_i \left\{ \widehat{\mu}_i(t-1) + 3\sigma\sqrt{\frac{\ln T}{n_i(t-1)}} + \frac{\beta_{t-1}^{(i)}}{n_i(t-1)} \right\},$$

where $\beta_{t-1}^{(i)}$ is the aggregated manipulation of arm i up to (including) $t-1$. The term $\widehat{\mu}_i(t-1) + 3\sigma\sqrt{\frac{\ln T}{n_i(t-1)}}$ is the standard *UCB term*[3] for any arm $i \in [K]$ at time t, which we denote as $\mathrm{UCB}_i(t)$. Let $\widetilde{\mathrm{UCB}}_i(t) = \mathrm{UCB}_i(t) + \beta_{t-1}^{(i)}/n_i(t-1)$ represent the *modified UCB term* for the strategic arm $i\,(i \neq i^*)$ with manipulation strategy $S^{(i)}$ (recall $\beta_t^{(i)}$ is induced by $S^{(i)}$, and $\beta_t^{(i^*)} = 0$ always).

The main result in this section is an upper bound for regret $\mathbb{E}[R(T)]$ under an arbitrary adaptive manipulation strategy S.

Theorem 3.1. *For any manipulation strategy S of the strategic arms, the regret of the UCB principal is bounded by*

$$\mathbb{E}[R(T)] \leq \sum_{i \neq i^*} \left[\max\left\{ 3B_i, \frac{81\sigma^2 \ln T}{\Delta_i} \right\} + \left(1 + 3\Delta_i\right] \right.$$

Theorem 3.1 reveals that the UCB algorithm is robust in our strategic model of arm manipulations. If the budget of each arm is bounded by $\mathcal{O}(\ln T)$, the regret of the principal is still bounded by $\mathcal{O}(\ln T)$. If $B_i = \Omega(\ln T)$ for some arm i's, the regret is upper bounded by $\mathcal{O}(\sum_{i \neq i^*} B_i)$. This is sublinear in T as long as $B = \sum_{i \neq i^*} B_i = o(T)$.

Theorem 3.1 strictly generalizes the regret bound of the standard UCB framework, which corresponds to a special case with no budgets. Fixing any manipulation strategy S, the proof starts by re-writing the regret in the following format:

$$\mathbb{E}[R(T)] = \sum_{i \neq i^*} \Delta_i \cdot \mathbb{E}[n_i^S(T)]. \tag{1}$$

What remains is to bound $\mathbb{E}[n_i^S(T)]$ for each arm i. For convenience, we omit the superscript S since it is clear that we focus on an arbitrary S. Lemma 3.2 gives the upper bound of $\mathbb{E}[n_i(T)]$ for each arm i, and combined with (1), yields a proof of Theorem 3.1.

Lemma 3.2. *Suppose the principal runs UCB. For any manipulation strategy S of strategic arms, the expected number of times that arm $i (i \neq i^*)$ is pulled up to time T can be bounded as follows,*

$$\mathbb{E}[n_i(T)] \leq \max\left\{ \frac{3B_i}{\Delta_i}, \frac{81\sigma^2 \ln T}{\Delta_i^2} \right\} + 3$$

Proof Sketch. The main difference from the analysis of the standard UCB is to choose a proper threshold $C_i(T)$ for $n_i(t-1)$ so that we can have the best trade-off between the two terms in the following decomposition of $\mathbb{E}[n_i(T)]$:

$$\mathbb{E}\left[n_i(T)\right] \leq 1 + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\{I_t = i, n_i(t-1) \leq C_i(T)\}\right] + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\{I_t = i, n_i(t-1) \geq C_i(T)\}\right].$$

[3] There is also a UCB variant that uses time-dependent confidence width $3\sigma\sqrt{\frac{\ln t}{n_i(t-1)}}$. Both versions are common in the literature. Our regret upper bound holds for both, but it appears that the $(\ln T)$ version is more convenient for the analysis of lower bounds in equilibrium.

After careful manipulation, it turns out that $C_i(T) = \max\left\{\frac{81\sigma^2 \ln T}{\Delta_i^2}, \frac{3B_i}{\Delta_i}\right\}$ gives the correct regret bound, after bounding the first term directly by $C_i(T)$ and bounding the second term via the Chernoff-Hoeffding inequality. The formal proof is shown in Appendix B.1. □

3.2. Tightness of the Regret Bounds at Equilibrium

The above regret bound for UCB holds for arbitrary adaptive manipulation strategies. This raises the following question: *is it possible to achieve better regret upper bounds by restricting arm manipulations to form a subgame perfect Nash equilibrium?* We provide a negative answer to this question, and prove that the regret upper bounds are tight even in equilibrium. We first prove that `LSI` is a *dominant strategy* for each arm in any subgame — an optimal strategy regardless of what strategies other arms use, given any realized history h_{t-1} — when the principal runs UCB. As a consequence, each arm playing `LSI` forms a dominant-strategy SPE. We then establish a lower bound on regret when each arm plays the `LSI` strategy, and show that this bound matches the upper bound.

Concretely, we first prove that the (random) number of times that arm i is pulled under strategy `LSI` first-order *stochastically dominates* the number of times pulled under any other adaptive manipulation strategy $S^{(i)}$, given any fixed history.

Theorem 3.3. *Suppose $T \geq K$, and the principal runs the UCB algorithm. For any arm i, any strategy $S^{(i)}$, and any strategy profile $S^{(-i)}$ of others, and for any time t and history $h_{t-1}^{(i)}$, we have*

$$\mathbb{P}[n_i^{(\texttt{LSI},S^{(-i)})}(t:T) \geq n] \geq \mathbb{P}[n_i^{S}(t:T) \geq n],\ \forall n \in \mathbb{N}, \tag{2}$$

where $n_i(t:T) = \sum_{\tau=t}^{T} \mathbb{I}\{I_\tau = i\}$ is the total number of pulls of arm i from t to T. That is, $n_i^{(\texttt{LSI},S^{(-i)})}(t:T)$ first-order stochastically dominates $n_i^S(t:T)$. Therefore, $\mathbb{E}[n_i^{(\texttt{LSI},S^{(-i)})}(t:T)] \geq \mathbb{E}[n_i^S(t:T)]$, and thus `LSI` is a best response to any $S^{(-i)}$.

It follows directly from Theorem 3.3 that each arm playing `LSI` forms a dominant-strategy SPE. The complete proof of Theorem 3.3 is quite involved, and can be found in Appendix B.2.

To see why this conclusion is not obvious, let us illustrate the trade-off in designing the optimal manipulation strategy. The advantage of the `LSI` strategy in UCB is to significantly increase the arm's UCB term and receive many pulls at the very beginning. This, however, also comes with a disadvantage— it quickly decreases the confidence width (the $3\delta\sqrt{\ln T}/\sqrt{n_i(t-1)}$ term) and the effect of the manipulation (the $\beta_{t-1}^i/n_i(t-1)$ term) in the UCB term, whereas other arms' confidence width and manipulation effect remain large. For this reason, it may also be beneficial for an arm to defer its manipulation to later rounds so that it avoids fierce competition in the early few rounds resulting from other arms' large confidence width, large manipulation effect, and possibly large rewards due to lucky draws.

The proof shows that in this intricate random process, the aforementioned advantage of using `LSI` always dominates its disadvantage. We make use of the *coupling technique* (Thorisson, 2000) to compare the random sequence of pulled arms when arm i uses `LSI` compared with an arbitrary strategy $S^{(i)}$. A crucial step is to show that under coupling of the two stochastic processes, either `LSI` results in more pulls of arm i than $S^{(i)}$ or they must result in each of the other arms to be pulled for the same number of times. We then argue that in the latter case, `LSI` must also be better than $S^{(i)}$ because they face the same outside competition but the modified UCB term of `LSI` is larger than the modified UCB term of $S^{(i)}$. As a consequence, `LSI` performs better than $S^{(i)}$ in both cases, yielding a proof of the theorem.

To show that the regret bounds in Section 3.1 are tight, it will suffice to develop a lower bound on regret for when each arm plays `LSI` , as shown in the following theorem.

Theorem 3.4 (Regret Lower Bound at Equilibrium)**.** *Suppose the principal uses UCB algorithm and each arm uses* `LSI`*. For any σ-sub-Gaussian reward distributions on arms, the regret of the principal satisfies,*

$$\mathbb{E}[R(T)] \geq \underline{\Delta} \sum_{i \neq i^*} \frac{B_i}{2\Delta_i} - \mathcal{O}\left(\frac{\ln T}{\underline{\Delta}}\right).$$

The proof of Theorem 3.4 differs from standard techniques in proving regret lower bounds, and is carefully tailored to achieve tight bounds with respect to budget B_i's. Classical regret lower bounds are typically proved by constructing a particular class of distributions, i.e., Bernoulli (Bubeck & Cesa-Bianchi, 2012), and then arguing that the given algorithm cannot do very well on these constructed instances. These bounds are usually distribution-dependent. Our proof takes a completely different route. Indeed, our technique results in a lower bound that holds for arbitrary σ-sub-Gaussian distributions and thus is distribution-independent.

The proof of Theorem 3.4 starts with a simple lower bound for the regret $\mathbb{E}[R(T)]$ by utilizing Equation (1):

$$\mathbb{E}[R(T)] = \sum_{i \neq i^*} \Delta_i \mathbb{E}[n_i(T)] \geq \underline{\Delta} \cdot \sum_{i \neq i^*} \mathbb{E}[n_i(T)]. \tag{3}$$

We then only need to focus on lower bounding $\sum_{i \neq i^*} \mathbb{E}[n_i(T)]$ when all the arms play strategy `LSI`. We prove an upper bound for $\mathbb{E}[n_{i^*}(T)]$, which translates to a lower bound for $\sum_{i \neq i^*} \mathbb{E}[n_i(T)]$. However, upper bounding $\mathbb{E}[n_{i^*}(T)]$ requires quite different techniques than upper

bounding $\mathbb{E}[n_i(T)]$ for any non-optimal arm i. A crucial step is to argue that when i^* has been pulled more than C times (for some carefully chosen threshold C), it will become much less likely to be pulled again. This differs from standard techniques for upper bounding $\mathbb{E}[n_i(T)]$ for non-optimal arm i, for two reasons: (1) we have to compare the UCB term of arm i^* with all the other non-optimal arms' UCB terms, whereas to upper bound $\mathbb{E}[n_i(T)]$, one typically compares i with only the optimal arm i^*; (2) we need to argue i^* is pulled with small probability despite $\mu_{i^*} > \mu_i$ whereas upper bounding $\mathbb{E}[n_i]$ is more natural when $\mu_{i^*} > \mu_i$. To overcome these challenges, we carefully decompose the $\mathbb{E}[n_{i^*}(T)]$ term and pick thresholds not only for $n_{i^*}(t-1)$, but also for $n_i(t-1)$ for each non-optimal arm $i \neq i^*$. A complete proof of Theorem 3.4 can be found in Appendix B.3.

Remarks: The lower bound holds for arbitrary σ-Gaussian distributions, and may be negative in value, and thus not meaningful when $B_i = o(\ln T)$. However, the bound can be easily converted to a distribution-dependent lower bound $\max\left\{\underline{\Delta}\sum_{i\neq i^*}\frac{B_i}{2\Delta_i} - \mathcal{O}\left(\frac{\ln T}{\underline{\Delta}}\right), \Omega\left(K\ln T\right)\right\}$ because there exist distributions such that any no-regret learning algorithm will suffer regret $\Omega\left(K\ln T\right)$ (Bubeck & Cesa-Bianchi, 2012) and the non-optimal arms' manipulation strategy would only increase the regret. This distribution-dependent lower bound precisely matches the upper bound $\mathcal{O}\big(\max\{B, K\ln T\}\big)$ in Section 3.1.

3.3. Generalization to Bounded Rewards

In many applications, such as where the rewards are ratings provided by customers on platforms such as those operated by Yelp and Amazon, the rewards are bounded within some known interval (e.g. $0 \sim 5$ stars rating). Suppose, for example, that the reward is bounded within $[0, 1]$. In such settings, the LSI strategy may be infeasible since the strategic arm can increase its reward to at most the upper bound. In this case, arms can use a natural variant of LSI for bounded rewards: each arm i spends its budget to promote the realized reward to the maximum limit of 1 whenever it is pulled, and does so until it runs out of budget B_i. We term this natural variant the *Lump Sum Investment for Bounded Rewards strategy*, or LSIBR for short.

Theorem 3.3 can be easily generalized to this bounded reward setting. Each arm playing LSIBR forms a dominant-strategy subgame perfect Nash equilibrium in the bounded reward setting. The more challenging task is to prove a similar lower bound on regret. To do so, we provide a unified reduction from any regret lower bound under LSI to a regret lower bound under LSIBR, with an additional loss of $\Theta(\ln T)$. Our reduction applies to any stochastic bandit algorithms.The main findings are summarized in Theorem 3.5.

Theorem 3.5. *For any stochastic bandit algorithm, let* $\mathbb{E}\left[R^{\texttt{LSI}}(T)\right]$ *(resp.* $\mathbb{E}\left[R^{\texttt{LSIBR}}(T)\right]$ *) denote the regret in the unbounded (resp. bounded) reward setting, where each arm uses* LSI *(resp.* LSIBR(T)*). We have*

$$\mathbb{E}\left[R^{\texttt{LSIBR}}(T)\right] \geq \mathbb{E}\left[R^{\texttt{LSI}}(T)\right] - O(\frac{\underline{\Delta}\ln T}{(1-\mu_{i^*})^2})$$

4. The Robustness of ε-Greedy and Thompson Sampling

In this section, we turn our attention to two other popular classes of MAB algorithms, i.e., ε-Greedy and Thompson Sampling (TS) (Thompson, 1933; Agrawal & Goyal, 2017). Unlike UCB, these are *randomized* algorithms: ε-Greedy algorithm involves a random exploration phase and TS employs random sampling during arm selection (note: the randomness when executing UCB comes purely from the random rewards and not the algorithm itself). We establish the same regret upper bound for ε-Greedy and Thompson Sampling, again for arbitrary adaptive manipulation strategies. However, the additional randomness involved in ε-Greedy and TS makes it much more challenging to exactly characterize the SPE in the induced games. Nevertheless, we show that the regret upper bounds remain tight under the LSI strategy.

4.1. Regret Upper Bound for ε-Greedy Principal

As with UCB, we assume that the algorithm pulls arm t when $t \leq K$, i.e., first exploring each arm once. At round $t > K$, the algorithm selects an arm as follows:

$$I_t = \begin{cases} \text{uniformly drawn from } [K], & \text{w.p. } \varepsilon_t \\ \arg\max_i \left\{\widehat{\mu}_i(t-1) + \dfrac{\beta^{(i)}_{t-1}}{n_i(t-1)}\right\}, & \text{o.w.} \end{cases}$$

The first step above is *Exploration*, while the second step is *Exploitation*. We choose $\varepsilon_t = \Theta\left(\frac{1}{t}\right)$, which guarantees the convergence of the algorithm (Auer et al., 2002b). We prove the following regret bound for ε-Greedy, again for an arbitrary adaptive manipulation strategy S. As with the UCB case, the result strictly generalizes previous analysis for ε-Greedy to incorporate the effect of manipulations.

Theorem 4.1. *For any adaptive manipulation strategy S of strategic arms, the regret of the ε-Greedy principal with $\epsilon_t = \min\{1, \frac{cK}{t}\}$ and $c = \max\{20, \frac{36\sigma^2}{\underline{\Delta}}\}$, is bounded by*

$$\mathbb{E}[R(T)] \leq \sum_{i\neq i^*}\left[3B_i + \mathcal{O}\left(\frac{\ln T}{\Delta_i}\right)\right].$$

4.2. Regret Upper Bound for Thompson Sampling Principal

We model rewards with Gaussian priors and likelihood. As with UCB and ε-Greedy, we also assume that the algorithm

pulls each arm once in the first K rounds. At round $t > K$, the algorithm selects an arm according to the following procedure:

(1) For each $i \in [K]$, sample $\theta_i(t-1)$ from a Gaussian distribution $\mathcal{N}(\widetilde{\mu}_i(t-1), \frac{1}{n_i(t-1)})$, where $\widetilde{\mu}_i(t-1) = \widehat{\mu}_i(t-1) + \frac{\beta_{t-1}^{(i)}}{n_i(t-1)}$.

(2) Select arm $I_t = \arg\max_i \theta_i(t-1)$.

The total manipulation by arm i until time t, $\beta_t^{(i)}$, is induced by a strategy profile S. TS is widely known to be challenging to analyze, and its regret bound was proved only recently (Agrawal & Goyal, 2017). This is because the algorithm does not directly depend on the empirical mean of each arm, but relies on random samples from the prior distribution centered at the empirical mean. This sampling process further complicates the analysis of the stochasticity in the algorithm. Moreover, it is unclear whether there exists an effective adversarial attack to TS. This was left as an open problem in Jun et al. (2018).

Nevertheless, we prove that TS admits the same regret upper bound as UCB and ϵ-Greedy for any adaptive manipulation, up to constant factors. These results serve as an evidence of the intrinsic robustness of stochastic bandits to strategic manipulations, regardless of which no regret learning algorithm is used.

Theorem 4.2. *For any manipulation strategy profile S of strategic arms, the regret of the Thompson Sampling principal can be bounded as*

$$\mathbb{E}[R(T)] \leq \sum_{i \neq i^*} \max\left\{6B_i, \frac{72\sigma^2 \ln T}{\Delta_i}\right\} + \mathcal{O}\left(\frac{\ln T}{\Delta_i}\right). \quad (4)$$

The proof of Theorem 4.2 is quite involved as it requires us to strictly generalize the analysis in Agrawal & Goyal (2017), which is already involved, and further incorporate each arm's manipulation. Here we describe the key lemma (Lemma 4.3) that leads to the above regret lower bound, and outline its proof. All formal proofs can be found in Appendix C.

Lemma 4.3. *For any manipulation strategy profile S, the expected number of times that arm i is pulled up to time T can be bounded as follows:*

$$\mathbb{E}[n_i(T)] \leq \max\left\{\frac{6B_i}{\Delta_i}, \frac{72\sigma^2 \ln T}{\Delta_i^2}\right\} + \mathcal{O}\left(\frac{\ln T}{\Delta_i^2}\right). \quad (5)$$

Proof Sketch. Let us start with some useful notation. For each arm $k \in [K]$, we pick two thresholds x_k and y_k such that $\mu_k \leq x_k \leq y_k \leq \mu_{i^*}$. Let $E_k^{\mu}(t)$ be the event $\widetilde{\mu}_k(t-1) \leq x_k$ and $\mathbb{E}_k^{\theta}(t)$ be the event $\theta_k(t) \leq y_k$. We also denote $\mathcal{F}_t$ as the history of plays until time t. Let $\tau_{k,s}$ be the time step at which arm k is played for the s^{th} time and $p_{k,t}$ be the probability that $p_{k,t} = \mathbb{P}(\theta_{i^*}(t) \geq y_k | \mathcal{F}_{t-1})$.

The key step is to carefully decompose $\mathbb{E}[n_i(T)]$, as follows:

$$\begin{aligned}\mathbb{E}[n_i(T)] \leq 1 &+ E\left[\sum_{t=K+1}^{T} \mathbb{I}\{I_t = i, E_i^{\mu}(t), \overline{\mathbb{E}_i^{\theta}(t)}\}\right] \\ &+ \sum_{t=K+1}^{T} \mathbb{P}\left(I_t = i, E_i^{\mu}(t), \mathbb{E}_i^{\theta}(t)\right) \quad (6) \\ &+ \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\{I_t = i, \overline{E_i^{\mu}(t)}\}\right].\end{aligned}$$

The proof then proceeds by bounding each of the above terms separately. We set $x_i = \mu_i + \frac{\Delta_i}{3}, y_i = \mu_{i^*} - \frac{\Delta_i}{3}$. The first term can be bounded by $(\frac{18 \ln T}{\Delta_i^2} + 1)$ using a result of Agrawal & Goyal (2017). The second term can be bounded by $\sum_{t=K+1}^{T-1} \mathbb{E}\left[\frac{1}{p_{i,\tau_{i^*,s}+1}} - 1\right]$ We then bound each summand by the following bounds (Lemma C.4 in the Appendix):

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_{i^*,s}+1}} - 1\right] \leq \begin{cases} e^{11/4\sigma^2} + \frac{\pi^2}{3}, \ \forall s, \\ \frac{4}{T\Delta_i^2}, \text{ if } s \geq \frac{72 \ln(T\Delta_i^2) \cdot \max\{1, \sigma^2\}}{\Delta_i^2}. \end{cases}$$

Finally, we bound the third term by $\max\left\{\frac{6B_i}{\Delta_i}, \frac{144\sigma^2 \ln T}{\Delta_i^2}\right\} + 1$ (Lemma C.5). □

4.3. Regret Lower Bound

It would again be natural to consider regret under a Nash equilibrium, and perhaps dominant strategy behavior. However, the equilibrium in the game induced by a ε-Greedy or TS principal is difficult to characterize. The main challenge comes from the additional stochasticity due to the random exploration phases in ε-Greedy and TS. Nevertheless, we are able to prove the following matching lower bound on regret under `LSI` manipulation by using similar ideas as in the proof of Theorem 3.4. This shows that our upper bound is indeed tight, but does not rule out the possibility of a better regret upper bound for ε-Greedyand TS when arms' manipulations are restricted to a Nash equilibrium. It remains a challenging open question to characterize the SPE under ε-Greedy and TS. The lower bound generalizes to bounded rewards, as shown in Theorem 3.5.

Proposition 4.4. *Suppose the principal runs ε-Greedy[4] or Thompson Sampling and each strategic arm uses* `LSI`. *For any σ-sub-Gaussian reward distributions on arms, the regret of the principal satisfies,*

$$\mathbb{E}[R(T)] \geq \underline{\Delta} \sum_{i \neq i^*} \frac{B_i}{2\Delta_i} - \mathcal{O}\left(\frac{\ln T}{\underline{\Delta}}\right).$$

[4] $\epsilon_t = \min\{1, \frac{cK}{t}\}$ where $c = \max\{20, \frac{36\sigma^2}{\underline{\Delta}}\}$

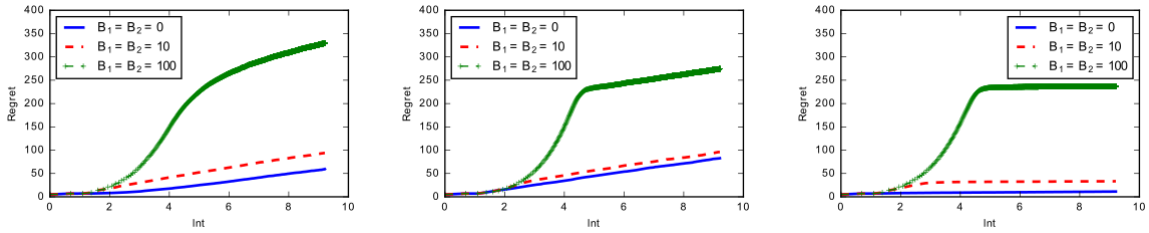

Figure 1: Total Regret as a function of $\ln t$ for the UCB principal (left), ε-Greedy principal (middle), and Thompson Sampling principal (right), for three different choices for budgets of arms 1 and 2. $B_3 = 0$ (the strongest arm).

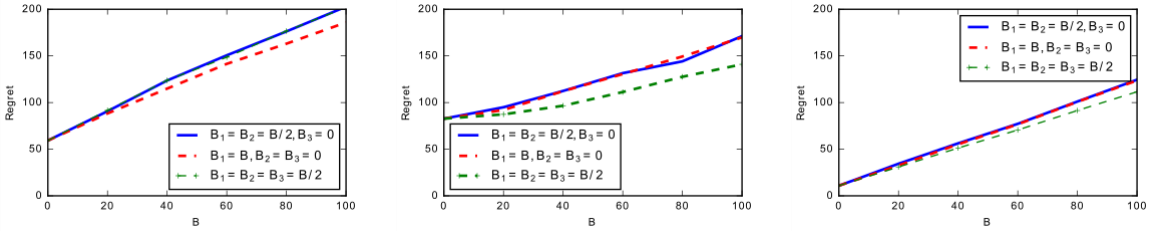

Figure 2: Total regret over $T = 10^4$ periods as a function of total budget B of arms 1 and 2, for the UCB principal (left), ε-Greedy principal (middle), and Thompson Sampling principal (right), for three different choices of how to divide the budget, and also allowing arm 3 to have budget in one scenario.

5. Simulations

In this section, we provide the results of simulations to validate our theoretical results. We only present only a representative sample here, and provide additional results in Appendix D.

Setup. There are three arms, with reward distributions $\mathcal{N}(\mu_1, \sigma^2)$, $\mathcal{N}(\mu_2, \sigma^2)$ and $\mathcal{N}(\mu_3, \sigma^2)$, respectively. We assume that $\mu_1 < \mu_2 < \mu_3$. In the ε-Greedy algorithm, we set $\varepsilon_t = \min\{1, \frac{4}{t}\}$. Throughout the simulations, we fix $\mu_1 = 5$, $\mu_2 = 8$, $\mu_3 = 10$, and $\sigma = 1$. All the arms use the `LSI` strategy. We run each bandit algorithm for $T = 10^4$ rounds, and this forms one trial. We repeat for 100 trials, and report the average results over these trials.

Regret of principal with different budgets. We consider the regret of UCB, ε-Greedy and Thompson Sampling with different budgets among the arms. For each algorithm, arm 1 and arm 2 have the same budget B_i, chosen from $\{0, 10, 100\}$. As explained earlier, it is WLOG to assume arm 3 has zero budget. We show the regret as a function of $\ln t$ in Figure 1. We observe that for small budgets (i.e., $B_i = 0, 10$), the $\Theta(\ln t)$ term dominates the regret, whereas for large budgets, the budget term B_i comes to dominate the regret as t becomes large. This is why we see a turning point in the regret curve for $B_1 = B_2 = 100$, where the regret transitions to a relatively flat curve since the budget is fixed. Interestingly, we find that Thompson sampling performs better than both UCB and ε-Greedy in this strategic manipulation scenario.

Regret is linear with total budget. We validate that the regret achieved by each stochastic bandit algorithm with strategic manipulations is linear in the total budget available to the strategic arms. We vary the budget $B = B_1 + B_2$ available to arms 1 and 2, and consider three settings: (1) $B_1 = B_2 = B/2, B_3 = 0$, (2) $B_1 = B, B_2 = B_3 = 0$, and (3) $B_1 = B_2 = B_3 = B/2$. For setting (1), we equally split the budget to arm 1 and arm 2. For setting (2), we give all the budget to arm 1. For setting (3), we also give the optimal arm some budget (and assume arm 3 uses strategy `LSI`), and want to understand the effect of the budget of the optimal arm.

Figure 2 shows the regret of each algorithm at the end of the $T = 10^4$ rounds, as budget $B = B_1 + B_2$ varies. The regret is generally linearly increasing with B, validating the theoretical findings. Interestingly, even if the optimal arm also has available budget, the regret still increases as the budget for arms 1 and 2 increase. In fact, the regret in this case, where the optimal arm also has budget, is similar to that when it does not, and the budget on optimal arm 3 does not affect the regret much. This is because the optimal arm will in any case be pulled many times, and its budget will be diluted significantly in later rounds, so that it has only a small effect on regret.

Acknowledgments

This work is supported in part through NSF award CCF-1841550, as well as a Google Fellowship for Zhe Feng. Haifeng Xu is supported by a Google Faculty Award. We would like to thank anonymous reviewers for their helpful feedback.

References

Agrawal, S. and Goyal, N. Near-optimal regret bounds for Thompson sampling. *J. ACM*, 64(5):30:1–30:24, September 2017. ISSN 0004-5411.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multi-armed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002a. ISSN 0885-6125.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Ben-Porath, E. The complexity of computing a best response automaton in repeated games with mixed strategies. *Games and Economic Behavior*, 2(1):1–12, 1990.

Besbes, O., Gur, Y., and Zeevi, A. Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards. *Stochastic Systems*, 9(4):319–337, 2019.

Blum, A., Kumar, V., Rudra, A., and Wu, F. Online learning in online auctions. *Theoretical Computer Science*, 324 (2-3):137–146, 2004.

Braverman, M., Mao, J., Schneider, J., and Weinberg, S. M. Multi-armed bandit problems with strategic arms. In *Conference on Learning Theory*, 2019.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Chapelle, O., Manavoglu, E., and Rosales, R. Simple and scalable response prediction for display advertising. *ACM Trans. Intell. Syst. Technol.*, 5(4):61:1–61:34, December 2014. ISSN 2157-6904.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Learning to optimize under non-stationarity. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1079–1087. PMLR, 16–18 Apr 2019.

Conitzer, V. and Sandholm, T. Complexity results about nash equilibria. In *Proceedings of the 18th international joint conference on Artificial intelligence*, pp. 765–771. Morgan Kaufmann Publishers Inc., 2003.

Feldman, M., Koren, T., Livni, R., Mansour, Y., and Zohar, A. Online pricing with strategic and patient buyers. In *Advances in Neural Information Processing Systems 29*, pp. 3864–3872. Curran Associates, Inc., 2016.

Feng, Z., Podimata, C., and Syrgkanis, V. Learning to bid without knowing your value. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, EC '18, pp. 505–522, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5829-3.

Feng, Z., Schrijvers, O., and Sodomka, E. Online learning for measuring incentive compatibility in ad auctions. In *Proceedings of the Web Conference*, WWW '19, 2019.

Frazier, P., Kempe, D., Kleinberg, J., and Kleinberg, R. Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 5–22, 2014.

Gupta, A., Koren, T., and Talwar, K. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pp. 1562–1578, 2019.

Immorlica, N., Mao, J., Slivkins, A., and Wu, Z. S. Bayesian exploration with heterogeneous agents. In *Proceedings of the 2019 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2019.

Jun, K.-S., Li, L., Ma, Y., and Zhu, J. Adversarial attacks on stochastic bandits. In *Advances in Neural Information Processing Systems 31*, pp. 3644–3653. Curran Associates, Inc., 2018.

Kveton, B., Szepesvári, C., Wen, Z., and Ashkan, A. Cascading bandits: Learning to rank in the cascade model. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 767–776. JMLR.org, 2015.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 661–670, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8.

Liu, Z., Wang, H., Shen, F., Liu, K., and Chen, L. Incentivized exploration for multi-armed bandits under reward drift. In *AAAI*, 2020.

Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, pp. 114–122, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5559-9.

Mansour, Y., Slivkins, A., and Syrgkanis, V. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pp. 565–582. ACM, 2015.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444.

Thorisson, H. *Coupling, Stationarity, and Regeneration*. Probability and Its Applications. Springer New York, 2000. ISBN 9780387987798.

Wang, S. and Huang, L. Multi-armed bandits with compensation. In *Advances in Neural Information Processing Systems*, pp. 5114–5122, 2018.

Weed, J., Perchet, V., and Rigollet, P. Online learning in repeated auctions. In *Conference on Learning Theory*, pp. 1562–1583, 2016.