Proceedings of Machine Learning Research vol 99:1–33, 2019 32nd Annual Conference on Learning Theory

Multi-armed Bandit Problems with Strategic Arms

Mark Braverman MBRAVERM@CS.PRINCETON.EDU
Department of Computer Science, Princeton University

Jieming Mao MAOJM517@GMAIL.COM
Warren Center, University of Pennsylvania

Jon Schneider JSCHNEI@GOOGLE.COM
Google

S. Matthew Weinberg SMWEINBERG@PRINCETON.EDU
Department of Computer Science, Princeton University

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We study a strategic version of the multi-armed bandit problem, where each arm is an individual strategic agent and we, the principal, pull one arm each round. When pulled, the arm receives some private reward v_a and can choose an amount x_a to pass on to the principal (keeping $v_a - x_a$ for itself). All non-pulled arms get reward 0. Each strategic arm tries to maximize its own utility over the course of T rounds. Our goal is to design an algorithm for the principal incentivizing these arms to pass on as much of their private rewards as possible.

When private rewards are stochastically drawn each round ($v_a^t \leftarrow D_a$), we show that:

- Algorithms that perform well in the classic adversarial multi-armed bandit setting necessarily perform poorly: For all algorithms that guarantee low regret in an adversarial setting, there exist distributions $D_1, \dots, D_k$ and an $o(T)$-approximate Nash equilibrium for the arms where the principal receives reward $o(T)$.
- There exists an algorithm for the principal that induces a game among the arms where each arm has a dominant strategy. Moreover, for every $o(T)$-approximate Nash equilibrium, the principal receives expected reward $\mu' T - o(T)$, where μ' is the second-largest of the means $\mathbb{E}[D_a]$. This algorithm maintains its guarantee if the arms are non-strategic ($x_a = v_a$), and also if there is a mix of strategic and non-strategic arms.

Keywords: multi-armed bandit, strategic learning, auction design

1. Introduction

Classically, algorithms for problems in machine learning assume that their inputs are drawn either stochastically from some fixed distribution or chosen adversarially. In many contexts, these assumptions do a fine job of characterizing the possible behavior of problem inputs. Increasingly, however, these algorithms are being applied to contexts (ad auctions, search engine optimization, credit scoring, etc.) where the quantities being learned are controlled by rational agents with external incentives. To this end, it is important to understand how these algorithms behave in *strategic* settings.

The multi-armed bandit problem is a fundamental decision problem in machine learning that models the trade-off between exploration and exploitation, and is used extensively as a building block in other machine learning algorithms (e.g. reinforcement learning). A learner (who we refer to as the *principal*) is a sequential decision maker who at each time step t, must decide which of k arms to 'pull'. Pulling this arm bestows a reward (either adversarially or stochastically generated) to the principal, and the principal would like to maximize his overall reward. Known algorithms for this problem guarantee that the principal can do approximately as well as the best individual arm.

In this paper, we consider a strategic model for the multi-armed bandit problem where each arm is an individual strategic agent and each round one arm is pulled by an agent we refer to as the *principal*. Each round, the pulled arm receives a private reward $v \in [0, 1]$ and then decides what amount x of this reward gets passed on to the principal (upon which the principal receives utility x and the arm receives utility $v - x$). Each arm therefore has a natural tradeoff between keeping most of its reward for itself and passing on the reward so as to be chosen more frequently. Our goal is to design mechanisms for the principal which simultaneously learns which arms are valuable while also incentivizing these arms to pass on most of their rewards.

This model captures a variety of dynamic agency problems, where at each time step the principal must choose to employ one of K agents to perform actions on the principal's behalf, where the agent's cost of performing that action is unknown to the principal (for example, hiring one of K contractors to perform some work, or hiring one of K investors with external information to manage some money - the important feature being that the principal doesn't know exactly how much they will pay/receive/etc. until the job is done, and the agent has a lot of freedom to set this ex-post). In this sense, this model can be thought of as a multi-agent generalization of the principal-agent problem in contract theory when agents are allowed private savings (see Section 1.2 for references). The model also captures, for instance, the interaction between consumers (as the principal) and many sellers deciding how steep a discount to offer the consumers - higher prices now lead to immediate revenue, but offering better discounts than your competitors will lead to future sales. In all domains, our model aims to capture settings where the principal has little domain-specific or market-specific knowledge, and can really only process the reward they get for pulling an arm and not any external factors that contributed to that reward.

There are two "obvious" approaches to try and solve these problems: Option one is to treat it like a procurement auction and run a reverse second-price auction. This doesn't quite work, however, in the case where the agents don't initially know how much reward they'll generate, so some amount of learning needs to enter the picture for a solution to be viable. Using the contractor as a *toy* running example: the contractor will not initially know how much it costs her to work on your home, but after working on your home several times *they* will start to learn how much the next one will cost (you will only learn how much they charge you). In any case, one cannot simply treat it like an auctions problem and ignore learning completely.

The second "obvious" approach is just to treat it as a learning problem, and ignore incentives completely. In fact, one oft-cited motivation for considering adversarial rewards in bandit settings is that arms might be strategic. Indeed, this is because even if the arms' rewards are stochastic, the utility they strategically pass on to the principal is unlikely to follow any distribution. Algorithms like EXP3 which guarantee low-regret in adversarial settings then seem like the natural "pure learning" approach. Interestingly, our main "negative result" shows that *any* adversarial learning algorithm admits a really bad approximate Nash equilibrium (more details below).

So auctions alone cannot solve the problem, nor can learning alone. To complement our main negative result, we show that the right combination of auctions and learning yields a positive result: an algorithm such that all approximate Nash result in good utility for the principal. We now overview our results in more detail.

1.1. Our results

1.1.1. Low-regret algorithms are far from strategyproof

Many algorithms for the multi-armed bandit problem are designed to work in worst-case settings, where an adversary can adaptively decide the value of each arm pull. Here, algorithms such as EXP3 (Auer et al. (2003)) guarantee that the principal receives almost as much as if he had only pulled the best arm. Formally, such algorithms guarantee that the principal experiences at most $O(\sqrt{T})$ regret over T rounds compared to any algorithm that only plays a single arm (when the adversary is oblivious).

Given these worst-case guarantees, one might naively expect low-regret algorithms such as EXP3 to also perform well in our strategic variant. It is important to note, however, that single arm strategies perform dismally in this strategic setting; if the principal only ever selects one arm, the arm has no incentive to pass along any surplus to the principal. In fact, we show that the objectives of minimizing adversarial regret and performing well in this strategic variant are fundamentally at odds.

Theorem 1 (informal restatement of Theorem 21) *Let M be a low-regret algorithm for the classic multi-armed bandit problem with adversarially chosen values. Then there exists an instance of the strategic multi-armed bandit problem and an $o(T)$-Nash equilibrium for the arms where a principal running M receives at most $o(T)$ revenue.*

While not immediately apparent from the statement of Theorem 1, these instances where low-regret algorithms fail are far from pathological; in particular, there is a problematic equilibrium for any instance where arm i receives a fixed reward v_i each round it is pulled, as long as the the gap between the largest and second-largest v_i is not too large (roughly 1/#arms).

Here we assume the game is played under a *tacit* observational model, meaning that arms can only observe which arms get pulled by the principal, but not how much value they give to the principal. In particular, this means that arms can achieve this equilibrium despite not communicating directly with each other and not observing the actions of the other arms. This rules out various sorts of "grim trigger" collusion strategies (similar to collusion that occurs in the setting of repeated auctions, see Skrzypacz and Hopenhayn (2004)), where arms agree on a protocol ahead of time and immediately defect as soon as one arm deviates from this protocol. (Indeed, in an *explicit* observational model, where arms can see both which arms get pulled and how much value they pass on, it is easy to show even stronger results via such strategies; see Appendix A.2 for details).

Instead, the strategies in the equilibrium of Theorem 1 take the form of *market-sharing strategies*, where arms calibrate their actions so that they each get played some proportion (e.g. $1/K$) of the time while passing on little utility to the principal. For example, consider a simple instance of this problem with two strategic arms, where the principal is using the low-regret EXP3 algorithm, and where arm 1 always gets private reward 1 if pulled and arm 2 always gets private reward 0.8. By always reporting some value slightly larger than 0.8, arm 1 can incentivize the principal to almost always pull it in the long run. This gains arm 1 roughly 0.2 utility per round (and arm 2 nothing). On the other hand, if arm 1 and arm 2 never pass along any surplus to the principal, they will likely be played equally often, gaining arm 1 roughly 0.5 utility per round and arm 2 0.4 utility per round.

To show such a market-sharing strategy works for general low-regret algorithms, much more work needs to be done. The arms must be able to enforce an even split of the principal's pulls (as soon as the principal starts lopsidedly pulling one arm more often than the others, the remaining arms can defect and start reporting their full value whenever pulled). As long as the principal guarantees good performance in the non-strategic adversarial case (achieving $o(T)$ regret), we show that the arms can (at $o(T)$ cost to themselves, and without explicitly communicating) cooperate so that they are all played equally often.

1.1.2. MECHANISMS FOR STRATEGIC ARMS WITH STOCHASTIC VALUES

We next show that, in contrast to Theorem 1, it is in fact possible for the principal to extract positive values from the arms per round, if we do not restrict the principal to use an adversarial low-regret algorithm (and hence there is a price to being adversarial low-regret).

We consider a setting where each arm i's reward when pulled is drawn independently from some distribution D_i with mean μ_i (unknown to the principal). In this case the principal can extract the value of the second-best arm (which is the best possible, as we show in Lemma 14). In the below statement, we are using the term "truthful mechanism" quite loosely as shorthand for "strategy that induces a game among the arms where each arm has a dominant strategy."

Theorem 2 (restatement of Corollary 16) *Let μ' be the second largest mean amongst the set of μ_is. Then there exists a truthful mechanism for the principal that guarantees revenue at least $\mu' T - o(T)$ when the arms are playing according to any $o(T)$-Nash equlibrium.*

The mechanism in Theorem 2 can be thought of as a combination of a second-price auction with the explore-then-exploit strategy from multi-armed bandits. The principal divides the time horizon into three "phases". In the first phase (of size $o(T)$), the principal begins by asking each arm i to simply report their value each round, thus allowing the principal to learn which arm is the most valuable. In the second phase (which comprises the vast majority of the rounds), the principal asks the most valuable arm (the arm with the highest mean in the first phase) to give him the second-largest mean worth of value per round. If this arm fails to comply in any round, the principal avoids picking this arm for the remainder of the rounds. Finally, in the third phase, the principal uses a proper scoring rule to recompensate all arms for reporting truthfully in the first phase. (A more detailed description of the mechanism can be seen in Mechanisms 1 and 2 in Section 4).

As an added bonus, we show that this mechanism has similar guarantees in the setting where some arms are strategic and some arms are non-strategic (and our mechanism does not know which arms are which).

Theorem 3 (restatement of Theorem 18) *Let μ_s be the second largest mean amongst the means of the strategic arms, and let μ_n be the largest mean amongst the means of the non-strategic arms. Then there exists a truthful mechanism for the principal that guarantees (with probability $1 - o(1/T)$) revenue at least $\max(\mu_s, \mu_n)T - o(T)$ when arms play according to any $o(T)$-Nash equilibrium.*

In particular, this implies that Mechanism 2 has low-regret in the classical *stochastic* multi-armed bandits setting, and so the adversarial aspect of the low-regret guarantees is actually essential for the proof of Theorems 1.

A fair critique of this mechanism is that most of the work of learning the distributions of the arms is offloaded to the beginning of the game. This is appealing because it makes it much feasible to "slide in" some auction design and scoring rules to handle incentives. It is an interesting problem whether learning can still be done adaptively over time in this model, as such a procedure would necessitate a much more sophisticated treatment of incentives; see Section 5 for further discussion.

1.2. Related work

The study of classical multi-armed bandit problems was initiated by Robbins (1952), and has since grown into an active area of study. The most relevant results for our paper concern the existence of low-regret bandit algorithms in the adversarial setting, such as the EXP 3 algorithm (Auer et al. (2003)), which achieves regret $\tilde{O}(\sqrt{KT})$. Other important results in the classical setting include the upper confidence bound (UCB) algorithm for stochastic bandits (Lai and Robbins (1985)) and the work of Gittins and Jones (1974) for Markovian bandits. For further details about multi-armed bandit problems, see the survey Bubeck and Cesa-Bianchi (2012).

One question that arises in the strategic setting (and other adaptive settings for multi-armed bandits) is what the correct notion of regret is; standard notions of regret guarantee little, since the best overall arm may still have a small total reward. Arora et al. (2012) considered the multi-armed bandit problem with an adaptive adversary and introduced the quantity of "policy regret", which takes the adversary's adaptiveness into account. They showed that any multi-armed bandit algorithm will get $\Omega(T)$ policy regret. This indicates that it is not enough to treat strategic behaviors as an instance of adaptively adversarial behavior; good mechanisms for the strategic multi-armed bandits problem must explicitly take advantage of the rational self-interest of the arms.

Our model bears some similarities to the principal-agent problem of contract theory, where a principal employs an more informed agent to make decisions on behalf of the principal, but where the agent may have incentives misaligned from the principal's interests when it gets private savings (for example Chassang (2013)). For more details on principal-agent problem, see the book Laffont and Martimort (2002). Our model can be thought of as a sort of multi-armed version of the principal-agent problem, where the principal has many agents to select from (the arms) and can try to use competition between the agents to align their interests with the principal.

Our negative results are closely related to results on collusions in repeated auctions. Existing theoretical work McAfee and McMillan (1992); Athey and Bagwell (2001); Johnson and Robert (1999); Aoyagi (2003, 2007); Skrzypacz and Hopenhayn (2004) has shown that collusive schemes exist in repeated auctions in many different settings, e.g., with/without side payments, with/without communication, with finite/infinite typespace. In some settings, efficient collusion can be achieved, i.e., bidders can collude to allocate the good to the bidders who values it the most and leave 0 asymptotically to the seller. Even without side payments and communication, Skrzypacz and Hopenhayn (2004) showed that tacit collusion exists and can achieve asymptotic efficiency with a large cartel.

Our setting is similar to settings considered in a variety of work on dynamic mechanism design, often inspired by online advertising. Bergemann and Vlimki (1996) considers the problem where a buyer wants to buy a stream of goods with an unknown value from two sellers, and examines Markov perfect equilibria in this model. Babaioff et al. (2009); Devanur and Kakade (2009); Babaioff et al. (2010) study truthful pay-per-click auctions where the auctioneer wishes to design a truthful mechanism that maximizes the social welfare. Kremer et al. (2014); Frazier et al. (2014) consider the scenario where the principal cannot directly choose which arm to pull, and instead must incentivize a stream of strategic players to prevent them from acting myopically. Amin et al. (2013, 2014) consider a setting where a seller repeatedly sells to a buyer with unknown value distribution, but the buyer is more heavily discounted than the seller. Kakade et al. (2013) develops a general method for finding optimal mechanisms in settings with dynamic private information. Nazerzadeh et al. (2008) develops an ex ante efficient mechanism for the Cost-Per-Action charging scheme in online advertising.

2. Our Model

2.1. Classic Multi-Armed Bandits

We begin by reviewing the definition of the classic multi-armed bandits problem and associated quantities.

In the classic multi-armed bandit problem a learner (the *principal*) chooses one of K choices (arms) per round, over T rounds. On round t, the principal receives some reward $v_{i,t} \in [0, 1]$ for pulling arm i. The values $v_{i,t}$ are either drawn independently from some distribution corresponding to arm i (in the case of *stochastic bandits*) or adaptively chosen by an adversary (in the case of *adversarial bandits*). Unless otherwise specified, we will assume we are in the adversarial setting.

Let I_t denote the arm pulled by the principal at round t. The *revenue* of an algorithm M is the random variable

$$\mathsf{Rev}(M) = \sum_{t=1}^{T} v_{I_t,t}$$

and the the *regret* of M is the random variable

$$\mathsf{Reg}(M) = \max_i \sum_{t=1}^{T} v_{i,t} - \mathsf{Rev}(M)$$

Definition 4 (δ-Low Regret Algorithm) *Mechanism M is a δ-low regret algorithm for the multi-armed bandit problem if*

$$\mathbb{E}[\mathsf{Reg}(M)] \leq \delta.$$

Here the expectation is taken over the randomness of M and the adversary.

Definition 5 ((ρ, δ)-Low Regret Algorithm) *Mechanism M is a (ρ, δ)-low regret algorithm for the multi-armed bandit problem if with probability $1 - \rho$,*

$$\mathsf{Reg}(M) \leq \delta.$$

There exist $O(\sqrt{KT\log K})$-low regret algorithms and $(\rho, O(\sqrt{KT\log(K/\rho)}))$-low regret algorithms for the multi-armed bandit problem; see Section 3.2 of Bubeck and Cesa-Bianchi (2012) for details.

2.2. Strategic Multi-Armed Bandits

The strategic multi-armed bandits problem builds upon the classic multi-armed bandits problem with the notable difference that now arms are strategic agents with the ability to withhold some payment from the principal. Instead of the principal directly receiving a reward $v_{i,t}$ when choosing arm i, now arm i receives this reward and passes along some amount $w_{i,t}$ to the principal, gaining the remainder $v_{i,t} - w_{i,t}$ as utility.

For simplicity, in the strategic setting, we will assume the rewards $v_{i,t}$ are generated stochastically; that is, each round, $v_{i,t}$ is drawn independently from a distribution D_i (where the distributions D_i are known to all arms but not to the principal). While it is possible to pose this problem in the adversarial setting (or other more general settings), this comes at the cost of there being no clear notion of strategic equilibrium for the arms.

This strategic variant comes with two additional modeling assumptions. The first is the informational model of this game; what information does an arm observe when some other arm is pulled. We define two possible observational models:

1. **Explicit:** After each round t, every arm sees the arm played I_t along with the quantity $w_{I_t,t}$ reported to the principal.

2. **Tacit:** After each round t, every arm only sees the arm played I_t.

In both cases, only arm i knows the size of the original reward $v_{i,t}$; in particular, the principal also only sees the value $w_{i,t}$ and learns nothing about the amount withheld by the arm. Collusion between arms is generally significantly easier in the explicit observational model than in the tacit observational model, and for this reason we will assume we are in the tacit observational model unless otherwise stated.

The second modeling assumption is whether to allow arms to go into debt while paying the principal. In the *restricted payment* model, we impose that $w_{i,t} \leq v_{i,t}$; an arm cannot pass along more than it receives in a given round. In the *unrestricted payment* model, we let $w_{i,t}$ be any value in $[0, 1]$. We prove our negative results in the restricted payment model and our positive results in the unrestricted payment model, but our proofs for our negative results work in both models (in particular, it is easier to collude and prove negative results in the unrestricted payment model) and Mechanism 2 can be adapted to work in the restricted payment model (see discussion in Section 4.2).

Finally, we proceed to define the set of strategic equilibria for the arms. We assume the mechanism M of the principal is fixed ahead of time and known to the K arms. If each arm i is using a (possibly adaptive) strategy S_i, then the expected utility of arm i is defined as

$$u_i(M, S_1, \dots, S_K) = \mathbb{E}\left[\sum_{t=1}^{T}(v_{i,t} - w_{i,t}) \cdot \mathbb{1}_{I_t=i}\right].$$

An ε-Nash equilibrium for the arms is then defined as follows.

Definition 6 (ε-Nash Equilibrium for the arms) *Strategies $(S_1, ..., S_K)$ form an ε-Nash equilibrium for the strategic multi-armed bandit problem if for all $i \in [n]$ and any deviating strategy S'_i,*

$$u_i(S_1, \dots, S_i, \dots, S_K) \geq u_i(S_1, \dots, S'_i, \dots, S_K) - \varepsilon.$$

Similarly as before, the revenue of the principal in this case is the random variable

$$\mathsf{Rev}(M, S_1, \dots, S_K) = \sum_{t=1}^{T} w_{I_t,t}.$$

The goal of the principal is to choose a mechanism M which guarantees large revenue in any ε-Nash Equilibrium for the arms.

In Section 4, we will construct mechanisms for the strategic multi-armed bandit problem which are truthful for the arms. We define the related terminology below.

Definition 7 (Dominant Strategy) *When the principal uses mechanism M, we say S_i is a dominant strategy for arm i if for any deviating strategy S'_i and any strategies for other arms $S_1, .., S_{i-1}, S_{i+1}, ..., S_K$,*

$$u_i(M, S_1, \dots, S_i, \dots, S_K) \geq u_i(M, S_1, \dots, S'_i, \dots, S_K).$$

Definition 8 (Truthfulness) *We say that a mechanism M for the principal is truthful, if all arms have some dominant strategies.*

3. Negative Results Overview

In this section we give a sketch of the proof of our main theorem, Theorem 21. The full list of our negative results and proofs can be found in Appendix A.

Theorem 9 *[(restatement of Theorem 21)] Let mechanism M be a (ρ, δ)-low regret algorithm for the multi-armed bandit problem with K arms, where $K \leq T^{1/3}/\log(T)$, $\rho \leq T^{-2}$, and $\delta \geq \sqrt{T \log T}$. Then in the strategic multi-armed bandit problem under the tacit observational model, there exist distributions D_i and an $O(\sqrt{KT\delta})$-Nash Equilibrium for the arms where the principal gets at most $O(\sqrt{KT\delta})$ revenue.*

Proof [Proof Sketch] The underlying idea here is that the arms work to try to maintain an equal market share, where each of the K arms are each played approximately $1/K$ of the time. To ensure this happens, arms collude so that arms that aren't as likely to be pulled pass along more than arms that have been pulled a lot or are more likely to be pulled; this ends up forcing any low-regret algorithm for the principal to choose all the arms equally often. Interestingly, this collusion strategy is *mechanism dependent*, as arms need to estimate the probability they will be pulled in the next round.

More formally, let μ_i denote the mean of the ith arm's distribution D_i. Without loss of generality, further assume that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. We will show that as long as $\mu_1 - \mu_2 \leq \frac{\mu_1}{K}$, there exists some $O(\sqrt{KT\delta})$-Nash equilibrium for the arms where the principal gets at most $O(\sqrt{KT\delta})$ revenue.

We begin by describing the equilibrium strategy S^* for the arms. Let $c_{i,t}$ denote the number of times arm i has been pulled up to time t. Set $B = 7\sqrt{KT\delta}$ and set $\theta = \sqrt{\frac{K\delta}{T}}$. The equilibrium strategy for arm i at time t is as follows:

1. If at any time $s \leq t$ in the past, there exists an arm j with $c_{j,s} - c_{i,s} \geq B$, defect and offer your full value $w_{i,t} = \mu_i$.

2. Compute the probability $p_{i,t}$, the probability that the principal will pull arm i conditioned on the history so far.

3. Offer $w_{i,t} = \theta(1 - p_{i,t})$.

The main technical challenge in proving that this strategy is an equilibrium involves showing that, if all arms are following this strategy and the principal is using a low-regret mechanism, then with high probability the arms will not defect. Here the low-regret property of the mechanism M is essential (indeed, as our positive results imply, the theorem is not true without this assumption). In particular, by the construction of $w_{i,t}$ in terms of $p_{i,t}$, the principal's expected total regret (here defined to be the sum of the principal's regrets with respect to each arm) will increase each round by some amount proportional to the variance of the $p_{i,t}$. Intuitively, this implies that the values $p_{i,t}$ cannot be too far from uniform for too many rounds, and therefore that each arm should be picked approximately the same proportion of the time. This is formalized in the following lemma and its proof can be found in Appendix A:

Lemma 10 *If all arms are using strategy S^*, then with probability $\left(1 - \frac{3}{T}\right)$, $|c_{i,t} - c_{j,t}| \leq B$ for all $t \in [T], i, j \in [K]$.*

Lemma 10 implies that if each arm plays strategy S^*, then each arm i will receive on average μ_i/K per round. To finish the proof, it suffices to note that by deviating and playing a different strategy S from S^*, one of two things can occur. If playing this different strategy S does not trigger the defect condition in (1), then still each arm will be played roughly $1/K$ of the time (and your total utility is unchanged up to $o(T)$ additive factors). On the other hand, once the defect condition is triggered, you can receive at most $\mu_1 - \mu_2$ utility per round (and only if you are arm 1). This implies that as long as $\mu_1/K > \mu_1 - \mu_2$, there is no incentive to deviate.

Additional details are provided in Appendix A. ■

While the theorem above merely claims that a bad set of distributions for the arms exists, the proof shows it is possible to collude in a wide range of instances - in particular, any collection of distributions which satisfies $\mu_1 - \mu_2 \leq \mu_1/K$. A natural question is whether we can extend the above results to show that it is possible to collude in any set of distributions.

One issue with the collusion strategy in the above proof is that if $\mu_1 - \mu_2 > \mu_1/K$, then arm 1 will have an incentive to defect in any collusive strategy that plays all the arms evenly (arm 1 can report a bit over μ_2 per round, and make $\mu_1 - \mu_2$ every round instead of μ_1 every K rounds). One solution to this is to design a collusive strategy that plays some arms more than others in equilibrium (for example, playing arm 1 90% of the time). We show how to modify our result for two arms to achieve an arbitrary market partition and thus work over a broad set of distributions.

Theorem 11 *Let mechanism M be a (ρ, δ)-low regret algorithm for the multi-armed bandit problem with two arms, where $\rho \leq T^{-2}$ and $\delta \geq \sqrt{T \log T}$. Then, in the strategic multi-armed bandit problem under the tacit observational model, for* any *distributions D_1, D_2 of values for the arms (supported on $[\sqrt{\delta/T}, 1]$), there exists an $O(\sqrt{T\delta})$-Nash Equilibrium for the arms where a principal using mechanism M gets at most $O(\sqrt{T\delta})$ revenue.*

Proof See Appendix A. ■

Unfortunately, it as not as easy to modify the proof of Theorem 21 to prove the same result for K arms. It is an interesting open question whether there exist collusive strategies for K arms that can achieve an arbitrary partition of the market.

4. Positive Results

In this section we will show that, in contrast to the previous results on collusion, there exists a mechanism for the principal that can obtain $\Theta(T)$ revenue from the arms when they play according to an $o(T)$-Nash equilibrium.

We begin by demonstrating a simpler version of our mechanism (Mechanism 1) that guarantees the principal $\Theta(T)$ revenue whenever the arms play according to their dominant strategies. In Section 4.2, we then show how to make this mechanism more robust (Mechanism 2) so that the principal is guaranteed $\Theta(T)$ revenue whenever the arms play according to any $o(T)$-approximate Nash equilibrium (thus showing a separation between the power of adversarial low-regret algorithms and general learning algorithms in this model). As an added bonus, we show that this mechanism also works for a combination of strategic and non-strategic arms (and therefore achieves low regret in the classical stochastic multi-armed bandits setting).

Throughout this section we will assume we are working in the tacit observational model and the unrestricted payment model (unless otherwise specified). We postpone all the proofs of this section to Appendix B.

4.1. Good dominant strategy equilibria

This mechanism essentially incentivizes each arm to report the mean of its distribution and then runs a second-price auction, asking the arm with the highest mean for the second-highest mean each round.

Define μ_i as the mean of distribution D_i for $i = 1, \dots, K$, let $\mu_{min} = \min_{i:\mu_i \neq 0}(\mu_i)$, and $u = -\log \mu_{min} + 1$. We assume throughout that $u = o(T/K)$.

Mechanism 1 Truthful mechanism for strategic arms with known stochastic values in the tacit model

---

1: Play each arm once (i.e. play arm 1 in the first round, arm 2 in the second round, etc.). Let w_i be the value arm i reports in round i.

2: Let $i^* = \arg\max w_i$ (breaking ties lexicographically), and let $w' = \max_{i \neq i^*} w_i$.

3: Tell arm i^* the value of w'. Play arm i^* for $R = T - (u+2)K - 1$ rounds. If arm i^* ever reports a value different from w', stop playing it immediately. If arm i^* always gives w', play it for one bonus round (ignoring the value it reports).

4: For each arm i such that $i \neq i^*$, play it for one round.

5: For each arm i satisfying $u + \log(w_i) \geq 0$, play it $\lfloor u + \log(w_i) \rfloor$ times. Then, with probability $u + \log(w_i) - \lfloor u + \log(w_i) \rfloor$, play arm i for one more round.

We will first show that the dominant strategy of each arm in this mechanism includes truthfully reporting their mean at the beginning, and then then compute the principal's revenue under this dominant strategy.

Lemma 12 *The following strategy is the dominant strategy for arm i in Mechanism 1:*

1. *(line 1 of Mechanism 1) Report the mean value μ_i of D_i the first time when arm i is played.*
2. *(lines 3,4 of Mechanism 1) If $i = i^*$, for the R rounds that the principal expects to see reported value w', report the value w'. For the bonus round, report 0. If $i \neq i^*$, report 0.*
3. *(line 5 of Mechanism 1) For all other rounds, report* 0.

Corollary 13 *Under Mechanism 1, the principal will receive revenue at least $\mu' T - o(T)$ when arms use their dominant strategies, where μ' is the second largest mean in the set of means μ_i.*

We additionally show that the performance of Mechanism 1 is as good as possible; no mechanism can do better than the second-best arm in the worst case.

Lemma 14 *Let μ and μ' be the largest and second largest values respectively among the μ_i. Then for any constant $\alpha > 0$, no truthful mechanism can guarantee $(\alpha\mu + (1-\alpha)\mu')T$ revenue in the worst case.*

4.2. Good approximate Nash equilbria

One issue with Mechanism 1 is that, while the principal achieves $\Theta(T)$ revenue when the arms play according to their dominant strategies, there can exist ϵ-Nash equilibria for the arms which still leave the principal with negligible revenue. For instance, if there are two arms with equal means $\mu_1 = \mu_2 = \mu$, one possible ϵ-Nash equilibrium is for them both to bid μ, and then for arm i^* to immediately defect after it is chosen. This is not a dominant strategy, since arm i^* surrenders its bonus for not defecting, but since this bonus is at most 1, this is still an ϵ-Nash equilibrium for any $\epsilon = o(T)$ which is larger than 1.

We can make Mechanism 2 more robust to strategies like this by increasing the size of the bonus with ϵ. If we additionally allow a tiny buffer between the current reported average and w', this mechanism has the added property that it works even when there are a mixture of strategic and non-strategic arms (and the principal does not know which are which). In particular, this Mechanism 2 obtains low-regret in the classical stochastic multi-armed bandits setting, which implies that our negative results in Section 3 are really due to the adversarial nature of the low-regret guarantees.

As before, define μ_i as the mean of distribution D_i for $i = 1, ..., K$. Our mechanism takes in two parameters, B (representing the size of the bonus) and M (representing the size of the buffer). We will set $B = 2\epsilon^{1/4}T^{3/4}/\mu_{min}$ and $M = 8B^{-1/2}\ln(KT)$. In addition, we will define $u = -\log\left(\min_{i:\mu_i \neq 0} \mu_i\right) + 2 + M$. We assume $u = o(\frac{T}{BK})$.

We begin by characterizing the dominant strategy for Mechanism 2. Similarly as in Lemma 12, we show that this dominant strategy involves each arm reporting their true mean in the beginning rounds.

Lemma 15 *The following strategy is the dominant strategy for arm i in Mechanism 2:*

1. *(line 1 of Mechanism 2) For the first B rounds, report a total sum of $(\mu_i + M)B$.*
2. *(lines 3,4 of Mechanism 2) If $i = i^*$, for the R rounds that the principal expects to see reported value w', report the value $w' - M$. For the B bonus rounds, report 0. If $i \neq i^*$, report 0.*

Mechanism 2 Truthful mechanism for strategic/non-strategic arms in the tacit model

1: Play each arm B times (i.e. play arm 1 in the first B rounds, arm 2 in the next B rounds, etc.). Let $\bar{w}_i$ be the average value arm i reported in its B rounds.
2: Let $i^* = \arg\max \bar{w}_i$ (breaking ties lexicographically), and let $w' = \max_{i \neq i^*} \bar{w}_i$.
3: Tell arm i^* the value of w'. Play arm i^* for $R = T - (u+3)BK$ rounds. If arm i^* ever reports values with average less than $w' - M$ in any round after B rounds in this step, stop playing it immediately. If arm i^* gives average no less than $w' - M$, play it for B bonus rounds (ignoring the value it reports).
4: For each arm i such that $i \neq i^*$, play it for B rounds.
5: For each arm i satisfying $u + \log(\bar{w}_i - M) \geq 0$, play it $B\lfloor (u + \log(\bar{w}_i - M)) \rfloor$ times. Then, with probability $u + \log(\bar{w}_i - M) - \lfloor u + \log(\bar{w}_i - M) \rfloor$, play arm i for B more rounds.

3. *(line 5 of Mechanism 2) For all other rounds, report* 0.

We use this to show that under any $o(T)$-Nash equilibrium, the principal receives $\mu' T - o(T)$ revenue under Mechanism 2.

Corollary 16 *Under Mechanism 2, the principal will receive revenue at least $\mu' T - o(T)$ whenever arms play according to an ϵ-Nash equilibrium, where μ' is the second largest mean in the set of means μ_i and $\epsilon = o(T)$.*

The dominant strategy in Lemma 15, as written, requires the arms to know their own means μ_i (in particular for step 1). However, if the arms don't initially know their means, they can instead simply report their value (plus M) each round, and still report a total sum of $(\mu_i + M)B$ in expectation. This no longer results in a strictly dominant strategy, but instead an $o(T)$-dominant strategy.

Lemma 17 *The following strategy is a prior-independent $o(T)$-dominant strategy for arm i in Mechanism 2:*

1. *(line 1 of Mechanism 2) For each round t in the first B rounds, report $v_{i,t} + M$.*

2. *(lines 3,4 of Mechanism 2) If $i = i^*$, for the R rounds that the principal expects to see reported value w', report the value $w' - M$. For the B bonus rounds, report 0. If $i \neq i^*$, report 0.*

3. *(line 5 of Mechanism 2) For all other rounds, report* 0.

It is an interesting question whether a more clever stochastic bandit algorithm can be embedded without destroying dominant strategies, and also whether a solution exists in exact dominant strategies for this model.

Similarly, the dominant strategy in Lemma 15 assumes we are in the unrestricted payment regime, because sometimes the value you must report (whether it is $\mu_i + M$ or $w' - M$) might be larger than the value received in that round. However, again it is possible to adapt the mechanism (by setting $M = 0$) and dominant strategy in Lemma 15 to work for arms in the restricted payment regime at the cost of transforming it into a $o(T)$-dominant strategy. To do this, arms (as in the previous paragraph) simply report their value each round in the first phase of the mechanism. In the second phase of the mechanism, instead of reporting w' each round, they again report their full value, until they have reported a total of Rw' (at which point they start reporting 0 for the rest of the game).

Finally, we consider the case when some arms are strategic and other arms are non-strategic. Importantly, the principal does not know which arms are strategic and which are non-strategic. We show in this case that the principal can get (per round) the larger of the largest mean of the non-strategic arms and the second largest mean of the strategic arms.

Theorem 18 *If the strategic arms all play according to in Lemma 15, then the principal will get at least* $\max(\mu_s, \mu_n)T - o(T)$ *with probability* $1 - o(1/T)$*. Here* μ_s *is the second largest mean of the strategic arms and* μ_n *is the largest mean of the non-strategic arms.*

5. Conclusions and Future Directions

We consider the multi-armed bandit problem with strategic arms: arms obtain a reward when pulled and may pass any of it onto the principal. Our first main result shows that treating this purely as a learning problem results in undesirable approximate Nash equilibria for the principle (guaranteeing only $o(T)$ reward over T rounds). Our second main result shows that a careful combination of auctions, learning, and scoring rules provides a learning algorithm such that every approximate Nash equilibrium guarantees the principal $\Omega(T)$ reward (and even better - the arms have a dominant strategy). Still, we are far from understanding the complete picture of multi-armed bandit problems in strategic settings. Many questions remain, both in our model and related models.

One limitation of our negative results is that they only show there exists some 'bad' approximate Nash equilibrium for the arms, i.e., one where any low-regret principal receives little revenue. This, however, says nothing about the space of all approximate Nash equilibria. Does there exist a low-regret mechanism for the principal along with an approximate Nash equilibria for the arms where the principal extracts significant utility? An affirmative answer to this question would raise hope for the possibility of a mechanism that can perform well in both the adversarial and strategic setting, whereas a negative answer would strengthen our claim that these settings are fundamentally at odds.

One limitation of our positive results is that all of the learning takes place at the beginning of the protocol. As a result, our mechanism fails in cases where the arms' distributions can change over time. Is it possible to design good mechanisms for such settings? Ideally, any good mechanism should learn the arms' values continually throughout the T rounds, but accommodating this would require novel tools to handle incentives.

Throughout this paper, whenever we consider strategic bandits we assume their rewards are stochastically generated. Can we say anything about strategic bandits with adversarially generated rewards? The key barrier here seems to be defining what a strategic equilibrium is in this case - arms need some underlying priors to reason about their future expected utility.

Finally, there are other quantities one may wish to optimize instead of the utility of the principal. For example, is it possible to design an efficient principal, who almost always picks the best arm (even if the arm passes along little to the principal)? Theorem 21 implies the answer is no if the principal also has to be efficient in the adversarial case, but are there other models where we can answer this question affirmatively?

References

Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Learning prices for repeated auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, pages 1169–1177, 2013.

Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Repeated contextual auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, pages 622–630, 2014.

Masaki Aoyagi. Bid rotation and collusion in repeated auctions. *Journal of Economic Theory*, 112 (1):79–105, 2003. URL http://EconPapers.repec.org/RePEc:eee:jetheo:v:112:y:2003:i:1:p:79-105.

Masaki Aoyagi. Efficient collusion in repeated auctions with communication. *Journal of Economic Theory*, 134(1):61–92, 2007. URL http://EconPapers.repec.org/RePEc:eee:jetheo:v:134:y:2007:i:1:p:61-92.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1503–1510, New York, NY, USA, 2012. ACM. URL http://icml.cc/2012/papers/749.pdf.

Susan Athey and Kyle Bagwell. Optimal collusion with private information. *RAND Journal of Economics*, 32(3):428–65, 2001. URL http://EconPapers.repec.org/RePEc:rje:randje:v:32:y:2001:i:3:p:428-65.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, January 2003. ISSN 0097-5397. doi: 10.1137/S0097539701398375. URL http://dx.doi.org/10.1137/S0097539701398375.

Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms: Extended abstract. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, EC '09, pages 79–88, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-458-4. doi: 10.1145/1566374.1566386. URL http://doi.acm.org/10.1145/1566374.1566386.

Moshe Babaioff, Robert D. Kleinberg, and Aleksandrs Slivkins. Truthful mechanisms with implicit payment computation. In *Proceedings of the 11th ACM Conference on Electronic Commerce*, EC '10, pages 43–52, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-822-3. doi: 10.1145/1807342.1807349. URL http://doi.acm.org/10.1145/1807342.1807349.

Dirk Bergemann and Juuso Vlimki. Learning and strategic pricing. *Econometrica*, 64(5):1125–49, 1996. URL http://EconPapers.repec.org/RePEc:ecm:emetrp:v:64:y:1996:i:5:p:1125-49.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. doi: 10.1561/2200000024. URL http://dx.doi.org/10.1561/2200000024.

Sylvain Chassang. Calibrated incentive contracts. *Econometrica*, 81(5):1935–1971, 2013.

Nikhil R. Devanur and Sham M. Kakade. The price of truthfulness for pay-per-click auctions. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, EC '09, pages 99–106, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-458-4. doi: 10.1145/1566374.1566388. URL http://doi.acm.org/10.1145/1566374.1566388.

Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, pages 5–22, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2565-3. doi: 10.1145/2600057.2602897. URL http://doi.acm.org/10.1145/2600057.2602897.

J.C. Gittins and D.M. Jones. A dynamic allocation index for the sequential design of experiments. In J. Gani, editor, *Progress in Statistics*, pages 241–266. North-Holland, Amsterdam, 1974.

Paul Johnson and Jacques Robert. Collusion in a model of repeated auctions. Cahiers de recherche, Universite de Montreal, Departement de sciences economiques, 1999. URL http://EconPapers.repec.org/RePEc:mtl:montde:9909.

Sham M. Kakade, Ilan Lobel, and Hamid Nazerzadeh. Optimal dynamic mechanism design and the virtual-pivot mechanism. *Operations Research*, 61(4):837–854, 2013.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988 – 1012, 2014. URL http://EconPapers.repec.org/RePEc:ucp:jpolec:doi:10.1086/676597.

Jean-Jacques Laffont and David Martimort. *The Theory of Incentives: The Principal-Agent Model*. 2002.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22, March 1985. ISSN 0196-8858. doi: 10.1016/0196-8858(85)90002-8. URL http://dx.doi.org/10.1016/0196-8858(85)90002-8.

Randolph McAfee and John McMillan. Bidding rings. *American Economic Review*, 82(3):579–99, 1992. URL http://EconPapers.repec.org/RePEc:aea:aecrev:v:82:y:1992:i:3:p:579-99.

John McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42(9):654–655, 1956.

Hamid Nazerzadeh, Amin Saberi, and Rakesh Vohra. Dynamic cost-per-action mechanisms and applications to online advertising. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 179–188, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367522. URL http://doi.acm.org/10.1145/1367497.1367522.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952. URL http://www.projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.bams/1183517370.

Andrzej Skrzypacz and Hugo Hopenhayn. Tacit collusion in repeated auctions. *Journal of Economic Theory*, 114(1):153–169, 2004. URL http://EconPapers.repec.org/RePEc:eee:jetheo:v:114:y:2004:i:1:p:153-169.

Appendix A. Negative Results

In this section, we show that algorithms that achieve low-regret in the multi-armed bandits problem with adversarial values perform poorly in the strategic multi-armed bandits problem. Throughout this section, we will assume we are working in the restricted payment model (i.e., arms can only pass along a value $w_{i,t}$ that is at most $v_{i,t}$), but all proofs also work in the unrestricted payment model (and in fact are much easier there).

A.1. Tacit Observational Model

We begin by showing that in the tacit observational model, where arms cannot see the amounts passed on by other arms, it is still possible for the arms to collude and leave the principal with $o(T)$ revenue.

We begin by proving this result for the case of two arms, where the proof is slightly simpler.

Theorem 19 *Let mechanism M be a (ρ, δ)-low regret algorithm for the multi-armed bandit problem with two arms, where $\rho \leq T^{-2}$ and $\delta \geq \sqrt{T \log T}$. Then in the strategic multi-armed bandit problem under the tacit observational model, there exist distributions D_1, D_2 and an $O(\sqrt{T\delta})$-Nash Equilibrium where a principal using mechanism M gets at most $O(\sqrt{T\delta})$ revenue.*

Proof Let D_1 and D_2 be distributions with means μ_1 and μ_2 respectively, such that $|\mu_1 - \mu_2| \leq \max(\mu_1, \mu_2)/2$. Additionally, assume both D_1 and D_2 are supported on $[\sqrt{\delta/T}, 1]$. We now describe the equilibrium strategy S^* (the below description is for arm 1; S^* for arm 2 is symmetric):

1. Set parameters $B = 6\sqrt{T\delta}$ and $\theta = \sqrt{\frac{\delta}{T}}$.
2. Define $c_{1,t}$ to be the number times arm 1 is pulled in rounds $1, ..., t$. Similarly define $c_{2,t}$ to be the number times arm 2 is pulled in rounds $1, ..., t$.
3. For $t = 1, \ldots, T$:
   (a) If there exists a $t' \leq t - 1$ such that $c_{1,t'} < c_{2,t'} - B$, set $w_{1,t} = v_{1,t}$.
   (b) If the condition in (a) is not true, let $p_{1,t}$ be the probability that the principal will pick arm 1 in this round conditioned on the history (assuming player 2 is also playing S^*), and let $p_{2,t} = 1 - p_{1,t}$. Then:
      i. If $c_{1,t-1} < c_{2,t-1}$ and $p_{1,t} < p_{2,t}$, set $w_{1,t} = \theta$.
      ii. Otherwise, set $w_{1,t} = 0$.

We will now show that (S^*, S^*) is an $O(\sqrt{T\delta})$-Nash equilibrium. To do this, for any deviating strategy S', we will both lower bound $u_1(M, S^*, S^*)$ and upper bound $u_1(M, S', S^*)$, hence bounding the net utility of deviation.

We begin by proving that $u_1(M, S^*, S^*) \geq \frac{\mu_2 T}{2} - O(\sqrt{T\delta})$. We need the following lemma.

Lemma 20 *If both arms are using strategy S^*, then with probability $\left(1 - \frac{4}{T}\right)$, $|c_{1,t} - c_{2,t}| \leq B$ for all $t \in [T]$.*

Proof Assume that both arms are playing the strategy S^* with the modification that they never defect (i.e. condition (a) in the above strategy is removed). This does not change the probability that $|c_{1,t} - c_{2,t}| \leq B$ for all $t \in [T]$.

Define $R_{1,t} = \sum_{s=1}^{t} w_{1,s} - \sum_{s=1}^{t} w_{I_s,s}$ be the regret the principal experiences for not playing only arm 1. Define $R_{2,t}$ similarly. We will begin by showing that with high probability, these regrets are bounded both above and below. In particular, we will show that with probability at least $1 - \frac{2}{T}$, $R_{i,t}$ lies in $[-2\theta\sqrt{T \log T} - \delta, \delta]$ for all $t \in [T]$ and $i \in \{1, 2\}$.

To do this, note that there are two cases where the regrets $R_{1,t}$ and $R_{2,t}$ can possibly change. The first is when $p_{1,t} > p_{2,t}$ and $c_{1,t} > c_{2,t}$. In this case, the arms offer $(w_{1,t}, w_{2,t}) = (0, \theta)$. With probability $p_{1,t}$ the principal chooses arm 1 and the regrets update to $(R_{1,t+1}, R_{2,t+1}) = (R_{1,t}, R_{2,t} + \theta)$, and with probability $p_{2,t}$ the principal chooses arm 2 and the regrets update to $(R_{1,t+1}, R_{2,t+1}) = (R_{1,t} - \theta, R_{2,t})$. It follows that $\mathbb{E}[R_{1,t+1} + R_{2,t+1} | R_{1,t} + R_{2,t}] = R_{1,t} + R_{2,t} + (p_{1,t} - p_{2,t})\theta \geq R_{1,t} + R_{2,t}$.

In the second case, $p_{1,t} < p_{2,t}$ and $c_{2,t} < c_{1,t}$, and a similar calculation shows again that $\mathbb{E}[R_{1,t+1} + R_{2,t+1} | R_{1,t} + R_{2,t}] = R_{1,t} + R_{2,t} + (p_{2,t} - p_{1,t})\theta \geq R_{1,t} + R_{2,t}$. It follows that $R_{1,t} + R_{2,t}$ forms a submartingale.

From the above analysis, it is also clear that $|(R_{1,t+1} + R_{2,t+1}) - (R_{1,t} + R_{2,t})| \leq \theta$. It follows from Azuma's inequality that, for any fixed $t \in [T]$,

$$\Pr\left[R_{1,t} + R_{2,t} \leq -2\theta\sqrt{T \log T}\right] \leq \frac{1}{T^2}$$

Applying the union bound, with probability at least $1 - \frac{1}{T}$, $R_{1,t} + R_{2,t} \geq -2\theta\sqrt{T \log T}$ for all $t \in [T]$. Furthermore, since the principal is using a $\left(T^{-2}, \delta\right)$-low-regret algorithm, it is also true that with probability at least $1 - T^{-2}$ (for any fixed t) both $R_{1,t}$ and $R_{2,t}$ are at most δ. Applying the union bound again, it is true that $R_{1,t} \leq \delta$ and $R_{2,t} \leq \delta$ for all t with probability at least $1 - \frac{1}{T}$. Finally, combining this with the earlier inequality (and applying union bound once more), with probability at least $1 - \frac{2}{T}$, $R_{i,t} \in \left[-2\theta\sqrt{T \log T} - \delta, \delta\right]$, as desired. For the remainder of the proof, condition on this being true.

We next proceed to bound the probability that (for a fixed t) $c_{1,t} - c_{2,t} \leq B$. Define the random variable τ to be the largest value $s \leq t$ such that $c_{1,\tau} - c_{2,\tau} = 0$ – note that if $c_{1,t} - c_{2,t} \geq 0$, then $c_{1,s} - c_{2,s} \geq 0$ for all s in the range $[\tau, t]$. Additionally let Δ_s denote the ± 1 random variable given by the difference $(c_{1,s} - c_{2,s}) - (c_{1,s-1} - c_{2,s-1})$. We can then write

$$\begin{aligned} c_{1,t} - c_{2,t} &\leq \sum_{s=\tau+1}^{t} \Delta_s \\ &\leq \sum_{s=\tau+1}^{t} \Delta_s \cdot \mathbb{1}_{p_{1,s} > p_{2,s}} + \sum_{s=\tau+1}^{t} \Delta_s \cdot \mathbb{1}_{p_{1,s} \leq p_{2,s}} \end{aligned}$$

Here the first summand corresponds to times s where one of the arms offers θ (and hence the regrets change), and the second summand corresponds to times where both arms offer 0. Note that since $c_{1,s} \geq c_{2,s}$ in this interval, the regret $R_{2,s}$ increases by θ whenever $\Delta_s = 1$ (i.e., arm 1 is chosen), and furthermore no choice of arm can decrease $R_{2,s}$ in this interval. Since we know that $R_{2,s}$ lies in the interval $\left[-2\theta\sqrt{T\log T} - \delta, \delta\right]$ for all s, this bounds the first sum by

$$\sum_{s=\tau+1}^{t} \Delta_s \cdot \mathbb{1}_{p_{1,s} > p_{2,s}} \leq \frac{2\delta + 2\theta\sqrt{T\log T}}{\theta} = \frac{2\delta}{\theta} + 2\sqrt{T\log T}$$

On the other hand, when $p_{1,s} \leq p_{2,s}$, then $\mathbb{E}[\Delta_s] = p_{1,s} - p_{2,s} \leq 0$. By Hoeffding's inequality, it then follows that with probability at least $1 - \frac{1}{T^2}$,

$$\sum_{s=\tau+1}^{t} \Delta_s \cdot \mathbb{1}_{p_{1,s} \leq p_{2,s}} \leq 2\sqrt{T\log T}$$

Altogether, this shows that with probability at least $1 - \frac{1}{T^2}$,

$$c_{1,t} - c_{2,t} \leq \frac{2\delta}{\theta} + 4\sqrt{T\log T} \leq 6\sqrt{T\delta} = B$$

The above inequality therefore holds for all t with probability at least $1 - \frac{1}{T}$. Likewise, we can show that $c_{2,t} - c_{1,t} \leq B$ also holds for all t with probability at least $1 - \frac{1}{T}$. Since we are conditioned on the regrets $R_{i,t}$ being bounded (which is true with probability at least $\frac{2}{T}$), it follows that $|c_{1,t} - c_{2,t}| \leq B$ for all t with probability at least $1 - \frac{4}{T}$.

■

By Lemma 20, we know that with probability $1 - \frac{4}{T}$, $|c_{1,t} - c_{2,t}| \leq B$ throughout the mechanism. In this case, arm 1 never uses step (a), and $c_{1,T} \geq (T - B)/2$. Therefore

$$\begin{aligned}
u_1(M, S^*, S^*) &\geq \left(1 - \frac{4}{T}\right) \cdot (\mu_1 - \theta) \cdot (T - B)/2 \\
&\geq \frac{\mu_1 T}{2}\left(1 - \frac{4}{T} - \frac{\theta}{\mu_1} - \frac{B}{T}\right) \\
&= \frac{\mu_1 T}{2} - 2\mu_1 - \frac{\theta T}{2} - \frac{B\mu_1}{2} \\
&\geq \frac{\mu_1 T}{2} - O(\sqrt{T\delta}).
\end{aligned}$$

Now we will show that $u_1(M, S', S^*) \leq \frac{\mu_1 T}{2} + O(\sqrt{T\delta})$. Without loss of generality, we can assume S' is deterministic. Let M_R be the deterministic mechanism when M's randomness is fixed to some outcome R. Consider the situation when arm 1 is using strategy S', arm 2 is using strategy S^* and the principal is using mechanism M_R. There are two cases:

1. $c_{1,t} - c_{2,t} \leq B$ is true for all $t \in [T]$. In this case, we have

$$u_1(M_R, S', S^*) \leq c_{1,T} \cdot \mu_1 \leq \mu_1(T + B)/2.$$

2. There exists some t such that $c_{1,t}-c_{2,t} > B$: Let τ_R+1 be the smallest t such that $c_{1,t}-c_{2,t} > B$. We know that $c_{1,\tau_R} - c_{2,\tau_R} \leq B$. Therefore we have

$$
\begin{aligned}
u_1(M_R, S', S^*) &= \sum_{t=1}^{T}(\mu_1 - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \\
&= \sum_{t=1}^{T}(\mu_1 - w_{2,t}) \cdot \mathbb{1}_{I_t=1} + \sum_{t=1}^{T}(w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \\
&\leq c_{1,\tau_R}\mu_1 + \mu_1 + (T - \tau_R - 1)\max(\mu_1 - \mu_2, 0) + \sum_{t=1}^{T}(w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \\
&\leq \mu_1(\tau_R + B)/2 + \mu_1 + (T - \tau_R - 1)(\mu_1/2) + \sum_{t=1}^{T}(w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \\
&\leq \mu_1 T/2 + \mu_1(B+1)/2 + \sum_{t=1}^{T}(w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1}.
\end{aligned}
$$

In general, we thus have that

$$
u_1(M_R, S', S^*) \leq \mu_1 T/2 + \mu_1(B+1)/2 + \max\left(0, \sum_{t=1}^{T}(w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1}\right).
$$

Therefore

$$
\begin{aligned}
u_1(M, S', S^*) &= \mathbb{E}_R[u_1(M_R, S', S^*)] \\
&\leq \mu_1 T/2 + \mu_1(B+1)/2 + \mathbb{E}_R\left[\max\left(0, \sum_{t=1}^{T}(w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1}\right)\right].
\end{aligned}
$$

Notice that $\sum_{t=1}^{T}(w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1}$ is the regret of not playing arm 2 (i.e., R_2 in the proof of Lemma 20). Since the mechanism M is (ρ, δ) low regret, with probability $1-\rho$, this sum is at most δ (and in the worst case, it is bounded above by $T\mu_2$). We therefore have that:

$$
\begin{aligned}
u_1(M, S', S^*) &\leq \frac{\mu_1 T}{2} + \frac{\mu_1(B+1)}{2} + \delta + \rho T \mu_2 \\
&\leq \frac{\mu_1 T}{2} + O(\sqrt{T\delta})
\end{aligned}
$$

From this and our earlier lower bound on $u_1(M, S^*, S^*)$, it follows that $u_1(M, S', S^*) - u_1(M, S^*, S^*) \leq O(\sqrt{T\delta})$, thus establishing that (S^*, S^*) is an $O(\sqrt{T\delta})$-Nash equilibrium for the arms.

Finally, to bound the revenue of the principal, note that if the arms both play according to S^* and $|c_{1,t} - c_{2,t}| \leq B$ for all t (so they do not defect), the principal gets a maximum of $T\theta = O(\sqrt{T\delta})$ revenue overall. Since (by Lemma 20) this happens with probability at least $1 - \frac{4}{T}$ (and the total amount of revenue the principal is bounded above by T), it follows that the total expected revenue of the principal is at most $O(\sqrt{T\delta})$.

■

We now extend this proof to the K arm case, where K can be as large as $T^{1/3}/\log(T)$.

Theorem 21 *Let mechanism M be a (ρ, δ)-low regret algorithm for the multi-armed bandit problem with K arms, where $K \leq T^{1/3}/\log(T)$, $\rho \leq T^{-2}$, and $\delta \geq \sqrt{T \log T}$. Then in the strategic multi-armed bandit problem under the tacit observational model, there exist distributions D_i and an $O(\sqrt{KT\delta})$-Nash Equilibrium for the arms where the principal gets at most $O(\sqrt{KT\delta})$ revenue.*

Proof [Proof of Theorem 21] As in the proof of Theorem 19, let μ_i denote the mean value of the ith arm's distribution D_i (supported on $[\sqrt{K\delta/T}, 1]$). Without loss of generality, further assume that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. We will show that as long as $\mu_1 - \mu_2 \leq \frac{\mu_1}{K}$, there exists some $O(\sqrt{KT\delta})$-Nash equilibrium for the arms where the principal gets at most $O(\sqrt{KT\delta})$ revenue.

We begin by describing the equilibrium strategy S^* for the arms. Let $c_{i,t}$ denote the number of times arm i has been pulled up to time t. As before, set $B = 7\sqrt{KT\delta}$ and set $\theta = \sqrt{\frac{K\delta}{T}}$. The equilibrium strategy for arm i at time t is as follows:

1. If at any time $s \leq t$ in the past, there exists an arm j with $c_{j,s} - c_{i,s} \geq B$, defect and offer your full value $w_{i,t} = \mu_i$.

2. Compute the probability $p_{i,t}$, the probability that the principal will pull arm i conditioned on the history so far.

3. Offer $w_{i,t} = \theta(1 - p_{i,t})$.

We begin, as before, by showing that if all parties follow this strategy, then with high probability no one will ever defect.

Lemma 22 *If all arms are using strategy S^*, then with probability $\left(1 - \frac{3}{T}\right)$, $|c_{i,t} - c_{j,t}| \leq B$ for all $t \in [T], i, j \in [K]$.*

Proof As before, assume that all arms are playing the strategy S^* with the modification that they never defect. This does not change the probability that $|c_{i,t} - c_{j,t}| \leq B$ for all $t \in [T], i, j \in [K]$.

Define $R_{i,t} = \sum_{s=1}^{t} w_{i,s} - \sum_{s=1}^{t} w_{I_s,s}$ be the regret the principal experiences for not playing only arm i up until time t. We begin by showing that with probability at least $1 - \frac{2}{T}$, $R_{i,t}$ lies in $[-K\theta\sqrt{T \log T} - (K-1)\delta, \delta]$ for all $t \in [T]$ and $i \in [K]$.

To do this, first note that since the principal is using a (T^{-2}, δ)-low-regret algorithm, with probability at least $1 - T^{-2}$ the regrets $R_{i,t}$ are all upper bounded by δ at any fixed time t. Via the union bound, it follows that $R_{i,t} \leq \delta$ for all i and t with probability at least $1 - \frac{1}{T}$.

To lower bound $R_{i,t}$, we will first show that $\sum_{i=1}^{K} R_{i,t}$ is a submartingale in t. Note that, with probability $p_{j,t}$, $R_{i,t+1}$ will equal $R_{i,t} + \theta((1 - p_{j,t}) - (1 - p_{i,t}))$. We then have

$$
\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{K} R_{i,t+1} \middle| \sum_{i=1}^{K} R_{i,t}\right] &= \sum_{i=1}^{K} R_{i,t} + \sum_{i=1}^{K} p_{i,t} \sum_{j=1}^{K} \theta((1-p_{j,t}) - (1-p_{i,t})) \\
&= \sum_{i=1}^{K} R_{i,t} + \sum_{i=1}^{K} p_{i,t} \sum_{j=1}^{K} \theta(p_{i,t} - p_{j,t}) \\
&= \sum_{i=1}^{K} R_{i,t} + \theta \sum_{i=1}^{K} p_{i,t}(Kp_{i,t} - 1) \\
&= \sum_{i=1}^{K} R_{i,t} + \theta \left(K \sum_{i=1}^{K} p_{i,t}^2 - \sum_{i=1}^{K} p_{i,t}\right) \\
&\geq \sum_{i=1}^{K} R_{i,t}
\end{aligned}
$$

where the last inequality follows by Cauchy-Schwartz. It follows that $\sum_{i=1}^{K} R_{i,t}$ forms a submartingale.

Moreover, note that (since $|p_i - p_j| \leq 1$) $|R_{i,t+1} - R_{i,t}| \leq \theta$. It follows that $\left|\sum_{i=1}^{K} R_{i,t+1} - \sum_{i=1}^{K} R_{i,t}\right| \leq K\theta$ and therefore by Azuma's inequality that, for any fixed $t \in [T]$,

$$
\Pr\left[\sum_{i=1}^{K} R_{i,t} \leq -2K\theta\sqrt{T \log T}\right] \leq \frac{1}{T^2}.
$$

With probability $1 - \frac{1}{T}$, this holds for all $t \in [T]$. Since (with probability $1 - \frac{1}{T}$) $R_{i,t} \leq \delta$, this implies that with probability $1 - \frac{2}{T}$, $R_{i,t} \in \left[-2K\theta\sqrt{T \log T} - (K-1)\delta, \delta\right]$.

We next proceed to bound the probability that $c_{i,t} - c_{j,t} > B$ for a i, j, and t. Define

$$
S_t^{(i,j)} = \left(c_{i,t} - c_{j,t} + \frac{1}{\theta}(R_{i,t} - R_{j,t})\right).
$$

We claim that $S_t^{(i,j)}$ is a martingale. To see this, we first claim that $R_{i,t+1} - R_{j,t+1} = R_{i,t} - R_{j,t} - \theta(p_{i,t} - p_{j,t})$. Note that, if arm k is pulled, then $R_{i,t+1} = R_{i,t} + \theta((1-p_{i,t}) - (1-p_{k,t})) = R_{i,t} + \theta(p_{k,t} - p_{i,t})$ and similarly, $R_{j,t+1} = R_{j,t} + \theta(p_{k,t} - p_{j,t})$. It follows that $R_{i,t+1} - R_{j,t+1} = R_{i,t} - R_{j,t} - \theta(p_{i,t} - p_{j,t})$.

Secondly, note that (for any arm k) $\mathbb{E}[c_{k,t+1} - c_{k,t}|p_t] = p_{k,t}$, and thus $\mathbb{E}[c_{i,t+1} - c_{j,t+1} - (c_{i,t} - c_{j,t})|p_t] = p_{i,t} - p_{j,t}$. It follows that

$$
\begin{aligned}
\mathbb{E}[S_{t+1}^{(i,j)} - S_t^{(i,j)}|p_t] &= \mathbb{E}[(c_{i,t+1} - c_{j,t+1}) - (c_{i,t} - c_{j,t})|p_t] \\
&\quad + \frac{1}{\theta}\mathbb{E}[(R_{i,t+1} - R_{j,t+1}) - (R_{i,t} - R_{j,t})|p_t] \\
&= (p_{i,t} - p_{j,t}) - (p_{i,t} - p_{j,t}) \\
&= 0
\end{aligned}
$$

and thus that $\mathbb{E}[S_{t+1}^{(i,j)}|S_t^{(i,j)}] = S_t^{(i,j)}$, and thus that $S_t^{(i,j)}$ is a martingale. Finally, note that $|S_{t+1}^{(i,j)} - S_t^{(i,j)}| \leq 2$, so by Azuma's inequality

$$\Pr\left[S_t^{(i,j)} \geq 4\sqrt{T\log(TK)}\right] \leq (TK)^{-2}$$

Taking the union bound, we find that with probability at least $1-\frac{1}{T}$, $S^{(i,j)} \leq 4\sqrt{T\log(TK)}$ for all i, j, and t. Finally, since with probability at least $1-\frac{2}{T}$ each $R_{i,t}$ lies in $\left[-2K\theta\sqrt{T\log T}-(K-1)\delta, \delta\right]$, with probability at least $1-\frac{3}{T}$ we have that (for all i, j, and t)

$$\begin{aligned}
c_{i,t}-c_{j,t} &= S_t^{(i,j)} - \frac{1}{\theta}(R_{i,t}-R_{j,t}) \\
&\leq 4\sqrt{T\log(TK)} + \frac{1}{\theta}\left|R_{i,t}-R_{j,t}\right| \\
&\leq 4\sqrt{T\log(TK)} + 2K\sqrt{T\log T} + \frac{K\delta}{\theta} \\
&\leq \frac{7K\delta}{\theta} \\
&= 7K\sqrt{T\delta} \\
&= B
\end{aligned}$$

■

By Lemma 22, we know that with probability $1-\frac{3}{T}$, $|c_{i,t}-c_{j,t}| \leq B$ for all $t \in [T], i, j \in [K]$. In this case, arm 1 never defect, and $c_{1,T} \geq T/K - B$. Therefore

$$\begin{aligned}
u_1(M, S^*, S^*) &\geq \left(1-\frac{3}{T}\right)\cdot(\mu_1-\theta)\cdot(T/K-B) \\
&\geq \frac{\mu_1 T}{K}\left(1-\frac{3}{T}-\frac{\theta}{\mu_1}-\frac{BK}{T}\right) \\
&= \frac{\mu_1 T}{K} - 3\mu_1/K - \frac{\theta T}{K} - B\mu_1 \\
&\geq \frac{\mu_1 T}{K} - O(\sqrt{KT\delta})
\end{aligned}$$

Now we are going to show that $u_1(M, S', S^*) \leq \frac{\mu_1 T}{K} + O(\sqrt{KT\delta})$. Without loss of generality, we can assume S' is deterministic. Let M_R be the deterministic mechanism when M's randomness is fixed to some outcome R. Consider the situation when arm 1 is using strategy S', arm 2 is using strategy S^* and the principal is using mechanism M_R. There are two cases:

1. $c_{i,t} - c_{j,t} \leq B$ is true for all $t \in [T]$ and $i, j \in [K]$. In this case, we have

$$u_1(M_R, S', S^*) \leq c_{1,T}\cdot\mu_1 \leq \mu_1(T+(K-1)B)/K.$$

2. There exists some $t \in [T]$ and $i, j \in [K]$ such that $c_{i,t} - c_{j,t} > B$: Let $\tau_R + 1$ be the smallest t such that $c_{i,t} - c_{j,t} > B$ for some $i, j \in [K]$. We know that $c_{1,\tau_R} - c_{i,\tau_R} \leq B$ for all $i \in [K]$. Therefore we have

$$
\begin{aligned}
u_1(M_R, S', S^*) &= \sum_{t=1}^{T} (\mu_1 - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \\
&= \sum_{t=1}^{T} (\mu_1 - w_{2,t}) \cdot \mathbb{1}_{I_t=1} + \sum_{t=1}^{T} (w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \\
&\leq c_{1,\tau_R}\mu_1 + \mu_1 + (T - \tau_R - 1)\max(\mu_1 - \mu_2, 0) + \sum_{t=1}^{T} (w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \\
&\leq \mu_1(\tau_R + B)/K + \mu_1 + (T - \tau_R - 1)(\mu_1/K) + \sum_{t=1}^{T} (w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \\
&\leq \mu_1 T/K + \mu_1(B+1)(K-1)/K + \sum_{t=1}^{T} (w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1}.
\end{aligned}
$$

In M_R, we also have

$$
\begin{aligned}
\sum_{t=1}^{T} (w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1} &= \sum_{t=1}^{T} (w_{2,t} - w_{I_t,t}) - \sum_{t=1}^{T} (w_{2,t} - w_{I_t,t}) \cdot \mathbb{1}_{I_t \neq 1} \\
&\leq \sum_{t=1}^{T} (w_{2,t} - w_{I_t,t}) + \sum_{t=1}^{\tau_R} w_{I_t,t} \cdot \mathbb{1}_{I_t \neq 1} - \sum_{t=\tau_R+1}^{T} (\mu_2 - \mu_{I_t}) \cdot \mathbb{1}_{I_t \neq 1} \\
&\leq \sum_{t=1}^{T} (w_{2,t} - w_{I_t,t}) + T(\theta + B/T) + 0.
\end{aligned}
$$

In general, we thus have that

$$
u_1(M_R, S', S^*) \leq \mu_1 T/K + \mu_1(B+1)(K-1)/K + \max\left(0, \sum_{t=1}^{T} (w_{2,t} - w_{I_t,t}) + T\theta + B\right).
$$

Therefore

$$
\begin{aligned}
u_1(M, S', S^*) &= \mathbb{E}_R[u_1(M_R, S', S^*)] \\
&\leq \mu_1 T/K + \mu_1(B+1)(K-1)/K + \mathbb{E}_R\left[\max\left(0, \sum_{t=1}^{T} (w_{2,t} - w_{I_t,t}) + T\theta + B\right)\right].
\end{aligned}
$$

Notice that $\sum_{t=1}^{T} (w_{2,t} - w_{I_t,t})$ is the regret of not playing arm 2. Since the mechanism M is (ρ, δ) low regret, with probability $1 - \rho$, this sum is at most δ (and in the worst case, it is bounded above by $T\mu_2$). We therefore have that:

$$\begin{aligned} u_1(M, S', S^*) &\leq \mu_1 T/K + \mu_1(B+1)(K-1)/K + \delta + \rho T\mu_+ T\theta + B \\ &\leq \frac{\mu_1 T}{K} + O(\sqrt{KT\delta}). \end{aligned}$$

From this and our earlier lower bound on $u_1(M, S^*, S^*)$, it follows that $u_1(M, S', S^*) - u_1(M, S^*, S^*) \leq O(\sqrt{KT\delta})$, thus establishing that (S^*, S^*) is an $O(\sqrt{KT\delta})$-Nash equilibrium for the arms.

Finally, to bound the revenue of the principal, note that if the arms both play according to S^* and $|c_{i,t} - c_{j,t}| \leq B$ for all $t \in [T], i, j \in [K]$ (so they do not defect), the principal gets a maximum of $T\theta = O(\sqrt{KT\delta})$ revenue overall. Since (by Lemma 20) this happens with probability at least $1 - \frac{3}{T}$ (and the total amount of revenue the principal is bounded above by T), it follows that the total expected revenue of the principal is at most $O(\sqrt{KT\delta})$. ■

While the theorems above merely claim that a bad set of distributions for the arms exists, note that the proofs above show it is possible to collude in a wide range of instances - in particular, any set of distributions which satisfy $\mu_1 - \mu_2 \leq \mu_1/K$. A natural question is whether we can extend the above results to show that it is possible to collude in any set of distributions.

One issue with the collusion strategies in the above proofs is that if $\mu_1 - \mu_2 > \mu_1/K$, then arm 1 will have an incentive to defect in any collusive strategy that plays all the arms evenly (arm 1 can report a bit over μ_2 per round, and make $\mu_1 - \mu_2$ every round instead of μ_1 every K rounds). One solution to this is to design a collusive strategy that plays some arms more than others in equilibrium (for example, playing arm 1 90% of the time). We show how to modify our result for two arms to achieve an arbitrary market partition and thus work over a broad set of distributions.

Theorem 23 *Let mechanism M be a (ρ, δ)-low regret algorithm for the multi-armed bandit problem with two arms, where $\rho \leq T^{-2}$ and $\delta \geq \sqrt{T \log T}$. Then, in the strategic multi-armed bandit problem under the tacit observational model, for* any *distributions D_1, D_2 of values for the arms (supported on $[\sqrt{\delta/T}, 1]$), there exists an $O(\sqrt{T\delta})$-Nash Equilibrium for the arms where a principal using mechanism M gets at most $O(\sqrt{T\delta})$ revenue.*

Unfortunately, it as not as easy to modify the proof of Theorem 21 to prove the same result for K arms. It is an interesting open question whether there exist collusive strategies for K arms that can achieve an arbitrary partition of the market.

Proof Let D_1 and D_2 be distributions with means μ_1 and μ_2 respectively, and both distributions supported on $[\sqrt{\delta/T}, 1]$. We now describe the equilibrium strategy S^* (the below description is for arm 1; S^* for arm 2 is symmetric):

1. Set parameters $B = 6\sqrt{T\delta}/\mu_2$ and $\theta = \sqrt{\frac{\delta}{T}}$.

2. Define $c_{1,t}$ to be the number times arm 1 is pulled in rounds $1, ..., t$. Similarly define $c_{2,t}$ to be the number times arm 2 is pulled in rounds $1, ..., t$.

3. For $t = 1, ..., T$.

   (a) If there exists a $t' \leq t - 1$ such that $c_{1,t'}/\mu_1 < c_{2,t'}/\mu_2 - B$, set $w_{1,t} = v_{1,t}$.

(b) If the condition in (a) is not true, let $p_{1,t}$ be the probability that the principal will pick arm 1 in this round conditioned on the history (assuming player 2 is also playing S^*), and let $p_{2,t} = 1 - p_{1,t}$. Then:

   i. If $c_{1,t-1}/\mu_1 < c_{2,t-1}/\mu_2$ and $p_{1,t}/\mu_1 < p_{2,t}/\mu_2$, set $w_{1,t} = \theta$.

   ii. Otherwise, set $w_{1,t} = 0$.

We will now show that (S^*, S^*) is an $O(\sqrt{T\delta})$-Nash equilibrium. To do this, for any deviating strategy S', we will both lower bound $u_1(M, S^*, S^*)$ and upper bound $u_1(M, S', S^*)$, hence bounding the net utility of deviation.

We begin by proving that $u_1(M, S^*, S^*) \geq \frac{\mu_1^2 T}{\mu_1+\mu_2} - O(\sqrt{T\delta})$. We need the following lemma.

Lemma 24 *If both arms are using strategy S^*, then with probability $\left(1 - \frac{4}{T}\right)$, $|c_{1,t}/\mu_1 - c_{2,t}/\mu_2| \leq B$ for all $t \in [T]$.*

Proof Assume that both arms are playing the strategy S^* with the modification that they never defect (i.e. condition (a) in the above strategy is removed). This does not change the probability that $|c_{1,t}/\mu_1 - c_{2,t}/\mu_2| \leq B$ for all $t \in [T]$.

Define $R_{1,t} = \sum_{s=1}^{t} w_{1,s} - \sum_{s=1}^{t} w_{I_s,s}$ be the regret the principal experiences for not playing only arm 1. Define $R_{2,t}$ similarly. We will begin by showing that with high probability, these regrets are bounded both above and below. In particular, we will show that with probability at least $1 - \frac{2}{T}$, $R_{i,t}$ lies in $\left[-\frac{\mu_1}{\mu_2}(2\theta\sqrt{T\log T} + \delta), \delta\right]$ for all $t \in [T]$ and $i \in \{1, 2\}$.

To do this, note that there are two cases where the regrets $R_{1,t}$ and $R_{2,t}$ can possibly change. The first is when $p_{1,t}/\mu_1 > p_{2,t}/\mu_2$ and $c_{1,t}/\mu_1 > c_{2,t}/\mu_2$. In this case, the arms offer $(w_{1,t}, w_{2,t}) = (0, \theta)$. With probability $p_{1,t}$ the principal chooses arm 1 and the regrets update to $(R_{1,t+1}, R_{2,t+1}) = (R_{1,t}, R_{2,t} + \theta)$, and with probability $p_{2,t}$ the principal chooses arm 2 and the regrets update to $(R_{1,t+1}, R_{2,t+1}) = (R_{1,t} - \theta, R_{2,t})$. It follows that $\mathbb{E}[R_{1,t+1}/\mu_2 + R_{2,t+1}/\mu_1 | R_{1,t}/\mu_2 + R_{2,t}/\mu_1] = R_{1,t}/\mu_2 + R_{2,t}/\mu_1 + (p_{1,t}/\mu_1 - p_{2,t}/\mu_2)\theta \geq R_{1,t}/\mu_2 + R_{2,t}/\mu_1$.

In the second case, $p_{1,t}/\mu_1 < p_{2,t}/\mu_2$ and $c_{2,t}/\mu_1 < c_{1,t}/\mu_2$, and a similar calculation shows again that $\mathbb{E}[R_{1,t+1}/\mu_2 + R_{2,t+1}/\mu_1 | R_{1,t}/\mu_2 + R_{2,t}/\mu_1] = R_{1,t}/\mu_2 + R_{2,t}/\mu_1 + (p_{2,t}/\mu_2 - p_{1,t}/\mu_1)\theta \geq R_{1,t} + R_{2,t}$. It follows that $R_{1,t}/\mu_2 + R_{2,t}/\mu_1$ forms a submartingale.

From the above analysis, it is also clear that $|(R_{1,t+1}/\mu_2 + R_{2,t+1}/\mu_1) - (R_{1,t}/\mu_2 + R_{2,t}/\mu_1)| \leq \theta/\mu_2$. It follows from Azuma's inequality that, for any fixed $t \in [T]$,

$$\Pr\left[R_{1,t}/\mu_2 + R_{2,t}/\mu_1 \leq -\frac{2\theta}{\mu_2}\sqrt{T\log T}\right] \leq \frac{1}{T^2}$$

Applying the union bound, with probability at least $1 - \frac{1}{T}$, $R_{1,t}/\mu_2 + R_{2,t}/\mu_1 \geq -\frac{2\theta}{\mu_2}\sqrt{T\log T}$ for all $t \in [T]$. Furthermore, since the principal is using a $\left(T^{-2}, \delta\right)$-low-regret algorithm, it is also true that with probability at least $1 - T^{-2}$ (for any fixed t) both $R_{1,t}$ and $R_{2,t}$ are at most δ. Applying the union bound again, it is true that $R_{1,t} \leq \delta$ and $R_{2,t} \leq \delta$ for all t with probability at least $1 - \frac{1}{T}$. Finally, combining this with the earlier inequality (and applying union bound once more), with probability at least $1 - \frac{2}{T}$, $R_{i,t} \in \left[-\frac{\mu_1}{\mu_2}(2\theta\sqrt{T\log T} + \delta), \delta\right]$, as desired. For the remainder of the proof, condition on this being true.

We next proceed to bound the probability that (for a fixed t) $c_{1,t}/\mu_1 - c_{2,t}/\mu_2 \leq B$. Define the random variable $\tau - 1$ to be the largest value $s \leq t$ such that $c_{1,\tau}/\mu_1 - c_{2,\tau}/\mu_2 \leq 0$ – note that if

$c_{1,t}/\mu_1 - c_{2,t}/\mu_2 \geq 0$, then $c_{1,s}/\mu_1 - c_{2,s}/\mu_2 \geq 0$ for all s in the range $[\tau, t]$. Additionally let Δ_s denote the ± 1 random variable given by the difference $(c_{1,s}/\mu_1 - c_{2,s}/\mu_2) - (c_{1,s-1}/\mu_1 - c_{2,s-1}/\mu_2)$. We can then write

$$\begin{aligned} c_{1,t}/\mu_1 - c_{2,t}/\mu_2 &\leq \sum_{s=\tau+1}^{t} \Delta_s \\ &\leq \sum_{s=\tau+1}^{t} \Delta_s \cdot \mathbb{1}_{p_{1,s}/\mu_1 > p_{2,s}/\mu_2} + \sum_{s=\tau+1}^{t} \Delta_s \cdot \mathbb{1}_{p_{1,s}/\mu_1 \leq p_{2,s}/\mu_2} \end{aligned}$$

Here the first summand corresponds to times s where one of the arms offers θ (and hence the regrets change), and the second summand corresponds to times where both arms offer 0. Note that since $c_{1,s}/\mu_1 \geq c_{2,s}/\mu_2$ in this interval, the regret $R_{2,s}$ increases by θ whenever $\Delta_s = 1/\mu_1$ (i.e., arm 1 is chosen), and furthermore no choice of arm can decrease $R_{2,s}$ in this interval. Since we know that $R_{2,s}$ lies in the interval $\left[-\frac{\mu_1}{\mu_2}(2\theta\sqrt{T\log T} + \delta), \delta\right]$ for all s, this bounds the first sum by

$$\sum_{s=\tau+1}^{t} \Delta_s \cdot \mathbb{1}_{p_{1,s} > p_{2,s}} \leq \frac{\delta + \frac{\mu_1}{\mu_2}(2\theta\sqrt{T\log T} + \delta)}{\theta} \cdot (1/\mu_1) = \frac{1}{\mu_2}\left(\frac{2\delta}{\theta} + 2\sqrt{T\log T}\right)$$

On the other hand, when $p_{1,s}/\mu_1 \leq p_{2,s}/\mu_2$, then $\mathbb{E}[\Delta_s] = p_{1,s}/\mu_1 - p_{2,s}/\mu_2 \leq 0$. By Hoeffding's inequality, it then follows that with probability at least $1 - \frac{1}{T^2}$,

$$\sum_{s=\tau+1}^{t} \Delta_s \cdot \mathbb{1}_{p_{1,s} \leq p_{2,s}} \leq \frac{2}{\mu_2}\sqrt{T\log T}$$

Altogether, this shows that with probability at least $1 - \frac{1}{T^2}$,

$$c_{1,t} - c_{2,t} \leq \frac{1}{\mu_2}\left(\frac{2\delta}{\theta} + 4\sqrt{T\log T}\right) \leq 6\sqrt{T\delta}/\mu_2 = B$$

The above inequality therefore holds for all t with probability at least $1 - \frac{1}{T}$. Likewise, we can show that $c_{2,t}/\mu_2 - c_{1,t}/\mu_1 \leq B$ also holds for all t with probability at least $1 - \frac{1}{T}$. Since we are conditioned on the regrets $R_{i,t}$ being bounded (which is true with probability at least $\frac{2}{T}$), it follows that $|c_{1,t}/\mu_1 - c_{2,t}/\mu_2| \leq B$ for all t with probability at least $1 - \frac{4}{T}$.

■

By Lemma 20, we know that with probability $1 - \frac{4}{T}$, $|c_{1,t}/\mu_1 - c_{2,t}/\mu_2| \leq B$ throughout the mechanism. In this case, arm 1 never uses step (a), and $c_{1,T} \geq \frac{\mu_1}{\mu_1+\mu_2}T - \frac{\mu_1\mu_2}{\mu_1+\mu_2}B$. Therefore

$$\begin{aligned} u_1(M, S^*, S^*) &\geq \left(1 - \frac{4}{T}\right) \cdot (\mu_1 - \theta) \cdot \left(\frac{\mu_1}{\mu_1+\mu_2}T - \frac{\mu_1\mu_2}{\mu_1+\mu_2}B\right) \\ &\geq \frac{\mu_1^2 T}{\mu_1+\mu_2} - O(\sqrt{T\delta}) \end{aligned}$$

Now we will show that $u_1(M, S', S^*) \leq \frac{\mu_1^2 T}{\mu_1+\mu_2} + O(\sqrt{T\delta})$. Without loss of generality, we can assume S' is deterministic. Let M_R be the deterministic mechanism when M's randomness is fixed to some outcome R. Consider the situation when arm 1 is using strategy S', arm 2 is using strategy S^* and the principal is using mechanism M_R. There are two cases:

1. $c_{1,t}/\mu_1 - c_{2,t}/\mu_2 \leq B$ is true for all $t \in [T]$. In this case, we have

$$u_1(M_R, S', S^*) \leq c_{1,T} \cdot \mu_1 \leq \frac{\mu_1}{\mu_1+\mu_2} T + \frac{\mu_1\mu_2}{\mu_1+\mu_2} B.$$

2. There exists some t such that $c_{1,t}/\mu_1 - c_{2,t}/\mu_2 > B$: Let $\tau_R + 1$ be the smallest t such that $c_{1,t}/\mu_1 - c_{2,t}/\mu_2 > B$. We know that $c_{1,\tau_R}/\mu_1 - c_{2,\tau_R}/\mu_2 \leq B$. Therefore we have

$$\begin{aligned}
u_1(M_R, S', S^*) &= \sum_{t=1}^{T} (\mu_1 - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \\
&= \sum_{t=1}^{T} (\mu_1 - w_{2,t}) \cdot \mathbb{1}_{I_t=1} + \sum_{t=1}^{T} (w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \\
&\leq c_{1,\tau_R}\mu_1 + \mu_1 + (T - \tau_R - 1)\max(\mu_1 - \mu_2, 0) + \sum_{t=1}^{T} (w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \\
&\leq \mu_1 \left(\frac{\mu_1}{\mu_1+\mu_2}\tau_R + \frac{\mu_1\mu_2}{\mu_1+\mu_2} B \right) + \mu_1 + (T - \tau_R - 1)\frac{q_1^2}{\mu_1+\mu_2} \\
&\quad + \sum_{t=1}^{T} (w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \\
&\leq \frac{\mu_1^2}{\mu_1+\mu_2} T + \frac{\mu_1\mu_2}{\mu_1+\mu_2} B + \mu_1 + \sum_{t=1}^{T} (w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1}.
\end{aligned}$$

In general, we thus have that

$$u_1(M_R, S', S^*) \leq \frac{\mu_1^2}{\mu_1+\mu_2} T + \frac{\mu_1\mu_2}{\mu_1+\mu_2} B + \mu_1 + \max\left(0, \sum_{t=1}^{T} (w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1}\right).$$

Therefore

$$\begin{aligned}
u_1(M, S', S^*) &= \mathbb{E}_R[u_1(M_R, S', S^*)] \\
&\leq \frac{\mu_1^2}{\mu_1+\mu_2} T + \frac{\mu_1\mu_2}{\mu_1+\mu_2} B + \mu_1 + \mathbb{E}_R\left[\max\left(0, \sum_{t=1}^{T} (w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1}\right)\right].
\end{aligned}$$

Notice that $\sum_{t=1}^{T} (w_{2,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1}$ is the regret of not playing arm 2 (i.e., R_2 in the proof of Lemma 20). Since the mechanism M is (ρ, δ) low regret, with probability $1 - \rho$, this sum is at most δ (and in the worst case, it is bounded above by $T\mu_2$). We therefore have that:

$$\begin{aligned} u_1(M, S', S^*) &\leq \frac{\mu_1^2}{\mu_1+\mu_2}T + \frac{\mu_1\mu_2}{\mu_1+\mu_2}B + \mu_1 + \delta + \rho T\mu_2 \\ &\leq \frac{\mu_1^2}{\mu_1+\mu_2}T + O(\sqrt{T\delta}) \end{aligned}$$

From this and our earlier lower bound on $u_1(M, S^*, S^*)$, it follows that $u_1(M, S', S^*) - u_1(M, S^*, S^*) \leq O(\sqrt{T\delta})$, thus establishing that (S^*, S^*) is an $O(\sqrt{T\delta})$-Nash equilibrium for the arms.

Finally, to bound the revenue of the principal, note that if the arms both play according to S^* and $|c_{1,t}/\mu_1 - c_{2,t}/\mu_2| \leq B$ for all t (so they do not defect), the principal gets a maximum of $T\theta = O(\sqrt{T\delta})$ revenue overall. Since (by Lemma 20) this happens with probability at least $1 - \frac{4}{T}$ (and the total amount of revenue the principal is bounded above by T), it follows that the total expected revenue of the principal is at most $O(\sqrt{T\delta})$.

■

A.2. Explicit Observational Model

In this section we show that in the explicit observational model, there is an approximate equilibrium for the arms that results in the principal receiving no revenue. Since arms can view other arms' reported values, it is easy to collude in the explicit model; simply defect and pass along the full amount as soon as you observe another arm passing along a positive amount.

Theorem 25 *Let mechanism M be a δ-low regret algorithm for the multi-armed bandit problem. Then in the strategic multi-armed bandit problem under the explicit observational model, there exist distributions D_i and a $(\delta + 1)$-Nash equilibrium for the arms where a principal using mechanism M receives zero revenue.*

Proof Consider the two-arm setting where D_1 and D_2 are both deterministic distributions supported entirely on $\{1\}$, so that $v_{i,t} = 1$ for all $i = 1, 2$ and $t \in [T]$. Consider the following strategy S^* for arm i:

1. Set $w_{i,t} = 0$ if at time $1, ..., t-1$, the other arm always reports 0 when pulled.
2. Set $w_{i,t} = 1$ otherwise.

We will show that (S^*, S^*) is a $(\delta + 1)$-Nash Equilibrium. It suffices to show that arm 1 can get at most $\delta + 1$ more utility by deviating. Consider any deviating strategy S' for arm 1. By convexity, we can assume S' is deterministic (there is some best deterministic deviating strategy). Since mechanism M might be randomized, let R be the randomness used by M and define M_R to be the deterministic mechanism when M uses randomness R. Now, consider the case when arm 1 plays strategy S', arm 2 plays strategy S^* and the principal is usings mechanism M_R.

1. If arm 1 never reports any value larger than 0 when pulled, then S' behaves exactly the same as S^*. Therefore,

$$u_1(M_R, S', S^*) = u_1(M_R, S^*, S^*).$$

2. If arm 1 ever reports some value larger than 0 when pulled, let τ_R be the first time it does so. We know that S' behaves the same as S^* before τ_R. Therefore,

$$\begin{aligned} u_1(M_R, S', S^*) &\leq u_1(M_R, S^*, S^*) + \sum_{t=\tau_R}^{T} (v_{1,t} - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \\ &\leq u_1(M_R, S^*, S^*) + 1 + \sum_{t=\tau_R+1}^{T} (\max(w_{1,t}, w_{2,t}) - w_{1,t}) \cdot \mathbb{1}_{I_t=1} \end{aligned}$$

So in general, we have

$$u_1(M_R, S', S^*) \leq u_i(M_R, S^*, S^*) + 1 + \sum_{t=\tau_R+1}^{T} (\max(w_{1,t}, w_{2,t}) - w_{1,t}) \cdot \mathbb{1}_{I_t=1}.$$

Therefore

$$\begin{aligned} u_1(M, S', S^*) &= \mathbb{E}_R[u_1(M_R, S', S^*)] \\ &\leq \mathbb{E}_R[u_1(M_R, S^*, S^*)] + 1 + \mathbb{E}_R\left[\sum_{t=\tau_R+1}^{T} (\max(w_{1,t}, w_{2,t}) - w_{1,t}) \cdot \mathbb{1}_{I_t=1}\right] \\ &= u_1(M, S^*, S^*) + 1 + \mathbb{E}_R\left[\sum_{t=\tau_R+1}^{T} (\max(w_{1,t}, w_{2,t}) - w_{1,t}) \cdot \mathbb{1}_{I_t=1}\right]. \end{aligned}$$

Notice that this expectation is at most the regret of M in the classic multi-armed bandit setting when the adversary sets rewards equal to the values $w_{1,t}$ and $w_{2,t}$ passed on by the arms when they play (S', S^*). Therefore, by our low-regret guarantee on M, we have that

$$\mathbb{E}_R\left[\sum_{t=\tau_R+1}^{T} (\max(w_{1,t}, w_{2,t}) - w_{1,t}) \cdot \mathbb{1}_{I_t=1}\right] \leq \delta.$$

Thus

$$u_1(M, S', S^*) \leq u_1(M, S^*, S^*) + 1 + \delta$$

and this is a $(1+\delta)$-approximate Nash equilibrium. Finally, it is easy to check that the principal receives zero revenue when both arms play according to this equilibrium strategy. ■

Appendix B. Omitted Results and Proofs of Section 4

B.1. All Strategic Arms with Stochastic Values

Proof [Proof of Lemma 12] Note that the mechanism is naturally divided into three parts (in the same way the strategy above is divided into three parts): (1) the start, where each arm is played once and reports its mean, (2) the middle, where the principal plays the best arm and extracts the second-best arm's value (and plays each other arm once), and (3) the end, where the principal plays each arm some number of times, effectively paying them off for responding truthfully in step (1). To show the above strategy is dominant, we will proceed by backwards induction, showing that each part of the strategy is the best conditioned on an arbitrary history.

We start with step (3). It is easy to check that these rounds don't affect how many times the arm is played or not. It follows that it is strictly dominant to just report 0 (and receive your full value for the turn). Note that the reward the arm receives in expectation for this round is $(u + \log(w_i))\mu_i$; we will use this later.

For step (2), assume that $i = i^*$; otherwise, arm i is played only once, and the dominant strategy is to report 0 and receive expected reward μ_i. Depending on what happened in step (1), there are two cases; either $w' \le \mu_i$, or $w' > \mu_i$. We will show that if $w' \le \mu_i$, the arm should play w' for the next R rounds (not defecting) and report 0 for the bonus round. If $w' > \mu_i$, the arm should play 0 (defecting immediately).

Note that we can recast step (2) as follows: arm i starts by receiving a reward from his distribution D_i. For the next R turns, he can pay w' for the privilege of drawing a new reward from his distribution (ending the game immediately if he refuses to pay). If $w' \le \mu_i$, then paying for a reward w' is positive in expectation, whereas if $w' > \mu_i$, then paying for a reward is negative in expectation. It follows that the dominant strategy is to continue to report w' if $w' \le \mu_i$ (receiving a total expected reward of $R(\mu_i - w') + \mu_i$) and to immediately defect and report 0 if $w' > \mu_i$ (receiving a total expected reward of μ_i).

Finally, we analyze step (1). We will show that, regardless of the values reported by the other players, it is a dominant strategy for arm i to report its true mean μ_i. If arm i reports w_i, and $i \neq i^*$, then arm i will receive in expectation reward

$$G = (\mu_i - w_i) + \mu_i + \max(u + \log(w_i), 0)\mu_i$$

If $u + \log(w_i) > 0$, then this is maximized when $w_i = \mu_i$ and $G = (u + \log(\mu_i) + 1)\mu_i$ (note that by our construction of u, $u + \log(\mu_i) \ge 1$). On the other hand, if $u + \log(w_i) \le 0$, then this is maximized when $w_i = 0$ and $G = 2\mu_i$. Since $u + \log(\mu_i) + 1 \ge 2$, the overall maximum occurs at $w_i = \mu_i$.

Similarly, when arm i reports w_i and $i = i^*$, then arm i receives in expectation reward

$$G' = (\mu_i - w_i) + \max(0, R(\mu_i - w')) + \mu_i + \max(u + \log(w_i), 0)\mu_i$$

which is similarly maximized at $w_i = \mu_i$. Finally, it follows that if $\mu_i \le w'$, $G = G'$, so it is dominant to report $w_i = \mu_i$. On the other hand, if $\mu_i > w'$, then reporting $w_i = \mu_i$ will ensure $i = i^*$ and so once again it is dominant to report $w_i = \mu_i$. ■

Proof [Proof of Lemma 14] Suppose there exists an truthful mechanism A guarantees $(\alpha\mu + (1 - \alpha)\mu')T$ revenue for any distributions. We will show this results in a contradiction.

We now consider $L > \exp(1/\alpha)$ inputs. The i-th input has $\mu = b_i = 1/2 + i/(2L)$ and $\mu' = 1/2$. Among these inputs, one arm (call it arm j^*) is always the arm with largest mean and another arm is always the arm with the second largest mean. Other arms have the same input distribution in all the inputs.

Consider all the arms are using their dominant strategies. For the i-th input, let x_iT be the expected number of pulls by A on the arm k^* and p_iT be the expected amount arm k^* gives to the principal. Because the mechanism is truthful, in the i-th distribution, arm k^* prefers its dominant strategy than the dominant strategy it uses in some j-th distribution ($i \neq j$). In other words, we have for $i \neq j$,

$$b_i x_i - p_i \geq b_i x_j - p_j.$$

We also have, for all i,

$$b_i x_i - p_i \geq 0.$$

By using these inequalities , we get for all i,

$$p_i \leq b_i x_i + \sum_{j=1}^{i-1} x_j (b_{j+1} - b_j).$$

On the other hand, A's revenue in the i-th distribution is at most $(p_i + (1 - x_i)\mu')T$. Therefore we have, for all i,

$$p_i + (1 - x_i)\mu' \geq \alpha \cdot b_i + (1 - \alpha)\mu'.$$

So we get

$$(1 - x_i)\mu' + b_i x_i + \sum_{j=1}^{i-1} x_j (b_{j+1} - b_j) \geq \alpha \cdot b_i + (1 - \alpha)\mu'.$$

It can be simplified as

$$x_i \geq \alpha + \sum_{j=1}^{i-1} x_j \frac{b_{j+1} - b_j}{b_i - \mu'} = \alpha + \frac{1}{i} \cdot \sum_{j=1}^{i-1} x_j.$$

By induction we get for all i,

$$x_i \geq \alpha \sum_{j=1}^{i} \frac{1}{i} > \alpha \ln(i).$$

Therefore we have

$$x_L > \alpha \ln(L) \geq 1.$$

Here we get a contradiction. ■

B.2. Strategic and Non-strategic Arms with Stochastic Values

Proof [Proof of Lemma 15] Similarly as the proof of Lemma 12, the mechanism is divided into three parts: (1) the start, where each arm is played B times and reports its mean, (2) the middle, where the principal plays the best arm and extracts the second-best arm's value (and plays each other arm B times), and (3) the end, where the principal plays each arm some number of times, effectively paying them off for responding truthfully in step (1). To show the above strategy is dominant, we will proceed by backwards induction, showing that each part of the strategy is the best conditioned on an arbitrary history.

For step (3), similarly as the proof of Lemma 12, it is strictly dominant for the arm to report 0. The reward the arm receives in expectation for this step is $(u + \log(\bar{w}_i - M))\mu_i B$.

For step (2), assume that $i = i^*$; otherwise, arm i is played B times, and the dominant strategy is to report 0 and receive expected reward $\mu_i B$. Depending on what happened in step (1), there are two cases; either $w' - M \le \mu_i$, or $w' - M > \mu_i$. Similarly as the proof of Lemma 12, we know that if $w' - M \le \mu_i$, the arm should play $w' - M$ for the next R rounds (not defecting) and report 0 for B bonus rounds. If $w' - M > \mu_i$, the arm should play 0 (defecting immediately).

For step (1), similar as the proof of Lemma 12, the expected reward of arm i is either

$$G = (\mu_i - \bar{w}_i)B + B\mu_i + \max(u + \log(\bar{w}_i - M), 0)B\mu_i$$

or

$$G' = \max(0, R(\mu_i - w' + M)) + (\mu_i - \bar{w}_i)B + B\mu_i + \max(u + \log(\bar{w}_i - M), 0)B\mu_i$$

Using the same argument as the proof of Lemma 12, we know arm i's dominant strategy is to make $\bar{w}_i = \mu_i + M$. ∎

Proof [Proof of Lemma 17] The only difference between the strategy in this lemma and the strategy in Lemma 15 is the first step, where instead of the arm reporting their mean every round (which they don't necessarily know), they instead report their value every round. It suffices to show that the expected difference in utility between running the above strategy and the strategy in Lemma 15 is at most $o(T)$.

To do this, let $\bar{w}'_i = \frac{1}{B}\sum_{t=1}^{B}(v_t + M)$ be the average value reported in the first phase by this new strategy, and let $\bar{w}_i = \mu_i + M$ be the optimal average to report. Let $\delta = \bar{w}'_i - \bar{w}$. From the formulas for net utility in the proof of Lemma 15, we note that reporting $\bar{w}'_i$ in the first phase instead of $\bar{w}_i$ results in at most $T\delta$ less utility overall. On the other hand, since $\mathbb{E}[v_t] = \mu_i$ for all t, by the Chernoff bound,

$$\Pr\left[|\delta| > 2\sqrt{\log T/B}\right] \le 2\exp\left(\frac{1}{2}\left(2\sqrt{\frac{\log T}{B}}\right)^2 B\right) = \frac{2}{T^2}.$$

It follows that the expected difference in utility is at most

$$2\sqrt{\frac{\log T}{B}}T + \frac{2}{T^2}T = O(\epsilon^{-1/8}T^{5/8}) = o(T).$$

∎

Proof [Proof of Corollary 16] Note that the proof of Lemma 15 works regardless of the values of B and M, so the strategy described in Lemma 15 is still a dominant strategy here. In an ϵ-Nash equilibrium, each player plays according to a strategy which gives them at least ϵ less than their payoff in the dominant equilibrium. We will show that if this is the case, then the principal gets at most $K\epsilon$ less than their payoff in the dominant equilibrium; since $K\epsilon = o(T)$, this proves the theorem.

Recall that $B = 2\epsilon^{1/4}T^{3/4}/\mu_{min}$ and define $\gamma = \epsilon^{1/3}/T^{1/3}$. We first claim that, similarly as in the proof of Lemma 15, if $i = i^*$ and $(1+\gamma)\mu_i \ge w'$, then if arm i is playing according to an ϵ-Nash equilibrium, it will not defect. This follows from the fact that modifying arm i's strategy to start repeatedly reporting w' as soon as arm i would have defected under the original strategy increases arm i's payoff by at least $B\mu_i - R\gamma\mu_i \ge 2\epsilon^{1/4}T^{3/4} - \epsilon^{1/3}T^{2/3} \ge \epsilon$ in expectation (where the additional $B\mu_i$ term comes via the payoff from the bonus rounds).

We next show that, in any ϵ-Nash equilibrium, each arm i reports an average value $\bar{w}_i$ between $\mu_i(1-\gamma)+M$ and $\mu_i(1+\gamma)+M$ with high probability.

To do this, we define

$$G_\mu(w) = ((\mu - (w+M)) + \mu + \max(u + \log w, 0)\mu) \cdot B.$$

Note that $G_\mu(w)$ upper bounds the expected reward an arm with mean μ which reports $w+M$ can get from all rounds except the R rounds in line 4 (but including the potential bonus rounds). Moreover, by the proof of Lemma 15, for $w = \mu$, $G_\mu(\mu)$ exactly equals the expected reward (in these rounds) of an arm following the dominant strategy. We'll first show that if $w < \mu(1-\gamma)$, then $G_\mu(\mu) - G_\mu(w) \geq \epsilon^{11/12}T^{1/12}$.

First, if $u + \log(w - M) < 0$, then $G_\mu(w) \leq 2B\mu - BM$, but by the proof of Lemma 15, $G_\mu(\mu) = B(u + \log\mu + 1)\mu - BM$. Since $u + \log\mu + 1 > 3$, $G_\mu(\mu) - G_\mu(w) \geq B\mu \geq \epsilon^{1/4}T^{3/4} \geq \epsilon^{11/12}T^{1/12}$. We can thus assume $\max(u + \log w, 0) = u + \log w$. Under this assumption

$$G_\mu(w) = ((\mu - w) + \mu + (u + \log w)\mu) \cdot B.$$

Then, if $w \leq \mu(1-\gamma)$,

$$\begin{aligned} G_\mu(\mu) - G_\mu(w) &= B(\mu\log\mu - \mu - \mu\log w + w) \\ &= B\mu(\log(1-\gamma) - \gamma) \\ &\geq B\mu\gamma^2 \\ &\geq 2\epsilon^{1/4}T^{3/4}\epsilon^{2/3}T^{-2/3} \\ &\geq \epsilon^{11/12}T^{1/12} \end{aligned}$$

Similarly, if $w > \mu(1+\gamma)$, we have that $G_\mu(\mu) - G_\mu(w) \geq \epsilon^{11/12}T^{1/12}$. Now, in expectation over w, $\mathbb{E}_w[G_\mu(\mu) - G_\mu(w)] \leq \epsilon$; otherwise, this player could increase their expected total reward by at least ϵ by switching to the dominant strategy (note that a player's expected reward from the R rounds in line 4 can only increase by switching to the dominant strategy). From Markov's inequality, it follows that

$$\Pr_{w_i}\left[w_i \in [\mu_i(1-\gamma), \mu_i(1+\gamma)]\right] \geq 1 - (\epsilon/T)^{1/12}.$$

Via the union bound, it follows that the probability that each w_i belongs to the interval $[\mu_i(1-\gamma)+M, \mu_i(1+\gamma)+M]$ is at least $1 - K(\epsilon/T)^{1/12} \geq 1 - o(1)$. Note that if this is the case, arm i^* will not defect, since $(1+\gamma)\mu_{i^*} \geq w_{i^*} \geq w'$. In addition, note that $w' \geq (1-\gamma)\mu' + M$ (since the two largest means are larger than μ', the two largest reported values w will be at least $(1-\gamma)\mu' + M$). It follows in this case that the principal receives at least $(1-\gamma)\mu' R = \mu' T - o(T)$. Since this occurs with probability $1 - o(1)$, it follows that the principal receives at least $\mu' T - o(T)$ in expectation, as desired. ∎

Proof [Proof of Theorem 18] Recall that $B = 2\epsilon^{1/4}T^{3/4}/\mu_{min}$ and $M = 8B^{-1/2}\ln(KT)$. We first show that with high probability non-strategic arms' reported values don't deviate too much from their means.

For each non-strategic arm i, by the Chernoff bound,

$$\Pr[|\bar{w}_i - \mu_i| \geq M/2] \leq 2\exp(-(M/2)^2 B/2) \leq 1/(KT)^8$$

By the union bound, with probability $1 - o(1/T)$, all non-strategic arms i satisfy $|\bar{w}_i - \mu_i| \leq M/2$. From now on, we will assume we are in the case when $|\bar{w}_i - \mu_i| < M/2$, for all i such that arm i is a non-strategic arm.

In the proof of Corollary 16, we showed that any strategic arm i playing according to an ϵ-Nash equilbrium, will report in Line 1 an average value $\bar{w}_i$ between $(1-\gamma)\mu_i + M$ and $(1+\gamma)\mu_i + M$ with high probability, where $\gamma = o(1)$. Note that this guarantee holds even in the presence of non-strategic arms, as we only use the fact that any strategy an arm plays in an ϵ-Nash equilibrium has an expected value of at least ϵ less than their dominant strategy's expected value. With this, we can consider two possible cases:

- **Case 1**: Arm i^* is a strategic arm. Then $w' \geq (1-\gamma)\mu_s + M$ and $w' \geq \mu_n - M/2$, and also $\mu_{i^*} = w_{i^*} - M \geq w' - M$. So, from only the third step of Mechanism 2, the principal will get reward at least

$$\begin{aligned} & (w' - M)R = \max((1-\gamma)\mu_s, \mu_n - 3M/2)R \geq (1-\gamma)\max(\mu_s, \mu_n)R - 3MR/2 \\ \geq\ & (1-\gamma)\max(\mu_s, \mu_n)T - \max(\mu_s, \mu_n)(u+3)BK - 3MR/2 \\ =\ & \max(\mu_s, \mu_n)T - o(T). \end{aligned}$$

- **Case 2**: Arm i^* is a non-strategic arm. We know that $\mu_{i^*} \geq w_{i^*} - M/2 \geq (w' - M) + M/2$. By using the Chernoff bound and union bound again, we know that arm i^* will defect in the line three with probability at most $o(1/T)$. We also know that $\mu_{i^*} \geq w_{i^*} - M/2 \geq (1-\gamma)\mu_s + M - M/2$ and $\mu_{i^*} \geq w_{i^*} - M/2 \geq u_n - M/2 - M/2$. It follows via the same argument as Case 1 that the principal will get reward at least $\max(u_s, u_n)T - o(T)$.

■