Robust Performance Incentivizing Algorithms for Multi-Armed Bandits with Strategic Agents

Seyed A. Esmaeili[1], **Suho Shin**[2], **Aleksandrs Slivkins**[3]
[1] University of Chicago
[2] University of Maryland
[3] Microsoft Research New York City
esmaeili@uchicago.edu, suhoshin@umd.edu, slivkins@microsoft.com

Abstract

Motivated by applications such as online labor markets we consider a variant of the stochastic multi-armed bandit problem where we have a collection of arms representing strategic agents with different performance characteristics. The platform (principal) chooses an agent in each round to complete a task. Unlike the standard setting, when an arm is pulled it can modify its reward by absorbing it or improving it at the expense of a higher cost. The principle has to solve a mechanism design problem to incentivize the arms to give their best performance. However, since even with an effective mechanism agents may still deviate from rational behavior, the principal wants a robust algorithm that also gives a non-vacuous guarantee on the total accumulated rewards under non-equilibrium behavior. In this paper, we introduce a class of bandit algorithms that meet the two objectives of performance incentivization and robustness simultaneously. We do this by identifying a collection of intuitive properties that a bandit algorithm has to satisfy to achieve these objectives. Finally, we show that settings where the principal has no information about the arms' performance characteristics can be handled by combining ideas from second price auctions with our algorithms.

1 Introduction

Various settings of the web economy involve a platform (principal) who selects workers (agents) to complete tasks. These agents differ in their intrinsic performance abilities and can in addition exert further effort to improve their performance at the expense of a higher cost. Clearly, if the principal selects the top performing agents that give their best effort he would obtain higher revenue. To do that the principal has to incentivize higher quality performance through effective *mechanism design/performance incentivization*. Furthermore, a deeper look would lead to a second important consideration, namely *robustness*. While mechanism design can be effective in settings where the agents are entities like firms and corporations who exhibit high levels of rationality, when the agents are instead average individuals, deviations from rational behavior are known and observed to exist. Therefore, a more effective solution should not only incentivize higher quality contributions, but should also be robust; adaptively learning and adjusting to the agents' performance in order to generate a non-trivial amount of revenue even when the agents deviate from rational behavior.

We model this problem as a multi-armed bandit (MAB) problem where the agents (workers) are the arms. As in the stochastic MAB setting each arm has a reward (performance) distribution associated with it. However, we allow the agent to modify his rewards after the realization. More specifically, the agent is able to modify the reward by absorbing (stealing) the reward leading to a low quality output or he may improve (lift up) the reward to give a better output to the principal. Since real-life agents are heterogeneous in their skill level, in our model each agent has a specific mean for his distribution and can lift up his rewards to a specific (not necessarily unique) maximum value. To have a more realistic model of the agents, we assume that there always exists a subset of *honest* agents, who either do not modify their rewards or possibly improve them.

In this paper, we show how to simultaneously achieve both desiderata of performance incentivization and robustness. Specifically, we show that a class of MAB algorithms which we call *sharply-adaptive, monotonic, and fair (SAMF)* are both performance incentivizing and robust. As the name suggests, SAMF algorithms satisfy a collection of properties which we show to be both natural and intuitive. A SAMF algorithm leads to an equilibrium where agents essentially give their top performance leading the principal to obtain optimal revenue. In fact, under a mild assumption on the size of market (number of agents) we show that no sub-optimal revenue equilibria can exist when using SAMF algorithms. We further consider a setting where the top performance level across the agents is not known either to the principal or the agents. We show that this setting can be handled by combining second price auction methods with SAMF algorithms, with the final mechanism leading to high revenue at equilibrium. Importantly, when agents deviate from equilibrium behavior, SAMF algorithms are robust and achieve revenue at least at the level of the honest agent with the highest reward mean. An advantage of our approach is its modularity. That is, we identify a set of properties that a MAB algorithm needs to satisfy to be a SAMF algorithm. Interestingly, we show that the standard UCB and ϵ-greedy algorithms are examples of SAMF algorithms. In comparison to prior work, our setting allows for more complicated scenar-

ios where agents may deviate from strategic behavior and interestingly we show that our algorithms can still achieve non-vacuous guarentees despite that.

2 Related Work

There is an increasing attention in the multi-armed bandit literature to study the dynamics of algorithms in the presence of strategic behavior. The most relevant paper, to the best of our knowledge, is due to Feng, Parkes, and Xu (2020), which considers a problem in which each arm can increase its reward given a budget constraint over the time horizon. Importantly however, the improvements the agents add to the reward in that setting are fake (strategic manipulations) although the principal does not realize that. The main objective of that paper is to see how susceptible MAB algorithms are to such manipulations. The main result of that paper is that standard algorithms like UCB, ϵ-greedy, and Thompson sampling are intrinsically robust to those strategic manipulations given that sub-optimal arms are equipped with a sublinear amount of manipulation budget. This is different from our model since an increase in the reward in our setting brings a real improvement and the principal wants to induce good equilibria where essentially all agents perform at their top level.

Braverman et al. (2019) also study a problem of a similar flavor to ours. They consider each arm as a strategic agent who wants to maximize the total rewards she obtains through the horizon. Once an arm is pulled, the corresponding reward is realized and the agent can strategically take a certain fraction of the reward and deliver the remainder to the principal. Each agent here tries to maximize the cumulative reward by balancing the trade-off between (i) increasing short-term revenue by taking large fractions of the rewards and (ii) increasing the probability to be selected by taking a small fraction of the rewards. They show that there are instances where any adversarial MAB algorithm would obtain a total of sublinear rewards (revenue). Essentially, the core cause behind this is that the arms can possibly engage in tacit collusion and therefore agree to take most of the rewards and send the principal a sublinear amount. In our setting, we assume a blind observation model (see Subsection 3.1 for the definition) which makes the arms unable to collude to achieve such strategies since they cannot see the arm selections. Further, Braverman et al. (2019) construct an algorithm that recovers the second-highest mean of the arms based on a second-price auction algorithm. Our algorithm under a certain regime (of information) is inspired by their algorithm, but is more involved as we combine it with a MAB algorithm. This makes our algorithm (unlike theirs) robust, i.e., having a minimum (non-vacuous) revenue guarantee even when the arms follow non-equilibrium behaviour.

There also exists a line of literature that considers strategic behavior in user-generated content platforms like Quora. Jain, Chen, and Parkes (2009) study how to incentivize strategic users in user-generated content platforms to contribute their own content immediately instead of postponing them. Ghosh and Hummel (2013), preceded by Ghosh and Hummel (2011); Ghosh and McAfee (2011), consider a problem of incentivizing high-quality contributions in user-generated content platforms, where each user wants to maximize her utility by exerting an optimized level of effort in constructing her content. A major difference from our model is that in the previous papers the user's strategic choice is static rather than dynamic and made in a one-shot manner when she registers the content (mean of the arm). Another set of papers are focused on capturing different aspects of the strategic interaction in the multi-armed bandit problem - such as Kremer, Mansour, and Perry (2014); Mansour, Slivkins, and Syrgkanis (2015); Bahar et al. (2020) of incentivized exploration, Shin, Lee, and Ok (2022); Esmaeili, Hajiaghayi, and Shin (2023) of strategic replication, Wang and Huang (2018) of known-compensation - however, we do not discuss the details as their models are significantly different from ours.

Our model is also conceptually related to the problem of incentivizing strategic workers to exert more costly effort also known as contract theory Holmström (1979). Designing an optimal and robust contract has been studied in various papers such as Carroll (2015), Dütting, Roughgarden, and Talgam-Cohen (2019), and Castiglioni, Marchesi, and Gatti (2022), and even in a repeated environment such as Rogerson (1985) and Chandrasekher (2015). Our model might be interpreted as a repeated contract design with multiple agents. Our model, however, does not assume a random mapping between the agent's action and the outcome, and precludes monetary transfers but algorithmically incentivizes the agents to behave more in favor of the principal.

3 Model, Objectives, and Main Results

3.1 Model Details

We model the agents participating in a platform as a set of k arms $\mathcal{A}$[1]. The interaction runs over a collection of rounds from $1, 2, \ldots, n$ with n being the *horizon*. The principal (algorithm) would like to obtain top level performance from the arms while the arms would like to be pulled (selected) as many times as possible while spending as little effort as possible on improving their performance.

Reward Modification: At round t, the algorithm chooses/pulls an arm (agent) I_t and a reward $r_{I_t}(t) \in [0, 1]$ is sampled from the static distribution $\mathcal{D}_{I_t}$ whose mean is μ_{I_t}. After seeing the reward, arm I_t can modify the reward by adding a value of effort $c_{I_t}(t)$ leading to the final reward $\tilde{r}_{I_t}(t) = r_{I_t}(t) + c_{I_t}(t)$. Note that the algorithm only observes the reward $\tilde{r}_{I_t}(t)$. If $c_{I_t}(t) \geq 0$, then the reward is improved whereas if $c_{I_t}(t) < 0$ then the reward is degraded. Clearly, $c_{I_t}(t) < 0$ implies that arm I_t has absorbed (a part of) its reward. Reward absorption is used to model the fact that in online labor markets a worker may intentionally worsen his performance[2]. Note that since we assume a setting where all realized and modified rewards are in $[0, 1]$

[1] For simplicity, we assume that k is a constant.

[2] I.e., the worker may submit a low quality or even random output on the task to minimize the cost and time spent on it. Note that low quality output is often found in online labor markets, see for example (Ipeirotis, Provost, and Wang 2010; Wais et al. 2010).

then we always have $1-r_{I_t}(t) \geq c_{I_t}(t) \geq -r_{I_t}(t)$. We refer to the total effort spent by arm i up to and including a given round t by $C_i(t) = \sum_{s=1}^{t} c_i(s)$. Given agent i her strategy S_i amounts to the values of $c_i(t)$ she chooses whenever she is pulled. Note that the agent's ability to modify the reward after seeing the realization gives the agent more control and follows related previous work such as (Feng, Parkes, and Xu 2020; Braverman et al. 2019).

Utility: First, we denote the number of pulls arm i receives up to and including round t by $T_i(t)$. Since the arm would like to be pulled more and spend as little cost as possible. A natural form for the utility is the following:

$$u_i = \mathbb{E}[T_i(n)] - \mathbb{E}[C_i(n)] \tag{1}$$

Clearly, from the previous paragraph on reward modification we see that reward absorption would lead to higher utility provided the arm still receives the same number of pulls while the reverse is true for reward improvement. This is in agreement with the fact that high quality effort is more costly. Note we may also generalize this by using a cost function leading to a utility of $u_i = \mathbb{E}[T_i(n)] - \mathbb{E}[\sum_{t=1}^{n} f_i(c_i(t))]$ where $f_i(.)$ is a cost function specific to agent i. In the main body we mostly consider $f_i(c_i(t)) = c_i(t)$. In Appendix **??** we discuss how some results generalize to a wider family of cost functions.

Heterogeneous Characteristics of the Arms: It is natural to expect that the agents in the platform will have different levels of productivity which we model by allowing different characteristics for each agent (arm). Specifically, in addition to arms having different reward means, we assume that each arm i can lift its reward to a maximum value of M_i. Further, $Support(\mathcal{D}_i) \subset [0, M_i]$, this implies that $\mu_i \leq M_i$ and that if i is pulled at round t then $c_i(t) \leq M_i - r_i(t)$. In general, given two different arms i and i' then we might have $M_i \neq M_{i'}$. We refer to the highest possible max reward value (*top performance*) by $M_{\text{top}} = \max_{i\in[k]} M_i$. Further, the set of top performing arms is $\mathcal{A}_{\text{top}} = \{i \in [k] : M_i = M_{\text{top}}\}$ and their cardinality is $k_{\text{top}} = |\mathcal{A}_{\text{top}}|$. Moreover, we assume that the performance levels are not dependent on the horizon. I.e., $\forall i \in [k] : M_i = \Theta(1)$.

Honest Agents: While it is true that in game-theoretic settings all agents are assumed to be rational utility maximizers (Roughgarden 2010; Osborne et al. 2004), in reality we find that humans may deviate from utility maximizing strategies and possibly even be altruistic (Camerer 1997; Gintis et al. 2003). Therefore, in online labor markets which tend to include many participating agents it is natural to assume that a subset of the agents are *honest*, meaning that they perform according to their productivity level and do not attempt to exploit the platform or give lower quality outputs. More formally, we assume that there always exists a set of *honest agents* $\mathcal{A}_{\mathcal{H}} \subset \mathcal{A}$ with $|\mathcal{A}_{\mathcal{H}}| \geq 1$ who always spend non-negative effort whenever they are pulled. I.e., if an honest agent i is pulled at around t, then $c_i(t) \geq 0$ and therefore $\tilde{r}_i(t) \geq r_i(t)$. Note that an honest arm can follow any strategy as long as the effort is non-negative ($c_i(t) \geq 0$). Further, we denote the maximum mean of reward distributions in the set of honest agents by $\mu^*_{\mathcal{H}}$ and the honest agent with that maximum mean by h^*. I.e., $\mu^*_{\mathcal{H}} = \max_{i\in\mathcal{A}_{\mathcal{H}}} \mu_i$ and $h^* = \{i \in \mathcal{A}_{\mathcal{H}} | \mu_i = \mu^*_{\mathcal{H}}\}$. Moreover, in settings where we ask agents to "report"[3] their top performance value M_i, we assume that honest agents would always report the value truthfully.

Looking ahead, the guarantees of our algorithms still hold without honest agents under equilibrium settings obtaining essentially top revenue, but we cannot guarantee any robustness when agents deviate from equilibrium behavior. The latter are highly desirable due to the possibility of irrational behavior and mistakes. Further, we only make a mild assumption: having at least one honest agent, and will guarantee a strong "robustness objective" (see Sections 3.3 and 4) whereby our algorithms compete with the best honest agent in the wors case.

Blind Observations: Unlike other settings such as advertisements, in online labor markets information about the decisions made by the platform are often unavailable or concealed. This is clearly the case in platforms such as Amazon Mechanical Turk or Uber where the agents do not observe the selections made by the platform. More formally, in a given round the agents who have not been selected (pulled) will not see which arm was pulled or the value of the reward it gave. Any arm is also not aware of the value of the round (temporal index) even when it is pulled. The arm is only aware of its own history (its own pulls, rewards, and effort). Although we consider this blind observation model, it remains non-trivial as the strategy of any arm is dynamic and dependent on its own history.

Figure 1 shows an example of the interaction between the principle (algorithm) and agents in our model. Note how each agent has full knowledge of his own performance parameters (μ_i and M_i) unlike the principle. However, unlike the principle because of the blind observation model the agents lack information about the arm selections and the temporal index.

Public and Private Information Settings: As stated earlier the principal is unaware of any of the agents' performance characteristics (none of the μ_i and M_i values). Further, while each agent $i \in \mathcal{A}$ is aware of his own performance characteristics (μ_i and M_i), he is not aware of the other agents characteristics. This pauses a significant additional difficulty in mechanism design for performance incentivization similar to that encountered in auctions.

We therefore consider two settings. The first, is the ***public-information*** setting where the top performance level M_{top} is a *public parameter known* by the principal and agents. Further, the market is large enough so that $k_{\text{top}} \geq 2$. Similar to auction theory, intuitively this setting alleviates the difficulty of the "bidding" issue among the top performing agents $\mathcal{A}_{\text{top}}$. However, designing an algorithm that is per-

[3] We will have settings where we elicit the agents to report their maximum value. However, in our setting reporting the value is done by pulling the arm and observing its reward which is different than for example auctions where a numeric value can be simply written by the agent.

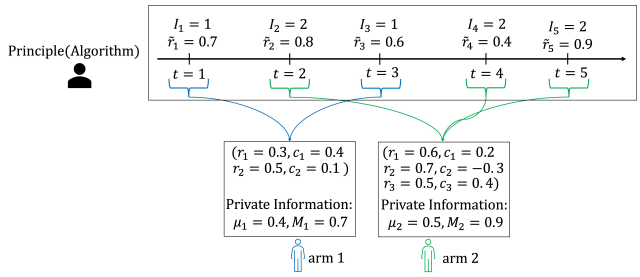

Figure 1: Model of the interaction. The information available to each agent and the principal (algorithm) is enclosed within their respective box. Since we are in a blind observation model, the agents do not have access to the same time index and use an "internal" time index that is different from that of the principle. Hence, agents 1 and 2 record their realized rewards and efforts by the indices $1, 2$ and $1, 2, 3$, respectively.

formance incentivizing while also being robust remains non-trivial. Moreover, this setting is strongly motivated by practical online labor markets such as Uber where it is known that a 5 star rating ($M_{\text{top}} = 5$) is achievable by many workers ($k_{\text{top}} \geq 2$).

The second setting we consider is the ***private-information*** setting where the top performance level is not known neither to the agents nor the principal, thereby not requiring the strong assumptions of the public-information setting. To better represent our algorithms we start with the public-information setting in Section 5 then the private-information setting in Section 6.

3.2 Mechanism Design Objective and Equilibria

The principal receives the rewards and therefore his primary measurement is the accumulated reward (revenue) through the horizon $P(n)$ defined as:

$$P(n) = \mathbb{E}[\sum_{t=1}^{n} \tilde{r}_{I_t}(t)] \tag{2}$$

Now we give the details for the principal's mechanism design objective.

Performance Incentivization: In light of the above model, the first natural objective for the principal is to accumulate revenue at the level of M_{top}, i.e., have $P(n) \geq M_{\text{top}}\, n - o(n)$. To do that the principal has to use an algorithm (mechanism) that is *performance incentivizing*, i.e., obtaining top performance revenue at equilibrium. More specifically, let the strategy profile over the arms be $S = (S_1, \ldots, S_k)$ then we should have:

$P(n, S) \geq M_{\text{top}}\, n - o(n)$, where S is an equilibrium profile.

We write $P(n, S)$ when we want to emphasize the revenue's dependence on the strategy. Now we formalize our notions of equilibrium. Note that our notions have similarities to that of (Braverman et al. 2019):

Definition 3.1. Asymptotic Equilibrium: A strategy profile S is an asymptotic equilibrium if given any arm $i \in [k]$ we have:

$$\lim_{n\to\infty} \frac{u_i(S'_i, S_{-i})}{u_i(S_i, S_{-i})} \leq 1 \quad \text{for any strategy } S'_i. \tag{3}$$

The above essentially states, that at equilibrium an agent is not significantly better off as the horizon becomes larger by deviating to another strategy.

Because of the heterogeneity in the agents' maximum performance levels, it is meaningful to introduce a more generalized notion of equilibrium. First, we establish some notation. Consider a partition of the set of arms $\mathcal{A}$ into $\mathcal{A}_{\text{prev}}$ (the "prevalent" set) and $\mathcal{A}_{\text{non-prev}}$ (the "non-prevalent" set) where the strategy profile is denoted by S_{prev} and $S_{\text{non-prev}}$ for the sets $\mathcal{A}_{\text{prev}}$ and $\mathcal{A}_{\text{non-prev}}$, respectively. Further, given an arm $i \in \mathcal{A}_{\text{prev}}$, we denote the strategy profile the other arms in the prevalent set ($\mathcal{A}_{\text{prev}} - \{i\}$) follow by $S_{(\text{prev},\text{-}i)}$ whereas i's strategy is $S_{(\text{prev},i)}$ under profile S_{prev}. We now define partial asymptotic equilibrium:

Definition 3.2. Partial Asymptotic Equilibrium: A partial asymptotic equilibrium holds if the set of arms $\mathcal{A}$ can be partitioned into $\mathcal{A}_{\text{prev}} \cup \mathcal{A}_{\text{non-prev}}$ such that $|\,\mathcal{A}_{\text{prev}}\,| \geq 1$ with the arms in $\mathcal{A}_{\text{prev}}$ following profile S_{prev} such that the following is true $\forall i \in \mathcal{A}_{\text{prev}}$:

$$\lim_{n\to\infty} \frac{u_i(S'_i, S_{(\text{prev},\text{-}i)}, S'_{\text{non-prev}})}{u_i(S_{(\text{prev},i)}, S_{(\text{prev},\text{-}i)}, S_{\text{non-prev}})} \leq 1 \tag{4}$$

for any strategy S'_i of i and any strategy profile $S'_{\text{non-prev}}$ over $\mathcal{A}_{\text{non-prev}}$.

The notion of partial asymptotic equilibrium essentially states that the arms in the prevalent set are at equilibrium regardless of the strategies followed by the non-prevalent set. This notion captures a natural setting where the arms which cannot give top performance ($M_i < M_{\text{top}}$) have little influence on the other arms and revenue.

Note that for an arm i the strategy S^*_i denotes the maximum performance strategy of always giving M_i. Further, S^* denotes a strategy profile where either all of the arms $\mathcal{A}$ or at least the subset of top arms $\mathcal{A}_{\text{top}}$ follow the strategy of giving the maximum reward value S^*_i.

3.3 Robustness Objective

While the above performance incentivization objective is clearly important and desirable, it is not sufficient. Specifically, it ignores non-equilibrium behavior which is known to be exhibited by agents in the real world. For example, agents may behave irrationally or they maybe non-strategic which is observed to happen in many settings (Camerer 1997; Güth, Schmittberger, and Schwarze 1982; Singh 2012). Therefore, one would want the algorithm to be ***robust***, i.e., achieving as a fallback a non-trivial amount of revenue if non-equilibrium strategies are followed. Based on our model, a natural choice would be that the principal always obtains rewards at least at the level of highest mean of honest agents. Formally, we should have:

$$P(n, S) \geq \mu^*_{\mathcal{H}}\, n - o(n) \;\text{ for any strategy profile } S. \tag{5}$$

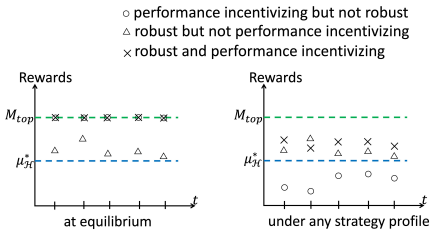

Figure 2: An illustrative figure showing the rewards we expect to obtain using different algorithms that satisfy different objectives. Notice how the algorithm that is both robust and performance incentivizing never falls below $\mu^*_{\mathcal{H}}$ and obtains rewards of M_{top} at equilibrium.

Figure 2 shows an illustration of the two desired objectives of performance incentivization and robustness.

3.4 Main Results

Our main results are the following:

- **Robustness:** In Section 4 we identify a property called *sharp adaptivity* and show that any sharply adaptive MAB algorithm obtains revenue (total rewards) of at least $P(n) \geq \mu^*_{\mathcal{H}}\, n - o(n)$ for any strategy profile. We further show that UCB and ϵ-greedy are examples of sharply adaptive MAB algorithms.
- **Public-Information Setting:** In Section 5 we identify two additional properties of MAB algorithms *monotinicity* and *fairness among the top arms (FATA)*, we abbreviate algorithms that satisfy these two properties in addition to sharp adaptivity by *SAMF*. We prove that UCB and ϵ-greedy are also SAMF algorithms. For the public-information setting we show that: (1) Under a condition which roughly corresponds to the number of top arms k_{top} being large, no equilibrium that leads to revenue less than $M_{\text{top}}\, n - o(n)$ exists when using SAMF algorithms. (2) SAMF algorithms lead to an equilibrium where all arms in $\mathcal{A}_{\text{top}}$ give the top performance regardless of whether the previous condition holds.
- **Private-Information Setting:** In Section 6 for the private-information setting we introduce an algorithm (SP+SAMF) that combines ideas from second price auctions with SAMF algorithms. This algorithm leads to an equilibrium with revenue of $M_{\text{top-1}} n - o(n)$ where $M_{\text{top-1}}$ is the second highest value of M_i.
- **Failure of a Pure Mechanism Design Approach:** In Section 7 we underscore the importance of the robustness objective. Specifically, we construct a benign example where a non-robust mechanism obtains a vacuous amount of revenue despite being performance incentivizing.

Further in Appendix **??** we show how some of our results generalize in the public-information setting to a range of cost functions. Due to space limits we delay the proofs to the appendix.

4 Obtaining a Robust Bandit Mechanism

Here we show how to satisfy robustness. Specifically, we define a property which we call *sharp adaptivity* and show that any sharply adaptive bandit algorithm is robust. I.e., it achieves revenue of $P(n) \geq \mu^*_{\mathcal{H}}\, n - o(n)$ for any strategy profile. At an intuitive and rough level when using a sharply adaptive algorithm if an arm i is pulled for a linear proportion of the horizon αn where α is a constant, then the total rewards accumulated must be at least at the level of $\mu^*_{\mathcal{H}}$, if i starts giving lower reward values then the algorithm quickly responds and no longer pulls i. Now we give the formal definition of sharp adaptivity.

Definition 4.1. Sharp Adaptivity: A MAB algorithm is sharply adaptive if for any given arm $i \neq h^*$, if $\mathbb{E}[T_i(n)] \geq \alpha n$ where $\alpha > 0$ is a constant, then the total expected effort spent by arm i is at least $\mathbb{E}[C_i(n)] \geq \alpha(\mu^*_{\mathcal{H}} - \mu_i)n - o(n)$, for any possible strategy profile followed by the arms.

We now prove that all sharply adaptive MAB algorithms satisfy robustness. This essentially follows since sharp adaptivity implies that if any arm $i \neq h^*$ is pulled for a linear proportion of the horizon (αn where α is some constant) then it must have given rewards of at least $\mu^*_{\mathcal{H}}$ for that proportion ($\mu^*_{\mathcal{H}}\, \alpha n - o(n)$), hence the total accumulated rewards through the horizon are at least $\mu^*_{\mathcal{H}}\, n - o(n)$.

Theorem 4.2. *For any strategy profile and arbitrary cost functions, a sharply adaptive MAB algorithm obtains revenue of* $P(n) \geq \mu^*_{\mathcal{H}}\, n - o(n)$.

Interestingly, we prove that UCB and ϵ-greedy are sharply adaptive. We note that while (Feng, Parkes, and Xu 2020) proved a related robustness result, their analysis is too loose. More concretely, in our setting their result implies that if $C_i(n) \leq \alpha(\mu^*_{\mathcal{H}} - \mu_i)n$, then $\mathbb{E}[T_i(n)] \leq \beta\alpha n$ where $\beta \geq 3$. A bound with $\beta > 1$ is not sufficient as it is too loose to guarantee robustness. We therefore have to establish the robustness of UCB and ϵ-greedy using different techniques.

Theorem 4.3. *UCB and ϵ-greedy are sharply-adaptive.*

Interestingly, one can also show that algorithms that are not sharply-adaptive can fail to be robust. Concrete examples include explore-then-commit and "successive elimination" (which is regret-optimal in the standard bandit setting) (Slivkins et al. 2019). This is because such algorithms only have an initial *limited* phase (time period) where different arms are explored, and then afterwards an arm with the "best performance" is permanently selected and pulled for the rest of the horizon with the other arms being "eliminated". Therefore, if an arm with mean below $\mu^*_{\mathcal{H}}$ gives best performance in that initial phase and then gives lower rewards (essentially below $\mu^*_{\mathcal{H}}$) after being permanently selected, then it would clearly break robustness.

5 Public-Information Setting

Having resolved the robustness objective by using a sharply-adaptive algorithm we now turn to the mechanism design issue in the public-information setting. We start this section by showing an interesting impossibility result. Specifically, for

any sublinear regret MAB algorithm[4] that the principle can use no arm i can have the maximum performance strategy S_i^* of always giving M_i as a *dominant strategy*. The intuition behind this result is that while an agent may receive most of the pulls in the horizon by giving the top performance level of M_{top}, because of the performance cost he would obtain higher utility by lowering his performance to only outperform the second highest agent.

Theorem 5.1. *Under any sublinear regret MAB algorithm, it is not a dominant strategy for an arm i to follow the maximum performance strategy S_i^*.*

Before we provide our main positive results, we introduce two properties of bandit algorithms that lead to good equilibria. We start with monotonicity which essentially states that an arm i is rewarded with more pulls if it upgrades its performance to the top performance level S_i^*.

Definition 5.2. Monotonicity: A MAB algorithm satisfies monotonicity if for any strategy profile $S = (S_1, \ldots, S_k)$, if an arm i was to deviate to the top performance strategy S_i^*, then for any reward seed and any strategy seed[5] $\forall j \neq i$: $T_j(n, S_i^*, S_{-i}) \leq T_j(n, S_i, S_{-i})$.

Furthermore, we introduce the Fairness Among the Top Arms (FATA) property which states that arms that follow the top performance strategy roughly receive an equal number of pulls.

Definition 5.3. Fairness Among the Top Arms (FATA): When there exists a subset of arms $\mathcal{A}_{r^*}$ such that whenever $i \in \mathcal{A}_{r^*}$ is pulled then a reward of $r^* = \max_{t\in[n]} r_t$ is given, then a MAB algorithm is fair among the top arms (FATA) if $\forall i, i' \in \mathcal{A}_{r^*} : |\,\mathbb{E}[T_i(n)] - \mathbb{E}[T_{i'}(n)]| \leq o(n)$.

Bandit algorithms that satisfy sharp adaptivity, monotonicity, and FATA will play a fundamental and we abbreviate them by *SAMF*.

Definition 5.4. SAMF algorithm: A MAB algorithm is SAMF if it satisfies sharp adaptivity, monotonicity, and FATA.

We now show that standard algorithms such as UCB and ϵ-greedy are SAMF.

Theorem 5.5. *UCB and ϵ-greedy are SAMF algorithms.*

In the next two subsections, we consider the public-information setting and prove positive equilibrium results with SAMF algorithms.

5.1 Non-Existence of Sub-Optimal Revenue Equilibria for Large k_{top}

In the public-information setting, we show that there exists no asymptotic equilibrium for which the principal receives revenue (total rewards) less than M_{top}. Specifically, we show that any MAB algorithm that is SAMF[6] has no asymptotic equilibrium where a principal obtains a revenue less than $P(n) = M_{\text{top}}\, n - o(n)$. Note that we actually show that no partial asymptotic equilibrium exists and therefore this implies that no asymptotic equilibrium exists where the principal obtains less than top performance revenue. All that is required is an assumption on the number of top arms as given below:

Condition 5.6. $k_{top} > \frac{2}{\min_{i\in\mathcal{A}_{top}}(1+\mu_i - M_{top})}(1+\epsilon)$ *where* $\epsilon > 0$ *is a constant.*

We now present the theorem which excludes bad equilibria.

Theorem 5.7. *Given a MAB algorithm that satisfies sharp adaptivity and monotonicity, if Condition 5.6 is satisfied, then there exists no partial asymptotic equilibrium strategy profile S such that $P(n, S) \leq \alpha\, M_{top}\, n + o(n)$ where $\alpha < 1$.*

Proof Sketch and Interpretation of Condition 5.6. At a rough level, the proof shows that there must be a linear proportion of the horizon of size at least $(1-\alpha)n - o(n)$ where sub-optimal revenue (less than M_{top}) is accumulated. Further, there must be a top arm that is pulled for at most $\frac{1}{k_{\text{top}}}$ of that proportion (at most $\frac{(1-\alpha)n - o(n)}{k_{\text{top}}}$). If the number of top arms k_{top} is large then the arm's utility is low. If the cost of performing at the top value which is $(M_{\text{top}} - \mu_i)$ in expectation is relatively small compared to the current utility as would be implied by Condition 5.6 then we would have $1 - (M_{\text{top}} - \mu_i) > \frac{(1+\mu_i)}{k_{\text{top}}}$, then this arm would have the incentive to deviate to a top performance level to take the $(1-\alpha)n - o(n)$ portion of the horizon.

This theorem also implies that the principal can make sure that he always obtains top performance revenue by increasing the number of top arms (higher k_{top}) and by making the gap needed for optimal performance $(M_{\text{top}} - \mu_i)$ small.

Although the above explanation is intuitive, since the MAB setting is dynamic the proof of the theorem is involved and requires some new techniques.

5.2 Achieving Top Performance Equilibrium for Any Value of k_{top}

Here we show that a SAMF algorithm leads to top performance revenue equilibrium even if Condition 5.6 does not hold. Specifically, we show that S^* is an equilibrium profile. At an intuitive level if a top arm $i \in \mathcal{A}_{\text{top}}$ follows strategy S_i^* of always giving M_{top} then by sharp adaptivity any other top arm $j \in \mathcal{A}_{\text{top}} - \{i\}$ needs to perform at that level (always give rewards of M_{top}) to receive a linear number of pulls. By additionally invoking the the monotonicity and FATA properties we can establish upper and lower bounds on j's utility through bounds on the number of pulls and cost. With the utility of a top arm j lower bounded under S^* and upper bounded under any deviation strategy the equilibrium is established.

[4] As in the standard terminology, a sublinear regret MAB algorithm is one where in the ordinary stochastic setting (arms cannot modify their realized rewards) the regret is sublinear, i.e., $R(n) = (\max_{i\in\mathcal{A}} \mu_i)n - \mathbb{E}[\sum_{t=1} r_{I_t}(t)] = o(n)$.

[5] Note that a strategy is in general randomized.

[6] More specifically, this result holds if the MAB algorithm is sharply adaptive and monotonic but the algorithm does not necessarily need to satisfy the FATA property.

Theorem 5.8. *Given the public-information setting, if a MAB algorithm is SAMF, then the strategy profile S^* where $\forall i \in \mathcal{A}_{top} : S_i = S_i^*$ is a partial asymptotic equilibrium leading to optimal revenue $P(n, S^*) = M_{top}\, n - o(n)$.*

6 Private-Information Setting: SAMF Algorithms with a Second Price Auction

Here we generalize our algorithm (mechanism) to deal with the case where the top performance level M_{top} is not publicly known neither to the principle nor to any of the arms[7]. Since our mechanism combines methods from second price (SP) auctions with SAMF algorithms we call it SP+SAMF (Algorithm 1). The rounds under this mechanism can be divided into 3 phases: (A) bidding phase rounds (line 1), (B) SAMF phase (line 4), and (C) reward phase (line 5). The rounds in the first and last phases (A) and (C) are only $o(n)$ while the middle phase (B) contains $\Omega(n)$ rounds. We assume that the SAMF algorithm we use has in the ordinary stochastic setting an instance-dependent logarithmic upper bound for the number of pulls of sub-optimal arms. I.e., in the ordinary stochastic (unmodifiable) reward setting, a sub-optimal arm i has $\mathbb{E}[T_i(n)] = O(\frac{\ln(n)}{\Delta_i^2})$. UCB is an example of such an algorithm. The mechanism tunes some parameters according to a logarithmic upper bound on the number of pulls, however other bounds can be accommodated as long as the MAB algorithm is a SAMF algorithm.

We give a brief intuitive explanation for the key ideas of the mechanism. The bidding process of phase (A) where each arm i reports its top performance value by giving a reward m_i resolves the difficulty of not knowing the top performance level. This is the case since the mechanism tells all of the arms the value of m' which is the second highest value among the m_i values[8]. In phase (B), if $k_{\text{top}} \geq 2$ then the setting essentially reduces to the public-information setting where the strategy profile S^* of always giving the top performance is an equilibrium. On the other hand, if $k_{\text{top}} = 1$ then we show that it is an equilibrium for the unique top arm to give a performance value of $m' + \frac{1}{\ln(n)}$ which is slightly above the second highest value and for the other arms to give their top performance value of M_i. For the unique top arm, this equilibrium is established since outperforming the rest of the arms enables it to essentially receive almost of the pulls ($n - o(n)$ pulls) whereas for the other arms giving their top performance results in higher pulls in the last reward phase (C) since each arm is pulled a number of times proportional to its performance (average reward it gives) in phase (B). Critically, our mechanism include the blocking condition of line (4A) in phase (B). The main objective behind this blocking condition is to make sure that the top arms do not gain a higher utility by bidding untruthfully with a higher value (see Appendix **??** for a concrete example). Note that this blocking condition is "punishing" for untruthful bidding instead of under performance as done in (Braverman et al. 2019). In the next section we consider a mechanism based on (Braverman et al. 2019) that punishes for under-performance instead. We demonstrate its failure due to its lack of robustness even under truthful bidding. Since we assume that honest agents always bid truthfully and we use a SAMF algorithm in phase (B) we always achieve robustness for any strategy profile.

[7]Note that as usual each arm i still knows its properties including its own maximum performance level M_i. But in this setting it would not know if $M_{\text{top}} = M_i$.

[8]Note that if more than one arm reports the same maximum value then m' would actually equal the maximum value.

Algorithm 1: SP+SAMF

(1) Pull each arm i once, let m_i be its reported value.
(2) Let m' be the second highest reported value.
(3) Tell all arms the value m'.
(4) Ignore all previous rewards and use a SAMF MAB algorithm with two modifications:
4A-if an arm i has $m_i \leq m'$, but gives reward $r > m'$, then block the arm (never play it again).
4B-stop at round $t = n - k\,\lceil \ln(n) \rceil^{\rho+3}$ for some fixed constant $\rho > 0$.
(5) Tell the arms that this is the reward phase then pull each arm i for a number of rounds N_i where $\mathbb{E}[N_i] = \hat{\mu}_i^{\text{SAMF}}\big(\ln(n)\big)^{\rho+3}$ and $\hat{\mu}_i^{\text{SAMF}}$ is the mean of arm i during SAMF phase rounds.

Below we give a full description of the equilibrium strategy profile. For convenience, we assume below that if $k_{\text{top}} = 1$, then the only agent with $M_i = M_{\text{top}}$ is not h^* (the honest agent with the maximum mean), this case leads to identical guarantees but using a more complicated strategy for the honest agent h^* since we assume that honest agents would never absorb (degrade) their rewards (see Appendix **??**).

Theorem 6.1. *Under SP+SAMF (Algorithm 1), the strategy profile $S^{\text{SP+SAMF}}$ is an asymptotic equilibrium that consists of the following: (1) In the bidding phase (line 1) each arm bids truthfully with $m_i = M_i$, (2) In the reward phase (line 5), all rewards are absorbed if the agent is not honest otherwise no positive effort is added. (3) In the SAMF phase (line 4) the strategy is:*

$$S_i^{\text{SP+SAMF}} = \begin{cases} \textit{if } M_i \leq m', \textit{ then always give a reward of } M_i \\ \textit{if } M_i > m', \textit{ then give } m' + \frac{1}{\ln(n)} \end{cases}$$

Denoting the second highest top performance value by $M_{\text{top-1}}$, the following theorem is immediately implied by Theorem 6.1:

Theorem 6.2. *The SP+SAMF mechanism (Algorithm (1)) leads to: (1) an equilibrium where the principal obtains revenue of $P(n) \geq M_{\text{top-1}} \cdot n - o(n)$. (2) Revenue of $P(n) \geq \mu^*_{\mathcal{H}}\, n - o(n)$ for any strategy profile.*

7 Failure of a Non-Robust Mechanism Design Approach

Here we show the failure of a non-robust mechanism design approach even in a setting where all agents are honest, we consider a mechanism similar to the one introduced in (Braverman et al. 2019). Specifically, consider PURE-SP

(Algorithm 2) which uses elements from second price auction similar to SP+SAMF but unlike SP+SAMF does not use a SAMF algorithm. More concretely, the mechanism starts with a bidding phase and then tells all arms the second highest reported value m'. It then proceeds to pull all of the arms which gave the maximum bid value ($\max_{i \in \mathcal{A}} m_i$) in a cycling manner. If a top bidding arm under-performs, it is deleted and not pulled again as shown in line (5A). The mechanism ends with a rewarding phase where each arm i is pulled a number of times proportional to its bid value m_i.

Algorithm 2: PURE-SP

(1) Pull each arm i once, let m_i be its reported value..
(2) Let m' be the second highest reported value.
(3) Tell all arms the value m'.
(4) Let $\mathcal{A}_{m\prime}$ be the subset of arms that gave the highest m_i value. I.e., $\mathcal{A}_{m\prime} = \{i \in \mathcal{A} \,|m_i = \arg\max_{j \in \mathcal{A}} m_j\}$
(5) Pull arms $i \in \mathcal{A}_{m\prime}$ in a cycling manner:
5A-delete arm i if it gives a value less than m' and update $\mathcal{A}_{m\prime}$.
5B-stop at round $t = n - 2k$.
(6) Tell the arms that this is the reward phase then pull each arm i for N_i rounds where $\mathbb{E}[N_i] = 2m_i$.

It is not difficult to show that this mechanism leads to a revenue of $P(n) \geq M_{\text{top-1}} n - o(n)$ under an equilibrium strategy profile (see Appendix **??**). However, if the agents deviate from the equilibrium strategy the revenue can easily vanish. Specifically, in the theorem below we show that even in an instance where all agents are honest the revenue can be sublinear with high probability if the agents deviate from the equilibrium in a given pull with a small constant probability of ϵ.

Theorem 7.1. *For* PURE-SP *(Algorithm 2) there exists an instance where even if the bidding is done truthfully and all of the agents are honest, if an agent at a subsequent pull deviates from the equilibrium strategy with probability $\epsilon > 0$ then with probability at least $1 - \frac{1}{n}$ the revenue $P(n) = o(n)$.*

References

Bahar, G.; Ben-Porat, O.; Leyton-Brown, K.; and Tennenholtz, M. 2020. Fiduciary bandits. In *International Conference on Machine Learning*, 518–527. PMLR.

Braverman, M.; Mao, J.; Schneider, J.; and Weinberg, S. M. 2019. Multi-armed Bandit Problems with Strategic Arms. In Beygelzimer, A.; and Hsu, D., eds., *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, 383–416. PMLR.

Camerer, C. F. 1997. Progress in behavioral game theory. *Journal of economic perspectives*, 11(4): 167–188.

Carroll, G. 2015. Robustness and linear contracts. *American Economic Review*, 105(2): 536–63.

Castiglioni, M.; Marchesi, A.; and Gatti, N. 2022. Designing Menus of Contracts Efficiently: The Power of Randomization. *arXiv preprint arXiv:2202.10966*.

Chandrasekher, M. 2015. Unraveling in a repeated moral hazard model with multiple agents. *Theoretical Economics*, 10(1): 11–49.

Dütting, P.; Roughgarden, T.; and Talgam-Cohen, I. 2019. Simple versus optimal contracts. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, 369–387.

Esmaeili, S.; Hajiaghayi, M.; and Shin, S. 2023. Replication-proof Bandit Mechanism Design. *arXiv preprint arXiv:2312.16896*.

Feng, Z.; Parkes, D.; and Xu, H. 2020. The intrinsic robustness of stochastic bandits to strategic manipulation. In *International Conference on Machine Learning*, 3092–3101. PMLR.

Ghosh, A.; and Hummel, P. 2011. A game-theoretic analysis of rank-order mechanisms for user-generated content. In *Proceedings of the 12th ACM conference on Electronic commerce*, 189–198.

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

Ghosh, A.; and McAfee, P. 2011. Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World wide web*, 137–146.

Gintis, H.; Bowles, S.; Boyd, R.; and Fehr, E. 2003. Explaining altruistic behavior in humans. *Evolution and human Behavior*, 24(3): 153–172.

Güth, W.; Schmittberger, R.; and Schwarze, B. 1982. An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization*, 3(4): 367–388.

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

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, 64–67.

Jain, S.; Chen, Y.; and Parkes, D. C. 2009. Designing incentives for online question and answer forums. In *Proceedings of the 10th ACM conference on Electronic commerce*, 129–138.

Kremer, I.; Mansour, Y.; and Perry, M. 2014. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5): 988–1012.

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

Osborne, M. J.; et al. 2004. *An introduction to game theory*, volume 3. Oxford university press New York.

Rogerson, W. P. 1985. Repeated moral hazard. *Econometrica: Journal of the Econometric Society*, 69–76.

Roughgarden, T. 2010. Algorithmic game theory. *Communications of the ACM*, 53(7): 78–86.

Shin, S.; Lee, S.; and Ok, J. 2022. Multi-armed Bandit Algorithm against Strategic Replication. In *International Conference on Artificial Intelligence and Statistics*, 403–431. PMLR.

Singh, S. 2012. Investor irrationality and self-defeating behavior: Insights from behavioral finance. *Journal of Global Business Management*, 8(1): 116.

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

Wais, P.; Lingamneni, S.; Cook, D.; Fennell, J.; Goldenberg, B.; Lubarov, D.; Marin, D.; and Simons, H. 2010. Towards building a high-quality workforce with mechanical turk. *Proceedings of computational social science and the wisdom of crowds (NIPS)*, 1–5.

Wang, S.; and Huang, L. 2018. Multi-armed bandits with compensation. *Advances in Neural Information Processing Systems*, 31.