Competing Bandits: The Perils of Exploration under Competition

GUY ARIDOR, Columbia University
YISHAY MANSOUR, Tel Aviv University and Google
ALEKSANDRS SLIVKINS, Microsoft Research New York City
ZHIWEI STEVEN WU, University of Minnesota - Twin Cities

Most modern firms strive to learn from interactions with consumers, and many engage in *exploration*: making potentially suboptimal choices for the sake of acquiring new information. We initiate a study of the interplay between *exploration* and *competition*—how such firms balance the exploration for learning and the competition for consumers. Here consumers play three distinct roles: they are customers that generate revenue, they are sources of data for learning, and they are self-interested agents which choose among the competing firms.

In our model, we consider competition between two firms facing the same multi-armed bandit instance who simultaneously choose a bandit algorithm. Users arrive one by one and choose between the two firms, so that each firm makes progress on its bandit instance if and only if it is chosen. We study to what extent competition leads to welfare increases for consumers and what algorithms firms adopt under competition. We find that stark competition induces firms to commit to a "greedy" algorithm that leads to low consumer welfare. However, we find that in order to incentivize "better" exploration strategies a firm needs to have some "free" consumers that visit them without having to be incentivized to do so. We investigate this effect via noise in consumer decision-making and asymmetries in timing of entry of firms into the market. Our findings shed light on the first-mover advantage in the digital economy by exploring the role that data can play as a barrier to entry in online markets and are closely related to the "competition vs. innovation" relationship, a well-studied theme in economics.

1 INTRODUCTION

Learning from interactions with consumers is ubiquitous in modern customer-facing systems, from product recommendations to web search to content selection to fine-tuning user interfaces. These interactions are essential for the development of high quality products, but, in order to learn what products are high quality, firms need to implement systems that engage in *exploration*: deploying potentially suboptimal products in order to acquire new information to make more informed product decisions tomorrow.

In this paper, we initiate a study of the interplay between *exploration* and *competition*. Competition induces firms to want to produce high quality goods, but they need to learn how to do so via exploration. However, while exploration may provide firms with the necessary information to allow them to produce high quality products in the long-run, it may lead to bad experiences for consumers today and adversely affect a firm's reputation. This may incentivize subsequent consumers to go to the firm's competitors tomorrow and leave the firm with a lower market share and less consumers to learn from. Consumers therefore serve three distinct roles: they are customers that generate revenue, they are sources of data for learning, and they are self-interested agents who choose among the competing firms.

Given the prevalence of these technologies and the increased importance of data in the digital economy, our main high-level question is: **does competition lead to the adoption of better exploration algorithms?** This translates into a number of more concrete questions. While it is commonly assumed that "better" technology always helps, is this so for our setting? Does increasing

Authors' addresses: Guy Aridor, Columbia University, New York, NY, USA. g.aridor@columbia.edu; Yishay Mansour, Tel Aviv University, Tel Aviv, Israel. mansour.yishay@gmail.com; Aleksandrs Slivkins, Microsoft Research, New York, NY, USA. slivkins@microsoft.com; Zhiwei Steven Wu, University of Minnesota - Twin Cities, Minneapolis, MN, USA. zsw@umn.edu.

competition lead to higher consumer welfare? To what extent is there a "data feedback loop" where one firm having more data leads to that firm attracting more users which leads to that firm having more data, etc.?

Our model. We investigate these questions with a stylized duopoly model where two firms commit to exploration strategies and compete for a stream of consumers. We define a game in which two firms (*principals*) simultaneously engage in exploration and compete for users (*agents*). These two processes are interlinked, as exploration decisions are experienced by users and informed by their feedback. We need to specify several conceptual pieces: how the principals and agents interact, what is the machine learning problem faced by each principal, and what is the information structure. Each piece can get rather complicated in isolation, let alone jointly, so we strive for simplicity. Thus, the basic model is as follows:

- A new agent arrives in each round $t = 1, 2, \ldots$, and chooses among the two principals. The principal chooses an action (*e.g.,* a list of web search results to show to the agent), the user experiences this action, and reports a reward. All agents have the same "decision rule" for choosing among the principals given the available information.
- Each principal faces a very basic and well-studied version of the multi-armed bandit problem: for each arriving agent, it chooses from a fixed set of actions (a.k.a. *arms*) and receives a reward drawn independently from a fixed distribution specific to this action.
- Principals simultaneously announce their learning algorithms before round 1, and cannot change them afterwards. There is a common Bayesian prior on the rewards (but the realized reward distributions are not observed by the principals or the agents). Each principal only observes agents that chose him. We consider two variants of the baseline model and the information set for the agent differs between the two:

(1) In the *expectation choice variant*, agents do not receive any other information and choose between the principals using their knowledge of t and the principals' algorithms.
(2) In the *reputation choice variant*, agents have access to a reputation score for each principal, which is a sliding window average of the rewards experienced by previous agents that have visited this principal.

Main Findings. We find that in a simultaneous entry duopoly a greedy (myopic) algorithm that does no purposeful exploration is incentivized in equilibrium. In the expectation choice variant, once a firm does any exploration it is immediately starved of consumers when its opponent plays a greedy algorithm since we show that the expected reward for every subsequent consumer is higher for the firm that plays the greedy algorithm. In the reputation choice variant, the same mechanism drives the greedy algorithm to be incentivized in equilibrium. In this variant, exploration hurts a firm's reputation and decrease its market share in the near term. This leaves the firm with less users to learn from, which further degrades the firm's performance relative to competitors who keep learning and improving from *their* customers, and so forth. Taken to the extreme, such dynamics lead to a "death spiral" effect when the vast majority of customers eventually switch to competitors. As a result, our model provides an example of the mechanisms that can lead to a "data feedback loop" that has been hypothesized in policy discussions. In both cases, consumer welfare is low since the greedy algorithm is known to be dramatically bad in many important cases of multi-armed bandits.

The primary mechanism that generates this stark result is that consumers need to be incentivized to select a firm over its competitors, leading firms that engage in exploration to be starved of consumers before they make enough progress on their learning problem. In order to incentivize "better" exploration strategies the key intuition is that the firm needs to have some "free" consumers

that visit them without the firm having to incentivize them to do so. This allows the firm to eventually overcome the initial losses in consumer perception from exploration.

We relax the decision rule of the consumers to allow for each firm to be chosen with some fixed baseline probability so that firms get a constant stream of "free" consumers; we call this choice rule `HardMax&Random`. We find that, in this setting, better algorithms help in a big way: a sufficiently better algorithm is guaranteed to win all non-random agents after an initial learning phase. While the precise notion of "sufficiently better algorithm" is rather subtle, we note that commonly known "smart" bandit algorithms typically defeat the commonly known "naive" ones, and the latter typically defeat the greedy algorithm. However, there is a substantial caveat: one can defeat any algorithm by interleaving it with the greedy algorithm. This has two undesirable corollaries: a better algorithm may sometimes lose, and a pure Nash equilibrium typically does not exist.

We use the reputation choice variant to analyze the effect of varying the timing of entry between the firms and allow one firm to have a first-mover advantage so that this firm (denoted as the incumbent) gets all the agents in the market before the other firm enters. We find that, if the first-mover period is sufficiently large, then the incumbent has a dominant strategy to play "smart" bandit algorithms and that consumer welfare is substantially higher than in the case of simultaneous entry. The intuition is simply that a sufficiently long incumbency period allows the firm to make sufficient progress on its learning problem that it can overcome the original drop in its reputation from exploration in the beginning rounds.

Additional Findings

Noise in consumer choice: We further relax the decision rule of the agents so that the probability of choosing a given principal varies smoothly as a function of the difference between principals' expected rewards; we call it `SoftMax`. For this decision rule, the "better algorithm wins" result holds under much weaker assumptions on what constitutes a better algorithm. This is the most technical result of the paper. The competition in this setting is necessarily much more relaxed: typically, both principals attract approximately half of the agents as time goes by (but a better algorithm may attract slightly more).

We use the reputation choice variant to explore numerically the extent to which better algorithms under the `HardMax&Random` decision rule are incentivized in realistic time-scales. We show that for small but still "relevant" parameter values, this effect does not show up for unless the time horizon is very long.

Reputation and Data First-Mover Advantage In the reputation choice variant, we investigate the "first-mover advantage" phenomenon in more detail. Being first in the market gives free data to learn from (a "data advantage") as well as a more definite, and possibly better reputation compared to an entrant (a "reputation advantage"). We run additional experiments so as to isolate and compare these two effects. We find that either effect alone leads to a significant advantage under competition. The data advantage is larger than reputation advantage when the incumbent commits to a more advanced bandit algorithm.

Data advantage is significant from an anti-trust perspective, as a possible barrier to entry. We find that even a small amount "data advantage" gets amplified under competition, causing a large difference in eventual market shares. This observation runs contrary to prior work [13, 34], which studied learning without competition, and found that small amounts of additional data do not provide significant improvement in eventual outcomes. We conclude that competition dynamics – that firms compete as they learn over time – are pertinent to these anti-trust considerations.

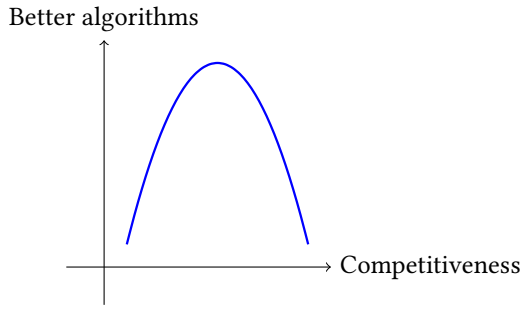


Fig. 1. Inverted-U relationship between competitiveness and algorithms.

Predicting Outcomes in Competition We also investigate how algorithms' performance "in isolation" (without competition) is predictive of the outcomes under competition in the reputation choice variant. We find that mean reputation – arguably, the most natural performance measure "in isolation" – is sometimes not a good predictor. We suggest a more refined performance measure, based on a comparison between the reputation of the two firms, and use it to explain some of the competition outcomes. **Add some discussion about comparison with "better" algorithms in analytical part**

Economic interpretation. The inverted-U relationship between the severity of competition among firms and the quality of technologies that they adopt is a familiar theme in the economics literature [*e.g.,* 3, 48].[1] We find it illuminating to frame our contributions in a similar manner, as illustrated in Figure 1.

In our model, competition varies based on the number of "free" consumers that a firm receives. One economic mechanism that can generate this comes from introducing "random" consumers where we vary from fully rational decisions with `HardMax` to relaxed rationality with `HardMax&Random` to an even more relaxed rationality with `SoftMax`. Indeed, with `HardMax` you lose all customers as soon as you fall behind in performance, with `HardMax&Random` you get some small market share no matter what, and with `SoftMax` you are further guaranteed a market share close to $\frac{1}{2}$ as long as your performance is not much worse than the competition. The uniform choice among principals corresponds to no rationality and no competition. Another economic mechanism that can generate this comes from the number of firms and the timing of entry in the market. Competition in this case ranges from monopoly to "incumbent" (first-mover in duopoly) to simultaneous entry duopoly to "late entrant" (last mover in duopoly). The same distinctions in both cases also control the severity of competition between the principals.

In both variants we identify an inverted-U relationship between competition and innovation in the spirit of Figure 1. These inverted-U relationships arise for a fundamentally different reason, compared to the existing literature on "competition vs. innovation." In the literature, better technology always helps in a competitive environment, other things being equal. Thus, the trade-off is between the costs of improving the technology and the benefits that the improved technology provides in the competition. Meanwhile, we find that a better exploration algorithm may sometimes

[1]The literature frames this relationship as one between "competition" and "innovation". In this context, "innovation" refers to adoption of a better technology, at a substantial R&D expense to a given firm. It is not salient whether similar ideas and/or technologies already exist outside the firm. It is worth noting that the adoption of exploration algorithms tends to require substantial R&D effort in practice, even if the algorithms themselves are well-known in the research literature; see [1] for an example of such R&D effort.

perform much worse under competition, even in the absence of R&D costs. This stems from the nature of exploration technologies in online markets which rely on learning from interactions with users. This leads to an implicit cost from exploration in the form of a reduced rate of users that a firm attracts and can learn from. However, interestingly, the economic mechanism in our model for incentivizing firms to engage in R&D has a qualitative similarity to the role that patents play in incentivizing innovation in standard R&D models. In these models, patents temporarily relax competition for the innovating firm by giving them exclusive access to their innovation for a limited period of time in order to incentivize them to invest in the better technology. In our model, temporarily relaxing competition in the form of giving firms free periods to learn incentivizes the firm to invest in the better technology.

Discussion We consider two separate variants of the model in order to provide a more thorough investigation of the tension between exploration and competition. The expectation choice model is tractable enough to allow us to obtain theoretical results with "asymptotic" flavor. However, for the sake of analytical tractability, we make the unrealistic simplification that users do not observe any signals about firms' ongoing performance and it is difficult to analyze important economic mechanisms such as asymmetries in the timing of entry.

The reputation choice model relaxes this simplification and accounts for competition in a more direct way as well as allows us to explore other relevant economic mechanisms in understanding the tension between exploration and competition. However, it now becomes considerably harder to analyze analytically. This is for several reasons: intricate feedback loop from performance to reputations to users to performance; mean reputation, most connected to our intuition, is sometimes a bad predictor in competition (see Sections 8 and 11); mathematical tools from regret-minimization would only produce "asymptotic" results, which do not seem to suffice. We therefore analyze our model using numerical simulation, which has several benefits. It allows us to analyze our model from a "non-asymptotic" perspective, looking for substantial effects within relevant time scales. Indeed, we start our investigation by determining what time scales are relevant in the context of this variant of the model. Further, it allows us to investigate important economic mechanisms that arise in environments where exploration and competition tensions are at play, such as the effect of incumbency, increasing the number of firms, and the extent to which data can serve as a barrier to entry.

Map of the paper. We survey related work (Section 2), lay out the model and preliminaries (Section 3), and proceed to analyze the different models. We start by analyzing the expectation choice model analytically. In Sections 4, 5, and 6 we analyze the expectation choice variant and characterize the equilibrium behavior under three different agent decision rules. Then, we turn to analyze the reputation choice variant of the model with details of the analysis described in Section 7. Sections 8 and 11 overview results from running different bandit algorithms in isolation (i.e. without competition). Sections 9 and 10 overview the results of the reputation choice variant. Section 12 presents results of varying the consumer choice rule in the reputation choice variant. Section 13 concludes.

2 RELATED WORK

Multi-armed bandits (*MAB*) is a particularly elegant and tractable abstraction for tradeoff between *exploration* and *exploitation*: essentially, between acquisition and usage of information. MAB problems have been studied in Economics, Operations Research and Computer Science for many decades; see [17, 24, 46] for background on regret-minimizing and Bayesian formulations, respectively. A discussion of industrial applications of MAB can be found in [1].

The literature on MAB is vast and multi-threaded. The most related thread concerns regret-minimizing MAB formulations with IID rewards [6, 33]. This thread includes "smart" MAB algorithms that combine exploration and exploitation, such as UCB1 [6] and Successive Elimination [21], and "naive" MAB algorithms that separate exploration and exploitation, including explore-first and ϵ-Greedy [*e.g.,* see 46].

The three-way tradeoff between exploration, exploitation and incentives has been studied in several other settings: incentivizing exploration in a recommendation system [12, 15, 19, 22, 32, 35, 36], dynamic auctions [*e.g.,* 5, 14, 29], pay-per-click ad auctions with unknown click probabilities [*e.g.,* 9, 10, 20], coordinating search and matching by self-interested agents [31], as well as human computation [*e.g.,* 23, 26, 45].

[16, 25, 30] studied models with self-interested agents jointly performing exploration, with no principal to coordinate them.

There is a superficial similarity (in name only) between this paper and the line of work on "dueling bandits" [*e.g.,* 50, 51]. The latter is not about competing bandit algorithms, but rather about scenarios where in each round two arms are chosen to be presented to a user, and the algorithm only observes which arm has "won the duel".

Our setting is closely related to the "dueling algorithms" framework [28] which studies competition between two principals, each running an algorithm for the same problem. However, this work considers algorithms for offline / full input scenarios, whereas we focus on online machine learning and the explore-exploit-incentives tradeoff therein. Also, this work specifically assumes binary payoffs (*i.e.,* win or lose) for the principals.

Other related work in economics. The competition vs. innovation relationship and the inverted-U shape thereof have been introduced in a classic book [44], and remained an important theme in the literature ever since [*e.g.,* 3, 48]. Production costs aside, this literature treats innovation as a priori beneficial for the firm. Our setting is very different, as innovation in exploration algorithms may potentially hurt the firm.

A line of work on *platform competition*, starting with [42], concerns competition between firms (*platforms*) that improve as they attract more users (*network effect*); see [49] for a recent survey. This literature is not concerned with *innovation*, and typically models network effects exogenously, whereas in our model network effects are endogenous: they are created by MAB algorithms, an essential part of the model. The "death spiral" effect that we find is similar to the increasing dominance notion found in the industrial organization literature (see e.g. [4, 18]). Further, there is also a line of work that investigates how buyer uncertainty about product quality can serve as a barrier to entry for late arrivers [11, 43]. In the reputation choice model, we observe a similar effect when we investigate the role that reputation can serve as a barrier to entry. However, in our model this effect is further strengthened by the fact that the firms have to learn while competing adding that the incumbent may not only have a reputational advantage but additionally the information it acquires serves as a further barrier to entry. The idea that data can serve as a barrier to entry, especially in online markets, has been studied theoretically in [34] and empirically in [13]. While these papers find that small amounts of additional data do not provide significant improvement, they focus on learning in isolation.

Relaxed versions of rationality similar to ours are found in several notable lines of work. For example, "random agents" (a.k.a. noise traders) can side-step the "no-trade theorem" [37], a famous impossibility result in financial economics. The `SoftMax` model is closely related to the literature on *product differentiation*, starting from [27], see [39] for a notable later paper.

There is a large literature on non-existence of equilibria due to small deviations (which is related to the corresponding result for `HardMax&Random`), starting with [40] in the context of health

insurance markets. Notable recent papers [8, 47] emphasize the distinction between HardMax and versions of SoftMax.

3 OUR MODEL AND PRELIMINARIES

Principals and agents. There are two principals and T agents. The game proceeds in rounds (we will sometimes refer to them as *global rounds*). In each round $t \in [T]$, the following interaction takes place. A new agent arrives and chooses one of the two principals. The principal chooses a recommendation: an action $a_t \in A$, where A is a fixed set of actions (same for both principals and all rounds). The agent follows this recommendation, receives a reward $r_t \in [0, 1]$, and reports it back to the principal.

The rewards are i.i.d. with a common prior. More formally, for each action $a \in A$ there is a parametric family $\psi_a(\cdot)$ of reward distributions, parameterized by the mean reward μ_a. (The paradigmatic case is 0-1 rewards with a given expectation.) The mean reward vector $\mu = (\mu_a : a \in A)$ is drawn from prior distribution $\mathcal{P}_{\texttt{mean}}$ before round 1. Whenever a given action $a \in A$ is chosen, the reward is drawn independently from distribution $\psi_a(\mu_a)$. The prior $\mathcal{P}_{\texttt{mean}}$ and the distributions $(\psi_a(\cdot) : a \in A)$ constitute the (full) Bayesian prior on rewards, denoted $\mathcal{P}$.

Each principal commits to a learning algorithm for making recommendations. This algorithm follows a protocol of *multi-armed bandits* (*MAB*). Namely, the algorithm proceeds in time-steps:[2] each time it is called, it outputs a chosen action $a \in A$ and then inputs the reward for this action. The algorithm is called only in global rounds when the corresponding principal is chosen.

The information structure is as follows. The prior $\mathcal{P}$ is known to everyone. The mean rewards μ_a are not revealed to anybody and each principal is completely unaware of the rounds when the other is chosen. We consider two variants of our model with different information structures. In the first, the *expectation choice* variant, each agent knows both principals' algorithms, and the global round when (s)he arrives, *but not* the rewards of the previous agents. In the second, the *reputation choice* variant, agents only make decisions based on the rewards of previous agents. Concretely, each of the two principals has a *reputation score*, and each agent's choice is driven by these two numbers. The reputation score is simply a sliding window average: an average reward of the last M agents that chose this firm.

Some terminology. The two principals are called "Principal 1" and "Principal 2". The algorithm of principal $i \in \{1, 2\}$ is called "algorithm i" and denoted $\texttt{alg}_i$. The agent in global round t is called "agent t"; the chosen principal is denoted i_t.

Throughout, $\mathbb{E}[\cdot]$ denotes expectation over all applicable randomness.

Bayesian-expected rewards. Consider the performance of a given algorithm $\texttt{alg}_i$, $i \in \{1, 2\}$, when it is run in isolation (*i.e.,* without competition, just as a bandit algorithm). Let $\texttt{rew}_i(n)$ denote its Bayesian-expected reward for the n-th step.

Now, going back to our game, fix global round t and let $n_i(t)$ denote the number of global rounds before t in which this principal is chosen. Then:

$$\mathbb{E}[r_t \mid \text{principal } i \text{ is chosen in round } t \text{ and } n_i(t) = n] = \texttt{rew}_i(n+1) \quad (\forall n \in \mathbb{N}).$$

Agents' response. Each agent t chooses principal i_t as follows: it chooses a distribution over the principals, and then draws independently from this distribution. Let p_t be the probability of choosing principal 1 according to this distribution. Below we specify p_t; we need to be careful so as to avoid a circular definition.

[2]These time-steps will sometimes be referred to as *local steps/rounds*, so as to distinguish them from "global rounds" defined before. We will omit the local vs. local distinction when clear from the context.

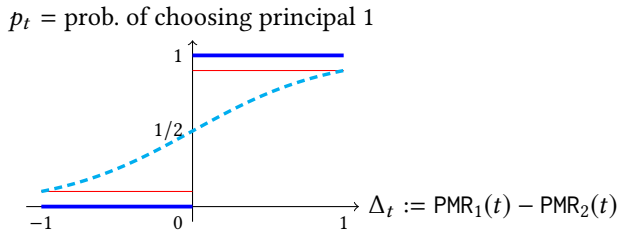


Fig. 2. The three models for agents' response function: HardMax is thick blue, HardMax&Random is slim red, and SoftMax is the dashed curve.

The form of p_t depends on the information structure that we consider. In the expectation choice model, where the agents only know the global round and the principals' algorithms, it is defined as follows.

Let $\mathcal{I}_t$ be the information available to agent t before the round. Assume $\mathcal{I}_t$ suffices to form posteriors for quantities $n_i(t)$, $i \in \{1, 2\}$, denote them by $\mathcal{N}_{i,t}$. Note that the Bayesian expected reward of each principal i is a function only of the number rounds he was chosen by the agents, so the posterior mean reward for each principal i can be written as

$$\mathsf{PMR}_i(t) := \mathbb{E}[r_t \mid \mathcal{I}_t \text{ and } i_t = i] = \mathbb{E}[\mathsf{rew}_i(n_i(t)+1) \mid \mathcal{I}_t] = \mathop{\mathbb{E}}_{n \sim \mathcal{N}_{i,t}}[\mathsf{rew}_i(n+1)].$$

This quantity represents the posterior mean reward for principal i at round t, according to information $\mathcal{I}_t$; hence the notation PMR. In general, probability p_t is defined by the posterior mean rewards $\mathsf{PMR}_i(t)$ for both principals. We assume a somewhat more specific shape:

$$p_t = f_{\mathsf{resp}}\left(\mathsf{PMR}_1(t) - \mathsf{PMR}_2(t)\right). \tag{1}$$

Here $f_{\mathsf{resp}} : [-1, 1] \to [0, 1]$ is the *response function*, which is the same for all agents. We assume that the response function is known to all agents.

To make the model well-defined, it remains to argue that information $\mathcal{I}_t$ is indeed sufficient to form posteriors on $n_1(t)$ and $n_2(t)$. This can be easily seen using induction on t.

Since all agents arrive with identical information (other than knowing which global round they arrive in), it follows that all agents have identical posteriors for $n_{i,t}$ (for a given principal i and a given global round t). This posterior is denoted $\mathcal{N}_{i,t}$.

In the reputation choice variant we consider, at a given time t each principal $i \in \{1, 2\}$ has a reputation score denoted as $\mathsf{REP}_i(t)$. In this case we have that the agent's responses take an analogous form as in the first case:

$$p_t = f_{\mathsf{resp}}\left(\mathsf{REP}_1(t) - \mathsf{REP}_2(t)\right). \tag{2}$$

Response functions. We use the response function f_{resp} to characterize the decision rule of the agents in our model. We assume that f_{resp} is monotonically non-decreasing, is larger than $\frac{1}{2}$ on the interval $(0, 1]$, and smaller than $\frac{1}{2}$ on the interval $[-1, 0)$. Beyond that, we consider three specific models (see Figure 2):

- HardMax: f_{resp} equals 0 on the interval $[-1, 0)$ and 1 on the interval $(0, 1]$. In other words, the agents will deterministically choose the principal with the higher posterior mean reward.
- HardMax&Random: f_{resp} equals ϵ_0 on the interval $[-1, 0)$ and $1 - \epsilon_0$ on the interval $(0, 1]$, where $\epsilon_0 \in (0, \frac{1}{2})$ are some positive constants. In words, each agent is a HardMax agent with probability $1 - 2\epsilon_0$, and with the remaining probability she makes a random choice.

- SoftMax: $f_{\texttt{resp}}(\cdot)$ lies in the interval $[\epsilon_0, 1-\epsilon_0]$, $\epsilon_0 > 0$, and is "smooth" around 0 (in the sense defined precisely in Section 6).

We say that $f_{\texttt{resp}}$ is *symmetric* if $f_{\texttt{resp}}(-x) + f_{\texttt{resp}}(x) = 1$ for any $x \in [0,1]$. This implies *fair tie-breaking*: $f_{\texttt{resp}}(0) = \frac{1}{2}$.

MAB algorithms. We characterize the inherent quality of an MAB algorithm in terms of its *Bayesian Instantaneous Regret* (henceforth, BIR), a standard notion from machine learning:

$$\texttt{BIR}(n) := \mathop{\mathbb{E}}_{\mu\sim\mathcal{P}_{\texttt{mean}}}\left[\max_{a\in A}\mu_a\right] - \texttt{rew}(n), \tag{3}$$

where $\texttt{rew}(n)$ is the Bayesian-expected reward of the algorithm for the n-th step, when the algorithm is run in isolation. We are primarily interested in how BIR scales with n; we treat K, the number of arms, as a constant unless specified otherwise.

We will emphasize several specific algorithms or classes thereof. In the second half of the paper we will :

- "smart" MAB algorithms that combine exploration and exploitation, such as UCB1 [6] and Thompson Sampling [41]. These algorithms achieve $\texttt{BIR}(n) \leq \tilde{O}(n^{-1/2})$ for all priors and all (or all but a very few) steps n. This bound is known to be tight for any fixed n. [3]
- "naive" MAB algorithms that separate exploration and exploitation, such as Explore-then-Exploit and ϵ-Greedy. These algorithms have dedicated rounds in which they explore by choosing an action uniformly at random. When these rounds are known in advance, the algorithm suffers constant BIR in such rounds. When the "exploration rounds" are instead randomly chosen by the algorithm, one can usually guarantee an inverse-polynomial upper bound BIR, but not as good as the one above: namely, $\texttt{BIR}(n) \leq \tilde{O}(n^{-1/3})$. This is the best possible upper bound on BIR for the two algorithms mentioned above.
- DynamicGreedy: at each step, recommends the best action according to the current posterior: an action a with the highest posterior expected reward $\mathbb{E}[\mu_a \mid \mathcal{I}]$, where $\mathcal{I}$ is the information available to the algorithm so far. DynamicGreedy has (at least) a constant BIR for some reasonable priors, *i.e.,* $\texttt{BIR}(n) > \Omega(1)$.
- StaticGreedy: always recommends the prior best action,*i.e.,* an action a with the highest prior mean reward $\mathbb{E}_{\mu\sim\mathcal{P}_{\texttt{mean}}}[\mu_a]$. This algorithm typically has constant BIR.

When we consider the analytical model, we focus on MAB algorithms such that $\texttt{BIR}(n)$ is non-increasing; we call such algorithms *monotone*. While some reasonable MAB algorithms may occasionally violate monotonicity, they can usually be easily modified so that monotonicity violations either vanish altogether, or only occur at very specific rounds (so that agents are extremely unlikely to exploit them in practice).

More background and examples can be found in Appendix A. In particular, we prove that DynamicGreedy is monotone.

Competition game between principals. Some of our results explicitly study the game between the two principals. We model it as a simultaneous-move game: before the first agent arrives, each principal commits to an MAB algorithm. Thus, choosing a pure strategy in this game corresponds to choosing an MAB algorithm (and, implicitly, announcing this algorithm to the agents).

Principal's utility is primarily defined as the market share, *i.e.,* the number of agents that chose this principal. Principals are risk-neutral, in the sense that they optimize their expected utility.

[3] This follows from the lower-bound analysis in [7].

Assumptions on the prior. We make some technical assumptions for the sake of simplicity. First, each action a has a positive probability of being the best action according to the prior:

$$\forall a \in A: \quad \Pr_{\mu \sim \mathcal{P}_{\texttt{mean}}} [\mu_a > \mu_{a'} \ \forall a' \in A] > 0. \tag{4}$$

Second, posterior mean rewards of actions are pairwise distinct almost surely. That is, the history h at any step of an MAB algorithm[4] satisfies

$$\mathbb{E}[\mu_a \mid h] \neq \mathbb{E}[\mu_{a'} \mid h] \quad \forall a, a' \in A, \tag{5}$$

except at a set of histories of probability 0. In particular, prior mean rewards of actions are pairwise distinct: $\mathbb{E}[\mu_a] \neq \mathbb{E}[\mu'_a]$ for any $a, a' \in A$.

We provide two examples for which property (5) is 'generic', in the sense that it can be enforced almost surely by a small random perturbation of the prior. Both examples focus on 0-1 rewards and priors $\mathcal{P}_{\texttt{mean}}$ that are independent across arms. The first example assumes Beta priors on the mean rewards, and is very easy.[5] The second example assumes that mean rewards have a finite support, see Appendix B for details.

Some more notation. Without loss of generality, we label actions as $A = [K]$ and sort them according to their prior mean rewards, so that $\mathbb{E}[\mu_1] > \mathbb{E}[\mu_2] > \ldots > \mathbb{E}[\mu_K]$.

Fix principal $i \in \{1, 2\}$ and (local) step n. The arm chosen by algorithm $\texttt{alg}_i$ at this step is denoted $a_{i,n}$, and the corresponding $\texttt{BIR}$ is denoted $\texttt{BIR}_i(n)$. History of $\texttt{alg}_i$ up to this step is denoted $H_{i,n}$.

Write $\texttt{PMR}(a \mid E) = \mathbb{E}[\mu_a \mid E]$ for posterior mean reward of action a given event E.

3.1 Generalizations

Our results can be extended compared to the basic model described above.

First, unless specified otherwise, our results allow a more general notion of principal's utility that can depend on both the market share and agents' rewards. Namely, principal i collects $U_i(r_t)$ units of utility in each global round t when she is chosen (and 0 otherwise), where $U_i(\cdot)$ is some fixed non-decreasing function with $U_i(0) > 0$. In a formula,

$$U_i := \sum_{t=1}^{T} \mathbf{1}_{\{i_t = i\}} \cdot U_i(r_t). \tag{6}$$

Second, our results carry over, with little or no modification of the proofs, to much more general versions of MAB, as long as it satisfies the i.i.d. property. In each round, an algorithm can see a *context* before choosing an action (as in *contextual bandits*) and/or additional feedback other than the reward after the reward is chosen (as in, *e.g., semi-bandits*), as long as the contexts are drawn from a fixed distribution, and the (reward, feedback) pair is drawn from a fixed distribution that depends only on the context and the chosen action. The Bayesian prior $\mathcal{P}$ needs to be a more complicated object, to make sure that $\texttt{PMR}$ and $\texttt{BIR}$ are well-defined. Mean rewards may also have a known structure, such as Lipschitzness, convexity, or linearity; such structure can be incorporated via $\mathcal{P}$. All these extensions have been studied extensively in the literature on MAB, and account for a substantial segment thereof; see [17] for background and details.

[4]The *history* of an MAB algorithm at a given step comprises the chosen actions and the observed rewards in all previous steps in the execution of this algorithm.

[5]Suppose the rewards are Bernouli r.v. and the mean reward μ_a for each arm a is drawn from some Beta distribution $\text{Beta}(\alpha_a, \beta_a)$. Given any history that contains h_a number of heads and t_a number of tails from arm a, the posterior mean reward is $\frac{\alpha_a + h_a}{\alpha_a + h_a + \beta_a + t_a}$. Note that h_a and t_a take integer values. Therefore, perturbing the parameters α_a and β_a independently with any continuous noise will induce a prior with property (5) with probability 1.

3.2 Chernoff Bounds

We use an elementary concentration inequality known as *Chernoff Bounds*, in a formulation from [38].

Theorem 3.1 (Chernoff Bounds). *Consider n i.i.d. random variables $X_1 \dots X_n$ with values in $[0,1]$. Let $X = \frac{1}{n}\sum_{i=1}^{n} X_i$ be their average, and let $v = \mathbb{E}[X]$. Then:*

$$\min\big(\Pr[X - v > \delta v], \quad \Pr[v - X > \delta v]\big) < e^{-vn\delta^2/3} \quad \text{for any } \delta \in (0,1).$$

4 FULL RATIONALITY (HARDMAX)

In this section, we will consider the version in which the agents are fully rational, in the sense that their response function is `HardMax`. We show that principals are not incentivized to *explore*—*i.e.,* to deviate from `DynamicGreedy`. The core technical result is that if one principal adopts `DynamicGreedy`, then the other principal loses all agents as soon as he deviates.

To make this more precise, let us say that two MAB algorithms *deviate* at (local) step n if there is an action $a \in A$ and a set of step-n histories of positive probability such that any history h in this set is feasible for both algorithms, and under this history the two algorithms choose action a with different probability.

Theorem 4.1. *Assume* `HardMax` *response function with fair tie-breaking. Assume that* $\texttt{alg}_1$ *is* `DynamicGreedy`*, and* $\texttt{alg}_2$ *deviates from* `DynamicGreedy` *starting from some (local) step $n_0 < T$. Then all agents in global rounds $t \geq n_0$ select principal* 1.

Corollary 4.2. *The competition game between principals has a unique Nash equilibirium: both principals choose* `DynamicGreedy`.

Remark 4.3. This corollary holds under a more general model which allows time-discounting: namely, the utility of each principal i in each global round t is $U_{i,t}(r_t)$ if this principal is chosen, and 0 otherwise, where $U_{i,t}(\cdot)$ is an arbitrary non-decreasing function with $U_{i,t}(0) > 0$.

4.1 Proof of Theorem 4.1

The proof starts with two auxiliary lemmas: that deviating from `DynamicGreedy` implies a strictly smaller Bayesian-expected reward, and that `HardMax` implies a "sudden-death" property: if one agent chooses principal 1 with certainty, then so do all subsequent agents do. We re-use both lemmas in later sections, so we state them in sufficient generality.

Lemma 4.4. *Assume that* $\texttt{alg}_1$ *is* `DynamicGreedy`*, and* $\texttt{alg}_2$ *deviates from* `DynamicGreedy` *starting from some (local) step $n_0 < T$. Then* $\texttt{rew}_1(n_0) > \texttt{rew}_2(n_0)$*. This holds for any response function* $f_{\texttt{resp}}$.

Lemma 4.4 does not rely on any particular shape of the response function because it only considers the performance of each algorithm without competition.

Proof of Lemma 4.4. Since the two algorithms coincide on the first $n_0 - 1$ steps, it follows by symmetry that histories H_{1,n_0} and H_{2,n_0} have the same distribution. We use a *coupling argument*: w.l.o.g., we assume the two histories coincide, $H_{1,n_0} = H_{2,n_0} = H$.

At local step n_0, `DynamicGreedy` chooses an action $a_{1,n_0} = a_{1,n_0}(H)$ which maximizes the posterior mean reward given history H: for any realized history $h \in \texttt{support}(H)$ and any action $a \in A$

$$\texttt{PMR}(a_{1,n_0} \mid H = h) \geq \texttt{PMR}(a \mid H = h). \tag{7}$$

[as: Rewrote the rest of the proof to account for positive-prob set of histories.]

By assumption (5), it follows that

$$\texttt{PMR}(a_{1,n_0} \mid H = h) > \texttt{PMR}(a \mid H = h) \quad \text{for any } h \in \texttt{support}(H) \text{ and } a \neq a_{1,n_0}(h). \tag{8}$$

Since the two algorithms deviate at step n_0, there is a set $S \subset \texttt{support}(H)$ of step-n_0 histories such that $\Pr[S] > 0$ and any history $h \in S$ satisfies $\Pr[a_{2,n_0} \neq a_{1,n_0} \mid H = h] > 0$. Combining this with (8), we deduce that

$$\texttt{PMR}(a_{1,n_0} \mid H = h) > \mathbb{E}\left[\mu_{a_{2,n_0}} \mid H = h\right] \quad \text{for each history } h \in S. \tag{9}$$

Using (7) and (9) and integrating over realized histories h, we obtain $\texttt{rew}_1(n_0) > \texttt{rew}_2(n_0)$. □

Lemma 4.5. *Consider* HardMax *response function with* $f_{\texttt{resp}}(0) \geq \frac{1}{2}$. *Suppose* $\texttt{alg}_1$ *is monotone, and* $\texttt{PMR}_1(t_0) > \texttt{PMR}_2(t_0)$ *for some global round* t_0. *Then* $\texttt{PMR}_1(t) > \texttt{PMR}_2(t)$ *for all subsequent rounds* t.

Proof. Let us use induction on round $t \geq t_0$, with the base case $t = t_0$. Let $\mathcal{N} = \mathcal{N}_{1,t_0}$ be the agents' posterior distribution for n_{1,t_0}, the number of global rounds before t_0 in which principal 1 is chosen. By induction, all agents from t_0 to $t-1$ chose principal 1, so $\texttt{PMR}_2(t_0) = \texttt{PMR}_2(t)$. Therefore,

$$\texttt{PMR}_1(t) = \mathop{\mathbb{E}}_{n\sim\mathcal{N}}[\texttt{rew}_1(n+1+t-t_0)] \geq \mathop{\mathbb{E}}_{n\sim\mathcal{N}}[\texttt{rew}_1(n+1)] = \texttt{PMR}_1(t_0) > \texttt{PMR}_2(t_0) = \texttt{PMR}_2(t),$$

where the first inequality holds because $\texttt{alg}_1$ is monotone, and the second one is the base case. □

Proof of Theorem 4.1. Since the two algorithms coincide on the first $n_0 - 1$ steps, it follows by symmetry that $\texttt{rew}_1(n) = \texttt{rew}_2(n)$ for any $n < n_0$. By Lemma 4.4, $\texttt{rew}_1(n_0) > \texttt{rew}_2(n_0)$.

Recall that $n_i(t)$ is the number of global rounds $s < t$ in which principal i is chosen, and $\mathcal{N}_{i,t}$ is the agents' posterior distribution for this quantity. By symmetry, each agent $t < n_0$ chooses a principal uniformly at random. It follows that $\mathcal{N}_{1,n_0} = \mathcal{N}_{2,n_0}$ (denote both distributions by $\mathcal{N}$ for brevity), and $\mathcal{N}(n_0 - 1) > 0$. Therefore:

$$\begin{aligned}
\texttt{PMR}_1(n_0) = \mathop{\mathbb{E}}_{n\sim\mathcal{N}}[\texttt{rew}_1(n+1)] &= \sum_{n=0}^{n_0-1} \mathcal{N}(n)\cdot\texttt{rew}_1(n+1) \\
&> \mathcal{N}(n_0-1)\cdot\texttt{rew}_2(n_0) + \sum_{n=0}^{n_0-2} \mathcal{N}(n)\cdot\texttt{rew}_2(n+1) \\
&= \mathop{\mathbb{E}}_{n\sim\mathcal{N}}[\texttt{rew}_2(n+1)] = \texttt{PMR}_2(n_0)
\end{aligned} \tag{10}$$

So, agent n_0 chooses principal 1. By Lemma 4.5 (noting that DynamicGreedy is monotone), all subsequent agents choose principal 1, too. □

4.2 HardMax with biased tie-breaking

The HardMax model is very sensitive to the tie-breaking rule. For starters, if ties are broken deterministically in favor of principal 1, then principal 1 can get all agents no matter what the other principal does, simply by using StaticGreedy.

Theorem 4.6. *Assume* HardMax *response function with* $f_{\texttt{resp}}(0) = 1$ *(ties are always broken in favor of principal* 1*). If* $\texttt{alg}_1$ *is* StaticGreedy*, then all agents choose principal* 1.

Proof. Agent 1 chooses principal 1 because of the tie-breaking rule. Since StaticGreedy is trivially monotone, all the subsequent agents choose principal 1 by an induction argument similar to the one in the proof of Lemma 4.5. □

A more challenging scenario is when the tie-breaking is biased in favor of principal 1, but not deterministically so: $f_{\texttt{resp}}(0) > \frac{1}{2}$. Then this principal also has a "winning strategy" no matter what the other principal does. Specifically, principal 1 can get all but the first few agents, under a mild technical assumption that DynamicGreedy deviates from StaticGreedy. Principal 1 can use DynamicGreedy, or any other monotone MAB algorithm that coincides with DynamicGreedy in the first few steps.

Theorem 4.7. *Assume* HardMax *response function with* $f_{\texttt{resp}}(0) > \frac{1}{2}$ *(i.e., tie-breaking is biased in favor of principal* 1*). Assume the prior* $\mathcal{P}$ *is such that* DynamicGreedy *deviates from* StaticGreedy *starting from some step* n_0*. Suppose that principal* 1 *runs a monotone MAB algorithm that coincides with* DynamicGreedy *in the first* n_0 *steps. Then all agents* $t \geq n_0$ *choose principal* 1*.*

Proof. The proof re-uses Lemmas 4.4 and 4.5, which do not rely on fair tie-breaking.

Because of the biased tie-breaking, for each global round t we have:

$$\text{if } \texttt{PMR}_1(t) \geq \texttt{PMR}_2(t) \text{ then } \Pr[i_t = 1] > \tfrac{1}{2}. \tag{11}$$

Recall that i_t is the principal chosen in global round t.

Let m_0 be the first step when $\texttt{alg}_2$ deviates from DynamicGreedy, or DynamicGreedy deviates from StaticGreedy, whichever comes sooner. Then $\texttt{alg}_2$, DynamicGreedy and StaticGreedy coincide on the first $m_0 - 1$ steps. Moreover, $m_0 \leq n_0$ (since DynamicGreedy deviates from StaticGreedy at step n_0), so $\texttt{alg}_1$ coincides with DynamicGreedy on the first m_0 steps.

So, $\texttt{rew}_1(n) = \texttt{rew}_2(n)$ for each step $n < m_0$, because $\texttt{alg}_1$ and $\texttt{alg}_2$ coincide on the first $m_0 - 1$ steps. Moreover, if $\texttt{alg}_2$ deviates from DynamicGreedy at step m_0 then $\texttt{rew}_1(m_0) > \texttt{rew}_2(m_0)$ by Lemma 4.4; else, we trivially have $\texttt{rew}_1(m_0) = \texttt{rew}_2(m_0)$. To summarize:

$$\texttt{rew}_1(n) \geq \texttt{rew}_2(n) \quad \text{for all steps } n \leq m_0. \tag{12}$$

We claim that $\Pr[i_t = 1] > \frac{1}{2}$ for all global rounds $t \leq m_0$. We prove this claim using induction on t. The base case $t = 1$ holds by (11) and the fact that in step 1, DynamicGreedy chooses the arm with the highest prior mean reward. For the induction step, we assume that $\Pr[i_t = 1] > \frac{1}{2}$ for all global rounds $t < t_0$, for some $t_0 \leq m_0$. It follows that distribution $\mathcal{N}_{1,t_0}$ stochastically dominates distribution $\mathcal{N}_{2,t_0}$.[6] Observe that

$$\texttt{PMR}_1(t_0) = \mathop{\mathbb{E}}_{n\sim\mathcal{N}_{1,t_0}} [\texttt{rew}_1(n+1)] \geq \mathop{\mathbb{E}}_{n\sim\mathcal{N}_{2,t_0}} [\texttt{rew}_2(n+1)] = \texttt{PMR}_2(t_0). \tag{13}$$

So the induction step follows by (11). Claim proved.

Now let us focus on global round m_0, and denote $\mathcal{N}_i = \mathcal{N}_{i,m_0}$. By the above claim,

$$\mathcal{N}_1 \text{ stochastically dominates } \mathcal{N}_2, \text{ and moreover } \mathcal{N}_i(m_0 - 1) > \mathcal{N}_i(m_0 - 1). \tag{14}$$

By definition of m_0, either (i) $\texttt{alg}_2$ deviates from DynamicGreedy starting from local step m_0, which implies $\texttt{rew}_1(m_0) > \texttt{rew}_2(m_0)$ by Lemma 4.4, or (ii) DynamicGreedy deviates from StaticGreedy starting from local step m_0, which implies $\texttt{rew}_1(m_0) > \texttt{rew}_1(m_0 - 1)$ by Lemma A.4. In both cases, using (12) and (14), it follows that the inequality in (13) is strict for $t_0 = m_0$.

Therefore, agent m_0 chooses principal 1, and by Lemma 4.5 so do all subsequent agents. □

5 RELAXED RATIONALITY: HARDMAX & RANDOM

This section is dedicated to the HardMax&Random response model, where each principal is always chosen with some positive baseline probability. The main technical result for this model states that a principal with asymptotically better BIR wins by a large margin: after a "learning phase" of constant duration, all agents choose this principal with maximal possible probability $f_{\texttt{resp}}(1)$.

[6] For random variables X, Y on $\mathbb{R}$, we say that X *stochastically dominates* Y if $\Pr[X \geq x] \geq \Pr[Y \geq x]$ for any $x \in \mathbb{R}$.

For example, a principal with $\texttt{BIR}(n) \leq \tilde{O}(n^{-1/2})$ wins over a principal with $\texttt{BIR}(n) \geq \Omega(n^{-1/3})$. However, this positive result comes with a significant caveat detailed in Section 5.1.

We formulate and prove a cleaner version of the result, followed by a more general formulation developed in a subsequent Remark 5.2. We need to express a property that $\texttt{alg}_1$ eventually catches up and surpasses $\texttt{alg}_2$, even if initially it receives only a fraction of traffic. For the cleaner version, we assume that both algorithms are well-defined for an infinite time horizon, so that their BIR does not depend on the time horizon T of the game. Then this property can be formalized as:

$$(\forall \epsilon > 0) \qquad \texttt{BIR}_1(\epsilon n)/\texttt{BIR}_2(n) \to 0. \tag{15}$$

In fact, a weaker version of (15) suffices: denoting $\epsilon_0 = f_{\texttt{resp}}(-1)$, for some constant n_0 we have

$$(\forall n \geq n_0) \qquad \texttt{BIR}_1(\epsilon_0 n/2)/\texttt{BIR}_2(n) < \tfrac{1}{2}. \tag{16}$$

We also need a very mild technical assumption on the "bad" algorithm:

$$(\forall n \geq n_0) \qquad \texttt{BIR}_2(n) > 4\, e^{-\epsilon_0 n/12}. \tag{17}$$

Theorem 5.1. *Assume* HardMax&Random *response function. Suppose both algorithms are monotone and well-defined for an infinite time horizon, and satisfy* (16) *and* (17)*. Then each agent $t \geq n_0$ chooses principal* 1 *with maximal possible probability $f_{\texttt{resp}}(1) = 1 - \epsilon_0$.*

Proof. Consider global round $t \geq n_0$. Recall that each agent chooses principal 1 with probability at least $f_{\texttt{resp}}(-1) > 0$.

Then $\mathbb{E}[n_1(t+1)] \geq 2\epsilon_0\, t$. By Chernoff Bounds (Theorem 3.1), we have that $n_1(t+1) \geq \epsilon_0 t$ holds with probability at least $1 - q$, where $q = \exp(-\epsilon_0 t/12)$.

We need to prove that $\texttt{PMR}_1(t) - \texttt{PMR}_2(t) > 0$. For any m_1 and m_2, consider the quantity

$$\Delta(m_1, m_2) := \texttt{BIR}_2(m_2 + 1) - \texttt{BIR}_1(m_1 + 1).$$

Whenever $m_1 \geq \epsilon_0 t/2 - 1$ and $m_2 < t$, it holds that

$$\Delta(m_1, m_2) \geq \Delta(\epsilon_0 t/2, t) \geq \texttt{BIR}_2(t)/2.$$

The above inequalities follow, resp., from algorithms' monotonicity and (16). Now,

$$\begin{aligned}
\texttt{PMR}_1(t) - \texttt{PMR}_2(t) &= \mathop{\mathbb{E}}_{m_1 \sim \mathcal{N}_{1,t},\ m_2 \sim \mathcal{N}_{2,t}} [\Delta(m_1, m_2)] \\
&\geq -q + \mathop{\mathbb{E}}_{m_1 \sim \mathcal{N}_{1,t},\ m_2 \sim \mathcal{N}_{2,t}} [\Delta(m_1, m_2) \mid m_1 \geq \epsilon_0 t/2 - 1] \\
&\geq \texttt{BIR}_2(t)/2 - q \\
&> \texttt{BIR}_2(t)/4 > 0 \qquad \text{(by (17))}.
\end{aligned}$$

□

Remark 5.2. Many standard MAB algorithms in the literature are parameterized by the time horizon T. Regret bounds for such algorithms usually include a polylogarithmic dependence on T. In particular, a typical upper bound for BIR has the following form:

$$\texttt{BIR}(n \mid T) \leq \text{polylog}(T) \cdot n^{-\gamma} \quad \text{for some } \gamma \in (0, \tfrac{1}{2}]. \tag{18}$$

Here we write $\texttt{BIR}(n \mid T)$ to emphasize the dependence on T.

We generalize (16) to handle the dependence on T: there exists a number T_0 and a function $n_0(T) \in \text{polylog}(T)$ such that

$$(\forall T \geq T_0,\ n \geq n_0(T)) \quad \frac{\texttt{BIR}_1(\epsilon_0 n/2 \mid T)}{\texttt{BIR}_2(n \mid T)} < \frac{1}{2}. \tag{19}$$

If this holds, we say that $\texttt{alg}_1$ *BIR-dominates* $\texttt{alg}_2$.

We provide a version of Theorem 5.1 in which algorithms are parameterized with time horizon T and condition (16) is replaced with (19); its proof is very similar and is omitted.

To state a game-theoretic corollary of Theorem 5.1, we consider a version of the competition game between the two principals in which they can only choose from a finite set $\mathcal{A}$ of monotone MAB algorithms. One of these algorithms is "better" than all others; we call it the *special* algorithm. Unless specified otherwise, it BIR-dominates all other allowed algorithms. The other algorithms satisfy (17). We call this game the *restricted competition game*.

Corollary 5.3. *Assume* `HardMax&Random` *response function. Consider the restricted competition game with special algorithm* `alg`*. Then, for any sufficiently large time horizon T, this game has a unique Nash equilibrium: both principals choose* `alg`*.*

5.1 A little greedy goes a long way

Given any monotone MAB algorithm other than `DynamicGreedy`, we design a modified algorithm which learns at a slower rate, yet "wins the game" in the sense of Theorem 5.1. As a corollary, the competition game with unrestricted choice of algorithms typically does not have a Nash equilibrium.

Given an algorithm $\texttt{alg}_1$ that deviates from `DynamicGreedy` starting from step n_0 and a "mixing" parameter p, we will construct a modified algorithm as follows.

(1) The modified algorithm coincides with $\texttt{alg}_1$ (and `DynamicGreedy`) for the first $n_0 - 1$ steps;
(2) In each step $n \geq n_0$, $\texttt{alg}_1$ is invoked with probability $1-p$, and with the remaining probability p does the "greedy choice": chooses an action with the largest posterior mean reward given the current information collected by $\texttt{alg}_1$.

For a cleaner comparison between the two algorithms, the modified algorithm does not record rewards received in steps with the "greedy choice". Parameter $p > 0$ is the same for all steps.

Theorem 5.4. *Assume symmetric* `HardMax&Random` *response function. Let $\epsilon_0 = f_{\texttt{resp}}(-1)$ be the baseline probability. Suppose $\texttt{alg}_1$ deviates from* `DynamicGreedy` *starting from some step n_0. Let $\texttt{alg}_2$ be the modified algorithm, as described above, with mixing parameter p such that $(1-\epsilon_0)(1-p) > \epsilon_0$. Then each agent $t \geq n_0$ chooses principal 2 with maximal possible probability $1-\epsilon_0$.*

Corollary 5.5. *Suppose that both principals can choose any monotone MAB algorithm, and assume the symmetric* `HardMax&Random` *response function. Then for any time horizon T, the only possible* pure *Nash equilibrium is one where both principals choose* `DynamicGreedy`*. Moreover, no pure Nash equilibrium exists when some algorithm "dominates"* `DynamicGreedy` *in the sense of* (19) *and the time horizon T is sufficiently large.*

Remark 5.6. The modified algorithm performs exploration at a slower rate. Let us argue how this may translate into a larger `BIR` compared to the original algorithm. Let $\texttt{BIR}'_1(n)$ be the `BIR` of the "greedy choice" after after $n-1$ steps of $\texttt{alg}_1$. Then

$$\texttt{BIR}_2(n) = \mathop{\mathbb{E}}_{m \sim (n_0-1)+\texttt{Binomial}(n-n_0+1,1-p)} \left[(1-p)\cdot \texttt{BIR}_1(m) + p \cdot \texttt{BIR}'_1(m) \right]. \tag{20}$$

In this expression, m is the number of times $\texttt{alg}_1$ is invoked in the first n steps of the modified algorithm. Note that $\mathbb{E}[m] = n_0 - 1 + (n - n_0 + 1)(1-p) \geq (1-p)n$.

Suppose $\texttt{BIR}_1(n) = \beta n^{-\gamma}$ for some constants $\beta, \gamma > 0$. Further, assume $\texttt{BIR}'_1(n) \geq c\ \texttt{BIR}_1(n)$, for some $c > 1 - \gamma$. Then for all $n \geq n_0$ and small enough $p > 0$ it holds that:

$$\begin{aligned}
\texttt{BIR}_2(n) &\geq (1 - p + pc)\ \mathbb{E}[\ \texttt{BIR}_1(m)\ ] \\
\mathbb{E}[\ \texttt{BIR}_1(m)\ ] &\geq \texttt{BIR}_1(\ \mathbb{E}[m]\ ) && \text{(By Jensen's inequality)} \\
&\geq \texttt{BIR}_1(\ (1-p)n\ ) && \text{(since } \mathbb{E}[m] \geq n(1-p)\text{)} \\
&\geq \beta \cdot n^{-\gamma} \cdot (1-p)^{-\gamma} && \text{(plugging in } \texttt{BIR}_1(n) = \beta n^{-\gamma}\text{)} \\
&> \texttt{BIR}_1(n)\ \ (1 - p\gamma)^{-1} && \text{(since } (1-p)^\gamma < 1 - p\gamma\text{).} \\
\texttt{BIR}_2(n) &> \alpha \cdot \texttt{BIR}_1(n), && \text{where} \quad \alpha = \tfrac{1-p+pc}{1-p\gamma} > 1.
\end{aligned}$$

(In the above equations, all expectations are over m distributed as in (20).)

Proof of Theorem 5.4. Let $\texttt{rew}'_1(n)$ denote the Bayesian-expected reward of the "greedy choice" after after $n-1$ steps of $\texttt{alg}_1$. Note that $\texttt{rew}_1(\cdot)$ and $\texttt{rew}'_1(\cdot)$ are non-decreasing: the former because $\texttt{alg}_1$ is monotone and the latter because the "greedy choice" is only improved with an increasing set of observations. Therefore, the modified algorithm $\texttt{alg}_2$ is monotone by (20).

By definition of the "greedy choice," $\texttt{rew}_1(n) \leq \texttt{rew}'_1(n)$ for all steps n. Moreover, by Lemma 4.4, $\texttt{alg}_1$ has a strictly smaller $\texttt{rew}(n_0)$ compared to DynamicGreedy; so, $\texttt{rew}_1(n_0) < \texttt{rew}_2(n_0)$.

Let alg denote a copy of $\texttt{alg}_1$ that is running "inside" the modified algorithm $\texttt{alg}_2$. Let $m_2(t)$ be the number of global rounds before t in which the agent chooses principal 2 *and* alg is invoked; in other words, it is the number of agents seen by alg before global round t. Let $\mathcal{M}_{2,t}$ be the agents' posterior distribution for $m_2(t)$.

We claim that in each global round $t \geq n_0$, distribution $\mathcal{M}_{2,t}$ stochastically dominates distribution $\mathcal{N}_{1,t}$, and $\texttt{PMR}_1(t) < \texttt{PMR}_2(t)$. We use induction on t. The base case $t = n_0$ holds because $\mathcal{M}_{2,t} = \mathcal{N}_{1,t}$ (because the two algorithms coincide on the first $n_0 - 1$ steps), and $\texttt{PMR}_1(n_0) < \texttt{PMR}_2(n_0)$ is proved as in (10), using the fact that $\texttt{rew}_1(n_0) < \texttt{rew}_2(n_0)$.

The induction step is proved as follows. The induction hypothesis for global round $t-1$ implies that agent $t-1$ is seen by alg with probability $(1-\epsilon_0)(1-p)$, which is strictly larger than ϵ_0, the probability with which this agent is seen by $\texttt{alg}_2$. Therefore, $\mathcal{M}_{2,t}$ stochastically dominates $\mathcal{N}_{1,t}$.

$$\begin{aligned}
\texttt{PMR}_1(t) &= \mathop{\mathbb{E}}_{n \sim \mathcal{N}_{1,t}} [\texttt{rew}_1(n+1)] \\
&\leq \mathop{\mathbb{E}}_{m \sim \mathcal{M}_{2,t}} [\texttt{rew}_1(m+1)] && (21) \\
&< \mathop{\mathbb{E}}_{m \sim \mathcal{M}_{2,t}} \left[(1-p) \cdot \texttt{rew}_1(m+1) + p \cdot \texttt{rew}'_1(m+1)\right] && (22) \\
&= \texttt{PMR}_2(t).
\end{aligned}$$

Here inequality (21) holds because $\texttt{rew}_1(\cdot)$ is monotone and $\mathcal{M}_{2,t}$ stochastically dominates $\mathcal{N}_{1,t}$, and inequality (22) holds because $\texttt{rew}_1(n_0) < \texttt{rew}_2(n_0)$ and $\mathcal{M}_{2,t}(n_0) > 0$.[7] □

6 SOFTMAX RESPONSE FUNCTION

This section is devoted to the SoftMax model. We recover a positive result under the assumptions from Theorem 5.1 (albeit with a weaker conclusion), and then proceed to a much more challenging result under weaker assumptions. We start with a formal definition:

Definition 6.1. A response function $f_{\texttt{resp}}$ is SoftMax if the following conditions hold:

- $f_{\texttt{resp}}(\cdot)$ is bounded away from 0 and 1: $f_{\texttt{resp}}(\cdot) \in [\epsilon, 1-\epsilon]$ for some $\epsilon \in (0, \frac{1}{2})$,

[7] If $\texttt{rew}_1(\cdot)$ is strictly increasing, then inequality (21) is strict, too; this is because $\mathcal{M}_{2,t}(t-1) > \mathcal{N}_{1,t}(t-1)$.

- the response function $f_{\texttt{resp}}(\cdot)$ is "smooth" around 0:

$$\exists \text{ constants } \delta_0, c_0, c_0' > 0 \qquad \forall x \in [-\delta_0, \delta_0] \qquad c_0 \le f'_{\texttt{resp}}(x) \le c_0'. \tag{23}$$

- fair tie-breaking: $f_{\texttt{resp}}(0) = \frac{1}{2}$.

Remark 6.2. This definition is fruitful when parameters c_0 and c_0' are close to $\frac{1}{2}$. Throughout, we assume that $\texttt{alg}_1$ is better than $\texttt{alg}_2$, and obtain results parameterized by c_0. By symmetry, one could assume that $\texttt{alg}_2$ is better than $\texttt{alg}_1$, and obtain similar results parameterized by c_0'.

Our first result is a version of Theorem 5.1, with the same assumptions about the algorithms and essentially the same proof. The conclusion is much weaker: we can only guarantee that each agent $t \ge n_0$ chooses principal 1 with probability slightly larger than $\frac{1}{2}$. This is essentially unavoidable in a typical case when both algorithms satisfy $\texttt{BIR}(n) \to 0$, by Definition 6.1.

Theorem 6.3. *Assume* SoftMax *response function. Suppose* $\texttt{alg}_1$ *has better* BIR *in the sense of* (16)*, and* $\texttt{alg}_2$ *satisfies the condition* (17)*. Then each agent* $t \ge n_0$ *chooses principal* 1 *with probability*

$$\Pr[i_t = 1] \ge \tfrac{1}{2} + \tfrac{c_0}{4}\, \texttt{BIR}_2(t). \tag{24}$$

Proof Sketch. We follow the steps in the proof of Theorem 5.1 to derive

$$\texttt{PMR}_1(t) - \texttt{PMR}_2(t) \ge \texttt{BIR}_2(t)/2 - q, \quad \text{where } q = \exp(-\epsilon_0 t/12).$$

This is at least $\texttt{BIR}_2(t)/4$ by (17). Then (24) follows by the smoothness condition (23). □

We recover a version of Corollary 5.3, if each principal's utility is the number of users (rather than the more general model in (6)). We also need a mild technical assumption that cumulative Bayesian regret (BReg) tends to infinity. BReg is a standard notion from the literature (along with BIR):

$$\texttt{BReg}(n) := n \cdot \mathop{\mathbb{E}}_{\mu \sim \mathcal{P}_{\texttt{mean}}} \left[\max_{a \in A} \mu_a \right] - \sum_{n=1}^{n} \texttt{rew}(n') = \sum_{n'=1}^{n} \texttt{BIR}(n'). \tag{25}$$

Corollary 6.4. *Assume that the response function is* SoftMax*, and each principal's utility is the number of users. Consider the restricted competition game with special algorithm* alg*, and assume that all other allowed algorithms satisfy* $\texttt{BReg}(n) \to \infty$*. Then, for any sufficiently large time horizon* T*, this game has a unique Nash equilibrium: both principals choose* alg*.*

Further, we prove a much more challenging result in which the condition (16) is replaced with a much weaker "BIR-dominance" condition. For clarity, we will again assume that both algorithms are well-defined for an infinite time horizon. The *weak BIR dominance* condition says there exist constants $\beta_0, \alpha_0 \in (0, 1/2)$ and n_0 such that

$$(\forall n \ge n_0) \quad \frac{\texttt{BIR}_1((1-\beta_0)\, n)}{\texttt{BIR}_2(n)} < 1 - \alpha_0. \tag{26}$$

If this holds, we say that $\texttt{alg}_1$ *weakly BIR-dominates* $\texttt{alg}_2$. Note that the condition (19) involves sufficiently small multiplicative factors (resp., $\epsilon_0/2$ and $\frac{1}{2}$), the new condition replaces them with factors that can be arbitrarily close to 1.

We make a mild assumption on $\texttt{alg}_1$ that its $\texttt{BIR}_1(n)$ tends to 0. Formally, for any $\epsilon > 0$, there exists some $n(\epsilon)$ such that

$$(\forall n \ge n(\epsilon)) \qquad \texttt{BIR}_1(n) \le \epsilon. \tag{27}$$

We also require a slightly stronger version of the technical assumption (17):for some n_0,

$$(\forall n \ge n_0) \qquad \texttt{BIR}_2(n) \ge \frac{4}{\alpha_0} \exp\left(\frac{-\min\{\epsilon_0, 1/8\} n}{12} \right) \tag{28}$$

Theorem 6.5. *Assume the* SoftMax *response function. Suppose* $\texttt{alg}_1$ *weakly-BIR-dominates* $\texttt{alg}_2$, $\texttt{alg}_1$ *satisfies* (27), *and* $\texttt{alg}_2$ *satisfies* (28). *Then there exists some* t_0 *such that each agent* $t \geq t_0$ *chooses principal* 1 *with probability*

$$\Pr[i_t = 1] \geq \tfrac{1}{2} + \tfrac{c_0 \alpha_0}{4}\, \texttt{BIR}_2(t). \tag{29}$$

The main idea behind our proof is that even though $\texttt{alg}_1$ may have a slower rate of learning in the beginning, it will gradually catch up and surpass $\texttt{alg}_2$. We will describe this process in two phases. In the first phase, $\texttt{alg}_1$ receives a random agent with probability at least $f_{\texttt{resp}}(-1) = \epsilon_0$ in each round. Since $\texttt{BIR}_1$ tends to 0, the difference in BIRs between the two algorithms is also diminishing. Due to the SoftMax response function, $\texttt{alg}_1$ attracts each agent with probability at least $1/2 - O(\beta_0)$ after a sufficient number of rounds. Then the game enters the second phase: both algorithms receive agents at a rate close to $\frac{1}{2}$, and the fractions of agents received by both algorithms — $n_1(t)/t$ and $n_2(t)/t$ — also converge to $\frac{1}{2}$. At the end of the second phase and in each global round afterwards, the counts $n_1(t)$ and $n_2(t)$ satisfy the weak BIR-dominance condition, in the sense that they both are larger than n_0 and $n_1(t) \geq (1-\beta_0)\, n_2(t)$. At this point, $\texttt{alg}_1$ actually has smaller BIR, which reflected in the PMRs eventually. Accordingly, from then on $\texttt{alg}_1$ attracts agents at a rate slightly larger than $\frac{1}{2}$. We prove that the "bump" over $\frac{1}{2}$ is at least on the order of $\texttt{BIR}_2(t)$.

Proof of Theorem 6.5. Let $\beta_1 = \min\{c_0' \delta_0, \beta_0/20\}$ with δ_0 defined in (23). Recall each agent chooses $\texttt{alg}_1$ with probability at least $f_{\texttt{resp}}(-1) = \epsilon_0$. By condition (27) and (28), there exists some sufficiently large T_1 such that for any $t \geq T_1$, $\texttt{BIR}_1(\epsilon_0 T_1/2) \leq \beta_1/c_0'$ and $\texttt{BIR}_2(t) > e^{-\epsilon_0 t/12}$. Moreover, for any $t \geq T_1$, we know $\mathbb{E}[n_1(t+1)] \geq \epsilon_0\, t$, and by the Chernoff Bounds (Theorem 3.1), we have $n_1(t+1) \geq \epsilon_0 t/2$ holds with probability at least $1 - q_1(t)$ with $q_1(t) = \exp(-\epsilon_0 t/12) < \texttt{BIR}_2(t)$. It follows that for any $t \geq T_1$,

$$\begin{aligned}
\texttt{PMR}_2(t) - \texttt{PMR}_1(t) &= \mathop{\mathbb{E}}_{m_1 \sim \mathcal{N}_{1,t},\, m_2 \sim \mathcal{N}_{2,t}} [\texttt{BIR}_1(m_1+1) - \texttt{BIR}_2(m_2+1)] \\
&\leq q_1(t) + \mathop{\mathbb{E}}_{m_1 \sim \mathcal{N}_{1,t}} [\texttt{BIR}_1(m_1+1) \mid m_1 \geq \epsilon_0 t/2 - 1] - \texttt{BIR}_2(t) \\
&\leq \texttt{BIR}_1(\epsilon_0 T_1/2) \leq \beta_1/c_0'
\end{aligned}$$

Since the response function $f_{\texttt{resp}}$ is c_0'-Lipschitz in the neighborhood of $[-\delta_0, \delta_0]$, each agent after round T_1 will choose $\texttt{alg}_1$ with probability at least

$$p_t \geq \frac{1}{2} - c_0' \left(\texttt{PMR}_2(t) - \texttt{PMR}_1(t)\right) \geq \frac{1}{2} - \beta_1.$$

Next, we will show that there exists a sufficiently large T_2 such that for any $t \geq T_1 + T_2$, with high probability $n_1(t) > \max\{n_0, (1-\beta_0) n_2(t)\}$, where n_0 is defined in (26). Fix any $t \geq T_1 + T_2$. Since each agent chooses $\texttt{alg}_1$ with probability at least $1/2 - \beta_1$, by Chernoff Bounds (Theorem 3.1) we have with probability at least $1 - q_2(t)$ that the number of agents that choose $\texttt{alg}_1$ is at least $\beta_0(1/2 - \beta_1)t/5$, where the function

$$q_2(x) = \exp\left(\frac{-(1/2-\beta_1)(1-\beta_0/5)^2 x}{3}\right).$$

Note that the number of agents received by $\texttt{alg}_2$ is at most $T_1 + (1/2 + \beta_1)t + (1/2 - \beta_1)(1 - \beta_0/5)t$.

Then as long as $T_2 \geq \frac{5T_1}{\beta_0}$, we can guarantee that $n_1(t) > n_2(t)(1-\beta_0)$ and $n_1(t) > n_0$ with probability at least $1 - q_2(t)$ for any $t \geq T_1 + T_2$. Note that the weak BIR-dominance condition in (26) implies that for any $t \geq T_1 + T_2$ with probability at least $1 - q_2(t)$,

$$\texttt{BIR}_1(n_1(t)) < (1-\alpha_0)\texttt{BIR}_2(n_2(t)).$$

It follows that for any $t \geq T_1 + T_2$,

$$\begin{aligned}\texttt{PMR}_1(t) - \texttt{PMR}_2(t) &= \mathop{\mathbb{E}}_{m_1 \sim \mathcal{N}_{1,t},\, m_2 \sim \mathcal{N}_{2,t}} [\texttt{BIR}_2(m_2+1) - \texttt{BIR}_1(m_1+1)] \\ &\geq (1 - q_2(t))\alpha_0 \texttt{BIR}_2(t) - q_2(t) \\ &\geq \alpha_0 \texttt{BIR}_2(t)/4\end{aligned}$$

where the last inequality holds as long as $q_2(t) \leq \alpha_0 \texttt{BIR}_2(t)/4$, and is implied by the condition in (28) as long as T_2 is sufficiently large. Hence, by the definition of our `SoftMax` response function and assumption in (23), we have

$$\Pr[i_t = 1] \geq \frac{1}{2} + \frac{c_0 \alpha_0 \texttt{BIR}_2(t)}{4}. \qquad \square$$

Similar to the condition (16), we can also generalize the weak BIR-dominance condition (26) to handle the dependence on T: there exist some T_0, a function $n_0(T) \in \text{polylog}(T)$, and constants $\beta_0, \alpha_0 \in (0, 1/2)$, such that

$$(\forall T \geq T_0, n \geq n_0(T)) \quad \frac{\texttt{BIR}_1((1-\beta_0)\, n \mid T)}{\texttt{BIR}_2(n \mid T)} < 1 - \alpha_0. \tag{30}$$

We also provide a version of Theorem 6.3 under this more general weak BIR-dominance condition; its proof is very similar and is omitted. The following is just a direct consequence of Theorem 6.3 with this general condition

Corollary 6.6. *Assume that the response function is* `SoftMax`*, and each principal's utility is the number of users. Consider the restricted competition game in which the special algorithm* `alg` *weakly-BIR-dominates the other allowed algorithms, and the latter satisfy* $\texttt{BReg}(n) \to \infty$*. Then, for any sufficiently large time horizon* T*, there is a unique Nash equilibrium: both principals choose* `alg`*.*

7 REPUTATION MODEL DETAILS

In this section we move to the reputation choice model. Recall that in this variant of the model, instead of making choices between firms based on the Bayesian expected reward, each agent chooses the firm with a maximal reputation score (breaking ties uniformly). The reputation score is simply a sliding window average: an average reward of the last M agents that chose this firm. We focus on numerical investigation of this model instead of analytical characterizations of equilibrium strategies.

Key Model Differences The timing of the model is the same as before where each firm commits to a MAB algorithm before the game starts and uses this algorithm to choose its actions. We focus on i.i.d. Bernoulli rewards: the reward of each arm a is drawn from $\{0, 1\}$ independently with expectation $\mu(a)$. The mean rewards $\mu(a)$ are the same for all rounds and both firms, but initially unknown. However, instead of starting from some prior, we suppose that each firm has a uniform, "fake", prior and that the initial information set of the firm is given by a "warm start". Each algorithm receives a "warm start": additional T_0 agents that arrive before the game starts, and interact with the firm as described above. The warm start ensures that each firm has a meaningful reputation and initial prior when competition starts..

In some of our experiments, one firm is the "incumbent" who enters the market before the other ("late entrant"), and therefore enjoys a *first-mover advantage*. Formally, the incumbent enjoys additional X rounds of the "warm start". We treat X as an exogenous element of the model, and study the consequences for a fixed X.

MAB algorithms. We consider the same three classes of bandit algorithms discussed in Section 3, but take a representative algorithm from each class and look for qualitative differences between the different classes. We will utilize Thompson Sampling (TS) from the "smart" MAB algorithms, ϵ-greedy (DEG)[8] from the "naive" algorithms, and DynamicGreedy (DG). For ease of comparison, all three algorithms are parameterized with the same fake prior: namely, the mean reward of each arm is drawn independently from a Beta(1, 1) distribution. Recall that Beta priors with 0-1 rewards form a conjugate family, which allows for simple posterior updates.

MAB instances. We consider instances with $K = 10$ arms. Since we focus on 0-1 rewards, an instance of the MAB problem is specified by the *mean reward vector* $(\mu(a) : a \in A)$. Initially this vector is drawn from some distribution, termed *MAB instance*. We consider three MAB instances:

(1) *Needle-In-Haystack*: one arm (the "needle") is chosen uniformly at random. This arm has mean reward .7, and the remaining ones have mean reward .5.
(2) *Uniform instance*: the mean reward of each arm is drawn independently and uniformly from $[1/4, 3/4]$.
(3) *Heavy-Tail instance*: the mean reward of each arm is drawn independently from Beta(.6, .6) distribution (which is known to have substantial "tail probabilities").

We argue that these MAB instances are (somewhat) representative. Consider the "gap" between the best and the second-best arm, an essential parameter in the literature on MAB. The "gap" is fixed in Needle-in-Haystack, spread over a wide spectrum of values under the Uniform instance, and is spread but focused on the large values under the Heavy-Tail instance. We also ran smaller experiments with versions of these instances, and achieved similar qualitative results.

Simulation details. For each MAB instance we draw $N = 1000$ mean reward vectors independently from the corresponding distribution. We use this same collection of mean reward vectors for all experiments with this MAB instance. For each mean reward vector we draw a table of realized rewards (*realization table*), and use this same table for all experiments on this mean reward vector. This ensures that differences in algorithm performance are not due to noise in the realizations but due to differences in the algorithms in the different experimental settings.

More specifically, the realization table is a 0-1 matrix W with K columns which correspond to arms, and $T + T_{\max}$ rows, which correspond to rounds. Here $T_{\max}$ is the maximal duration of the "warm start" in our experiments, *i.e.,* the maximal value of $X + T_0$. For each arm a, each value $W(\cdot, a)$ is drawn independently from Bernoulli distribution with expectation $\mu(a)$. Then in each experiment, the reward of this arm in round t of the warm start is taken to be $W(t, a)$, and its reward in round t of the game is $W(T_{\max} + t, a)$.

We fix the sliding window size $M = 100$. We found that lower values induced too much random noise in the results, and increasing M further did not make a qualitative difference. Unless otherwise noted, we used $T = 2000$.

Terminology. Following a standard game-theoretic terminology, algorithm Alg1 *(weakly) dominates* algorithm Alg2 for a given firm if Alg1 provides a larger (or equal) market share than Alg2 at the end of the game. An algorithm is a (weakly) dominant strategy for the firm if it (weakly) dominates all other algorithms. This is for a particular MAB instance and a particular selection of the game parameters.

Discussion While we experiment with various MAB instances and parameter settings, we only report on selected, representative experiments in the body of the paper. Additional plots and tables

[8]Throughout, we fix $\epsilon = 0.05$. Our pilot experiments showed that different ϵ did not qualitatively change the results.

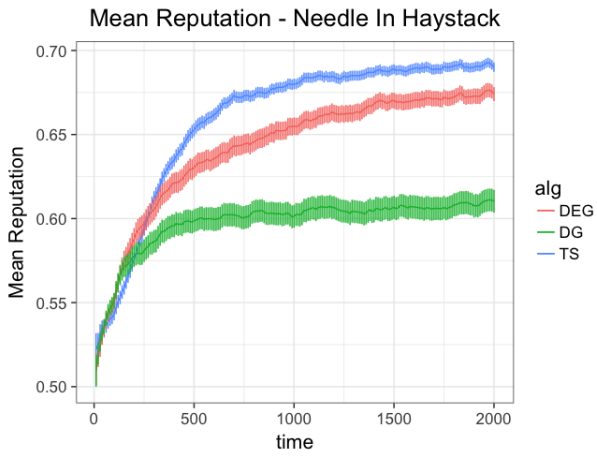


Fig. 3. Mean reputation trajectories for Needle-in-Haystack. The shaded area shows 95% confidence intervals.

are provided in the appendix. Unless noted otherwise, our findings are based on and consistent with all these experiments.

Even with a stylized model, numerical investigation is quite challenging. An "atomic experiment" is a competition game between a given pair of bandit algorithms, in a given competition model, on a given instance of a multi-armed bandit problem.[9] Accordingly, we have a three-dimensional space of atomic experiments one needs to run and interpret: {pairs of algorithms} x {competition models} x {bandit instances}, and we are looking for findings that are consistent across this entire space. It is essential to keep each of the three dimensions small yet representative. In particular, we need to capture a huge variety of bandit instances with only a few representative examples. Further, one needs succinct and informative summarization of results within one atomic experiment and across multiple experiments (*e.g.*, see Table 1).

The simulations are computationally intensive. An experiment on a particular MAB instance comprised multiple runs of the competition game: N mean reward vectors times 9 pairs of algorithms times three values for the warm start. We used a parallel implementation over a cluster of 12 2.2 GHz CPU cores, with 8 GB RAM per core. Each experiment took about 10 hours.

8 PERFORMANCE IN ISOLATION

We start with a pilot experiment in which we investigate each algorithm's performance "in isolation": in a stand-alone MAB problem without competition. We focus on reputation scores generated by each algorithm. We confirm that algorithms' performance is ordered as we'd expect: TS > DEG > DG for a sufficiently long time horizon. For each algorithm and each MAB instance, we compute the mean reputation score at each round, averaged over all mean reward vectors. We plot the *mean reputation trajectory*: how this score evolves over time. Figure 3 shows such a plot for the Needle-in-Haystack instance; for other MAB instances the plots are similar. We summarize this finding as follows:

Finding 1. *The mean reputation trajectories are arranged as predicted by prior work:* TS > DEG > DG *for a sufficiently long time horizon.*

[9]Each such experiment is run many times to reduce variance.

We also use Figure 3 to choose a reasonable time-horizon for the subsequent experiments, as $T = 2000$. The idea is, we want T to be large enough so that algorithms performance starts to plateau, but small enough such that algorithms are still learning.

The mean reputation trajectory is probably the most natural way to represent an algorithm's performance on a given MAB instance. However, we found that the outcomes of the competition game are better explained with a different "performance-in-isolation" statistic that is more directly connected to the game. Consider the performance of two algorithms, Alg1 and Alg2, "in isolation" on a particular MAB instance. The *relative reputation* of Alg1 (vs. Alg2) at a given time t is the fraction of mean reward vectors/realization tables for which Alg1 has a higher reputation score than Alg2. The intuition is that agent's selection in our model depends only on the comparison between the reputation scores.

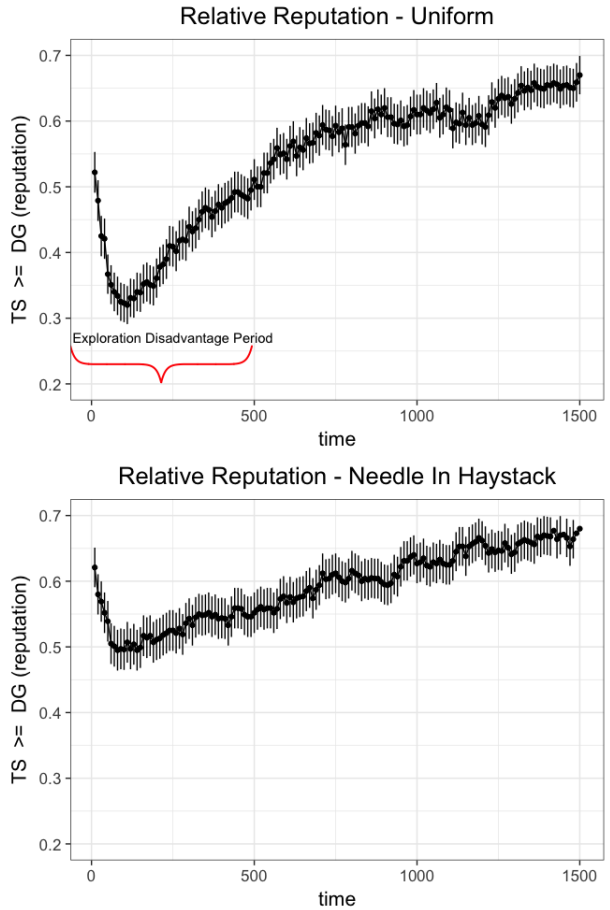


Fig. 4. Relative reputation trajectory for TS vs DG, on Uniform instance (top) and Needle-in-Haystack instance (bottom). Shaded area display 95% confidence intervals.

This angle allows a more nuanced analysis of reputation costs vs. benefits under competition. Figure 4 (top) shows the relative reputation trajectory for TS vs DG for the Uniform instance. The relative reputation is less than $\frac{1}{2}$ in the early rounds, meaning that DG has a higher reputation score

	Heavy-Tail			Needle-in-Haystack		
	$T_0 = 20$	$T_0 = 250$	$T_0 = 500$	$T_0 = 20$	$T_0 = 250$	$T_0 = 500$
TS vs DG	**0.29** ±0.03 EoG 55 (0)	**0.72** ±0.02 EoG 570 (0)	**0.76** ±0.02 EoG 620 (99)	**0.64** ±0.03 EoG 200 (27)	**0.6** ±0.03 EoG 370 (0)	**0.64** ±0.03 EoG 580 (122)
TS vs DEG	**0.3** ±0.03 EoG 37 (0)	**0.88** ±0.01 EoG 480 (0)	**0.9** ±0.01 EoG 570 (114)	**0.57** ±0.03 EoG 150 (14)	**0.52** ±0.03 EoG 460 (79)	**0.56** ±0.02 EoG 740 (628)
DG vs DEG	**0.62** ±0.03 EoG 410 (7)	**0.6** ±0.02 EoG 790 (762)	**0.57** ±0.03 EoG 730 (608)	**0.46** ±0.03 EoG 340 (129)	**0.42** ±0.02 EoG 650 (408)	**0.42** ±0.02 EoG 690 (467)

Table 1. **Simultaneous Entry Duopoly**, for Heavy-Tail and Needle-in-Haystack instances. Each cell describes a game between two algorithms, call them Alg1 vs. Alg2, for a particular value of the warm start T_0. Line 1 in the cell is the market share of Alg 1: the average (in bold) and the 95% confidence band. Line 2 specifies the "effective end of game" (EoG): the average and the median (in brackets). The time horizon is $T = 2000$.

in a majority of the simulations, and more than $\frac{1}{2}$ later on. The reason is the exploration in TS leads to worse decisions initially, but allows for better decisions later. The time period when relative reputation vs. DG dips below $\frac{1}{2}$ can be seen as an explanation for the competitive disadvantage of exploration. Such period also exists for the Heavy-Tail MAB instance. However, it does not exist for the Needle-in-Haystack instance, see Figure 4 (bottom).[10]

Finding 2. *Exploration can lead to relative reputation vs.* DG *going below* $\frac{1}{2}$ *for some initial time period. This happens for some MAB instances but not for some others.*

Definition 8.1. For a particular MAB algorithm, a time period when relative reputation vs. DG goes below $\frac{1}{2}$ is called *exploration disadvantage period.* An MAB instance is called *exploration-disadvantaged* if such period exists.

Uniform and Heavy-tail instance are exploration-disadvantaged, but Needle-in-Haystack is not.

9 COMPETITION VS. BETTER ALGORITHMS

Our main experiments are with the duopoly game defined in Section 3. As the "intensity of competition" varies from monopoly to "incumbent" to simultaneous entry duopoly to "late entrant", we find a stylized inverted-U relationship as in Figure 1. More formally, we look for equilibria in the duopoly game, where each firm's choices are limited to DG, DEG and TS. We do this for each "intensity level" and each MAB instance, and look for findings that are consistent across MAB instances. For cleaner results, we break ties towards less advanced algorithms (as they tend to have lower adoption costs [1, 2]). Note that DG is trivially the dominant strategy under monopoly.

Simultaneous entry duopoly. The basic scenario is when both firms are competing from round 1. A crucial distinction is whether an MAB instance is exploration-disadvantaged:

Finding 3. *Under simultaneous entry duopoly:*

(a) *(*DG,DG*) is the unique pure-strategy Nash equilibrium for exploration-disadvantaged MAB instances with a sufficiently small "warm start".*
(b) *This is not necessarily the case for MAB instances that are not exploration-disadvantaged. In particular,* TS *is a weakly dominant strategy for Needle-in-Haystack.*

[10] We see two explanations for this: TS identifies the best arm faster for the Needle-in-Haystack instance, and there are no "very bad" arms which make exploration very expensive in the short term.

We investigate the firms' market shares when they choose different algorithms (otherwise, by symmetry both firms get half of the agents). We report the market shares for Heavy-Tail and Needle-in-Haystack instances in Table 1 (see the first line in each cell), for a range of values of the warm start T_0. Table 2 reports similarly on the Uniform instance. We find that DG is a weakly dominant strategy for the Heavy-Tail and Uniform instances, as long as T_0 is sufficiently small. However, TS is a weakly dominant strategy for the Needle-in-Haystack instance. We find that for a sufficiently small T_0, DG yields more than half the market against TS, but achieves similar market share vs. DG and DEG. By our tie-breaking rule, (DG,DG) is the only pure-strategy equilibrium.

	$T_0 = 20$	$T_0 = 250$	$T_0 = 500$
TS vs DG	**0.46** ±0.03	**0.52** ±0.02	**0.6** ±0.02
TS vs DEG	**0.41** ±0.03	**0.51** ±0.02	**0.55** ±0.02
DG vs DEG	**0.51** ±0.03	**0.48** ±0.02	**0.45** ±0.02

Table 2. **Simultaneous entry duopoly**, for the Uniform MAB instance. Semantics are the same as in Table 1.

We attribute the prevalence of DG on exploration-disadvantaged MAB instances to its prevalence on the initial "exploration disadvantage period", as described in Section 8. Increasing the warm start length T_0 makes this period shorter: indeed, considering relative reputation trajectory in Figure 4 (top), increasing T_0 effectively shifts the starting time point to the right. This is why it helps DG if T_0 is small.

Temporary Monopoly. We turn our attention to the temporary monopoly scenario. Recall that the incumbent firm enters the market and serves as a monopolist until the entrant firm enters at round X. We make X large enough, but still much smaller than the time horizon T. We find that the incumbent is incentivized to choose TS, in a strong sense:

Finding 4. *Under temporary monopoly,* TS *is the dominant strategy for the incumbent. This holds across all MAB instances, if X is large enough.*

The simulation results for the Heavy-Tail MAB instance are reported in Table 3, for a particular $X = 200$. We see that TS is a dominant strategy for the incumbent. Similar tables for the other MAB instances and other values of X are reported in the supplement, with the same conclusion.

	TS	DEG	DG
TS	**0.003**±0.003	**0.083**±0.02	**0.17**±0.02
DEG	**0.045**±0.01	**0.25**±0.02	**0.23**±0.02
DG	**0.12**±0.02	**0.36**±0.03	**0.3**±0.02

User share of row player (entrant), 200 round head-start, Heavy-Tail Instance

DG is a weakly dominant strategy for the entrant, for Heavy-Tail instance in Table 3 and the Uniform instance, but not for the Needle-in-Haystack instance. We attribute this finding to exploration-disadvantaged property of these two MAB instance, for the same reasons as discussed above.

Finding 5. *Under temporary monopoly,* DG *is a weakly dominant strategy for the entrant for exploration-disadvantaged MAB instances.*

Inverted-U relationship. We interpret our findings through the lens of the inverted-U relationship between the "intensity of competition" and the "quality of technology". The lowest level of competition is monopoly, when DG wins out for the trivial reason of tie-breaking. The highest levels are simultaneous entry duopoly and "late entrant". We see that DG is incentivized for exploration-disadvantaged MAB instances. In fact, incentives for DG get stronger when the model transitions from simultaneous entry duopoly to "late entrant".[11] Finally, the middle level of competition, "incumbent" in the temporary monopoly creates strong incentives for TS. In stylized form, this relationship is captured in Figure 1.

Our intuition for why incumbency creates more incentives for exploration is as follows. During the temporary monopoly period, reputation consequences of exploration vanish. Instead, the firm wants to improve its performance as much as possible by the time competition starts. Essentially, the firm only faces a classical explore-exploit trade-off, and is incentivized to choose algorithms that are best at optimizing this trade-off.

Death spiral effect. Further, we investigate the "death spiral" effect mentioned in the Introduction. Restated in terms of our model, the effect is that one firm attracts new customers at a lower rate than the other, and falls behind in terms of performance because the other firm has more customers to learn from, and this gets worse over time until (almost) all new customers go to the other firm. With this intuition in mind, we define *effective end of game* (EoG) for a particular mean reward vector and realization table, as the last round t such that the agents at this and previous round choose different firms. Indeed, the game, effectively, ends after this round. We interpret low EoG as a strong evidence of the "death spiral" effect. Focusing on the simultaneous entry duopoly scenario, we specify the EoG values in Table 1 (the second line of each cell). We find that the EoG values are indeed small:

Finding 6. *Under simultaneous entry duopoly,* EoG *values tend to be much smaller than the time horizon T.*

We also see that the EoG values tend to increase as the warm start T_0 increases. We conjecture this is because larger T_0 tends to be more beneficial for a better algorithm (as it tends to follow a better learning curve). Indeed, we know that the "effective end of game" in this scenario typically occurs when a better algorithm loses, and helping it delays the loss.

Welfare implications. We study the effects of competition on consumer welfare: the total reward collected by the users over time. Rather than welfare directly, we find it more lucid to consider *market regret*:

$$T \max_a \mu(a) - \sum_{t\in[T]} \mu(a_t),$$

where a_t is the arm chosen by agent t. This is a standard performance measure in the literature on multi-armed bandits. Note that smaller regret means higher welfare.

We assume that both firms play their respective equilibrium strategies for the corresponding competition level. As discussed previously, these are:

- DG in the monopoly,
- DG for both firms in duopoly (Finding 3),
- TS for the incumbent (Finding 4) and DG for the entrant in temporary monopoly (Finding 5).

[11] For the Heavy-Tail instance, DG goes from a weakly dominant strategy to a strictly dominant one. For the Uniform instance, DG goes from a Nash equilibrium strategy to a weakly dominant one.

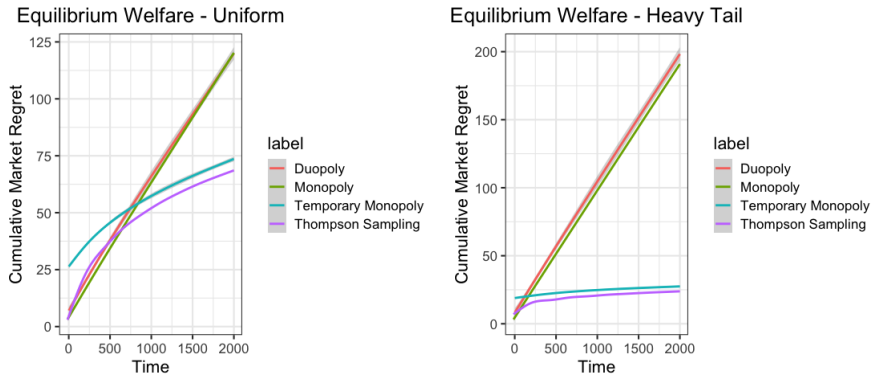


Fig. 5. Smoothed welfare plots resulting from equilibrium strategies in the different market structures. Note that welfare at $t = 0$ incorporates the regret incurred during the incumbent and warm start periods.

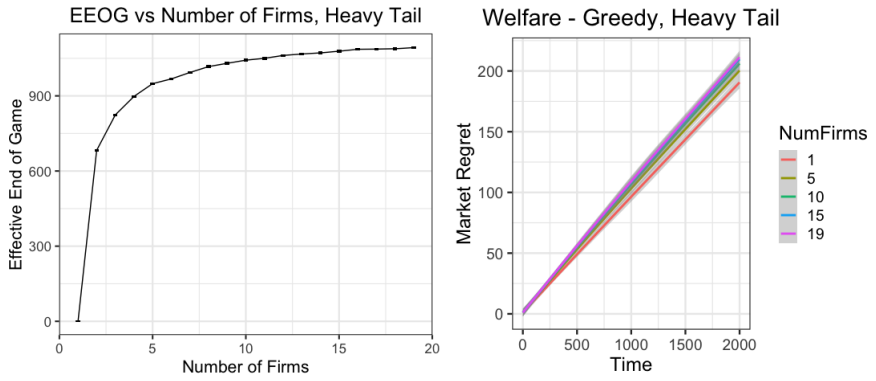


Fig. 6. Average welfare and EoG as we increase #firms playing DG

Figure 5 displays the market regret (averaged over multiple runs) under different levels of competition. Consumers are *better off* in the temporary monopoly case than in the duopoly case. Recall that under temporary monopoly, the incumbent is incentivized to play TS. Moreover, we find that the welfare is close to that of having a single firm for all agents and running TS. We also observe that monopoly and duopoly achieve similar welfare.

Finding 7. *In equilibrium, consumer welfare is (a) highest under temporary monopoly, (b) similar for monopoly and duopoly.*

Finding 7(b) is interesting because, in equilibrium, both firms play DG in both settings, and one might conjecture that the welfare should increase with the number of firms playing DG. Indeed, one run of DG may get stuck on a bad arm. However, two firms independently playing DG are less likely to get stuck simultaneously. If one firm gets stuck and the other does not, then the latter should attract most agents, leading to improved welfare.

To study this phenomenon further, we go beyond the duopoly setting to more than two firms playing DG (and starting at the same time). Figure 6 reports the average welfare across these simulations. Welfare not only does not get better, *but is weakly worse* as we increase the number of firms.

Finding 8. *When all firms deploy* DG*, and start at the same time, welfare is weakly decreasing as the number of firms increases*

We track the average EoG in each of the simulations and notice that it *increases* with the number of firms. This observation also runs counter of the intuition that with more firms running DG, one of them is more likely to "get lucky" and take over the market (which would cause EoG to *decrease* with the number of firms).

10 DATA AS A BARRIER TO ENTRY

Under temporary monopoly, the incumbent can explore without incurring immediate reputational costs, and build up a high reputation before the entrant appears. Thus, the early entry gives the incumbent both a *data* advantage and a *reputational* advantage over the entrant. We explore which of the two factors is more significant. Our findings provide a quantitative insight into the role of the classic "first mover advantage" phenomenon in the digital economy.

For a more succinct terminology, recall that the incumbent enjoys an extended warm start of $X + T_0$ rounds. Call the first X of these rounds the *monopoly period* (and the rest is the proper "warm start"). The rounds when both firms are competing for customers are called *competition period.*

We run two additional experiments to isolate the effects of the two advantages mentioned above. The *data-advantage experiment* focuses on the data advantage by, essentially, erasing the reputation advantage. Namely, the data from the monopoly period is not used in the computation of the incumbent's reputation score. Likewise, the *reputation-advantage experiment* erases the data advantage and focuses on the reputation advantage: namely, the incumbent's algorithm 'forgets' the data gathered during the monopoly period.

We find that either data or reputational advantage alone gives a substantial boost to the incumbent, compared to simultaneous entry duopoly. The results for the Heavy-Tail instance are presented in Table 4, in the same structure as Table 3. For the other two instances, the results are qualitatively similar.

	Reputation advantage (only)			Data advantage (only)		
	TS	DEG	DG	TS	DEG	DG
TS	**0.021**±0.009	**0.16**±0.02	**0.21** ±0.02	**0.0096**±0.006	**0.11**±0.02	**0.18**±0.02
DEG	**0.26**±0.03	**0.3**±0.02	**0.26**±0.02	**0.073**±0.01	**0.29**±0.02	**0.25**±0.02
DG	**0.34**±0.03	**0.4**±0.03	**0.33**±0.02	**0.15**±0.02	**0.39**±0.03	**0.33**±0.02

Table 4. Data advantage vs. reputation advantage experiment, on Heavy-Tail MAB instance. Each cell describes the duopoly game between the entrant's algorithm (the **row**) and the incumbent's algorithm (the **column**). The cell specifies the entrant's market share for the rounds in which hit was present: the average (in bold) and the 95% confidence interval. NB: smaller average is better for the incumbent.

We can quantitatively define the data (resp., reputation) advantage as the incumbent's market share in the competition period in the data-advantage (resp., reputation advantage) experiment, minus the said share under simultaneous entry duopoly, for the same pair of algorithms and the same problem instance. In this language, our findings are as follows.

Finding 9.
(a) Data advantage and reputation advantage alone are substantially large, across all algorithms and

all MAB instances.
(b) The data advantage is larger than the reputation advantage when the incumbent chooses TS.
(c) The two advantages are similar in magnitude when the incumbent chooses DEG *or* DG.

Our intuition for Finding 9(b) is as follows. Suppose the incumbent switches from DG to TS. This switch allows the incumbent to explore actions more efficiently – collect better data in the same number of rounds – and therefore should benefit the data advantage. However, the same switch increases the reputation cost of exploration in the short run, which could weaken the reputation advantage.

11 PERFORMANCE IN ISOLATION, REVISITED

We saw in Section 9 that mean reputation trajectories do not suffice to explain the outcomes under competition. Let us provide more evidence and intuition for this.

Mean reputation trajectories are so natural that one is tempted to conjecture that they determine the outcomes under competition. More specifically:

Conjecture 11.1. If one algorithm's mean reputation trajectory lies above another, perhaps after some initial time interval (*e.g.,* as in Figure 3), then the first algorithm prevails under competition, for a sufficiently large warm start T_0.

However, we find a more nuanced picture. For example, in Figure 1 we see that DG attains a larger market share than DEG even for large warm starts. We find that this also holds for $K = 3$ arms and longer time horizons, see the supplement for more details. We conclude:

Finding 10. *Conjecture 11.1 is false: mean reputation trajectories do not suffice to explain the outcomes under competition.*

To see what could go wrong with Conjecture 11.1, consider how an algorithm's reputation score is distributed at a particular time. That is, consider the empirical distribution of this score over different mean reward vectors.[12] For concreteness, consider the Needle-in-Haystack instance at time $t = 500$, plotted in Figure 7. (The other MAB instances lead to a similar intuition.)

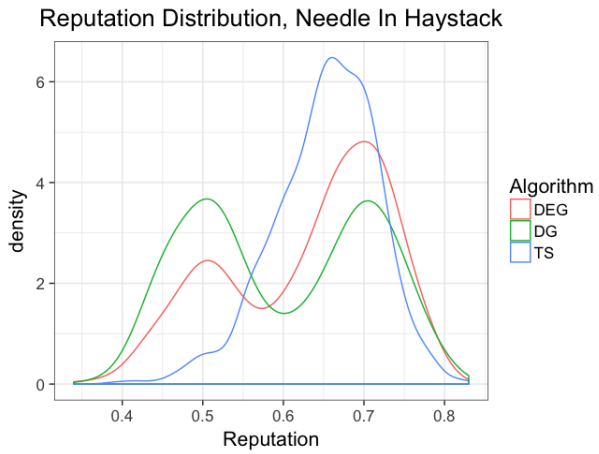


Fig. 7. Reputation scores for Needle-in-Haystack at $t = 500$ (smoothed using a kernel density estimate)

[12]Recall that each mean reward vector in our experimental setup comes with one specific realization table.

We see that the "naive" algorithms DG and DEG have a bi-modal reputation distribution, whereas TS does not. The reason is that for this MAB instance, DG either finds the best arm and sticks to it, or gets stuck on the bad arms. In the former case DG does slightly better than TS, and in the latter case it does substantially worse. However, the mean reputation trajectory may fail to capture this complexity since it simply takes average over different mean reward vectors. This may be inadequate for explaining the outcome of the duopoly game, given that the latter is determined by a simple comparison between the firm's reputation scores.

To further this intuition, consider the difference in reputation scores (*reputation difference*) between TS and DG on a particular mean reward vector. Let's plot the empirical distribution of the reputation difference (over the mean reward vectors) at a particular time point. Figure 8 shows such plots for several time points. We observe that the distribution is skewed to the right, precisely due to the fact that DG either does slightly better than TS or does substantially worse. Therefore, the mean is not a good measure of the central tendency, or typical value, of this distribution.

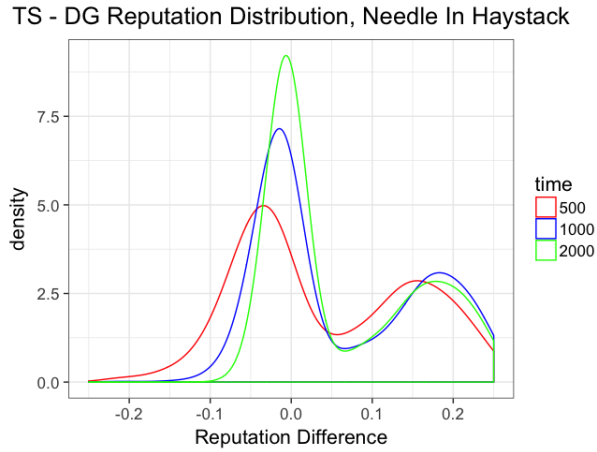


Fig. 8. Reputation difference TS − DG for Needle-in-Haystack (smoothed using a kernel density estimate)

12 NON-DETERMINISTIC CHOICE MODELS

Let us consider an extension in which the agents' choice is no longer deterministic. Recall that in our main model agents deterministically choose the firm with the higher reputation score; call this choice rule *HardMax* (HM). Now, we introduce some randomness: each agent selects between the firms uniformly with probability $\epsilon \in (0, 1)$, and takes the firm with the higher reputation score with the remaining probability; call this choice rule *HardMax with randomness* (HMR).

One can view HMR as a version of "warm start", where a firm receives some customers without competition, but these customers are dispersed throughout the game. The expected duration of this "dispersed warm start" is ϵT. If this quantity is large enough, we expect better algorithms to reach their long-term performance and prevail in competition. We confirm this intuition; we also find that this effect is negligible for smaller (but still relevant) values of ϵ or T.

Finding 11. TS *is weakly dominant under the* HMR *choice rule, if and only if* ϵT *is sufficiently large.* HMR *leads to lower variance in market shares, compared to* HM.

Table 5 shows the average market shares under the HM vs HMR choice rule. In contrast to the HM model, TS becomes weakly dominant under the HMR model, as T gets sufficiently large. These findings hold across all problem instances, see Table 6 (with the same semantics as in Table 5).

	Heavy-Tail (HMR with $\epsilon = .1$)			Heavy-Tail (HM)		
	TS vs DG	TS vs DEG	DG vs DEG	TS vs DG	TS vs DEG	DG vs DEG
$T = 2000$	**0.43** ± 0.02 Var: 0.15	**0.44** ± 0.02 Var: 0.15	**0.6** ± 0.02 Var: 0.1	**0.29** ± 0.03 Var: 0.2	**0.28** ± 0.03 Var: 0.19	**0.63** ± 0.03 Var: 0.18
$T = 5000$	**0.66** ± 0.01 Var: 0.056	**0.59** ± 0.02 Var: 0.092	**0.56** ± 0.02 Var: 0.098	**0.29** ± 0.03 Var: 0.2	**0.29** ± 0.03 Var: 0.2	**0.62** ± 0.03 Var: 0.19
$T = 10000$	**0.76** ± 0.01 Var: 0.026	**0.67** ± 0.02 Var: 0.067	**0.52** ± 0.02 Var: 0.11	**0.3** ± 0.03 Var: 0.21	**0.3** ± 0.03 Var: 0.2	**0.6** ± 0.03 Var: 0.2

Table 5. HM and HMR choice models on the Heavy-Tail MAB instance. Each cell describes the market shares in a game between two algorithms, call them Alg1 vs. Alg2, at a particular value of t. Line 1 in the cell is the market share of Alg 1: the average (in bold) and the 95% confidence band. Line 2 specifies the variance of the market shares across the simulations. The results reported here are with $T_0 = 20$.

	Uniform (HMR with $\epsilon = .1$)			Needle-In-Haystack (HMR with $\epsilon = .1$)		
	TS vs DG	TS vs DEG	DG vs DEG	TS vs DG	TS vs DEG	DG vs DEG
$T = 2000$	**0.42** ± 0.02 Var: 0.13	**0.45** ± 0.02 Var: 0.13	**0.49** ± 0.02 Var: 0.093	**0.55** ± 0.02 Var: 0.15	**0.61** ± 0.02 Var: 0.13	**0.46** ± 0.02 Var: 0.12
$T = 5000$	**0.48** ± 0.02 Var: 0.089	**0.53** ± 0.02 Var: 0.098	**0.46** ± 0.02 Var: 0.072	**0.56** ± 0.02 Var: 0.13	**0.63** ± 0.02 Var: 0.12	**0.43** ± 0.02 Var: 0.11
$T = 10000$	**0.54** ± 0.01 Var: 0.055	**0.6** ± 0.02 Var: 0.073	**0.44** ± 0.02 Var: 0.064	**0.58** ± 0.02 Var: 0.083	**0.65** ± 0.02 Var: 0.096	**0.4** ± 0.02 Var: 0.1

Table 6. HMR choice model for Uniform and Needle-In-Haystack MAB instances.

However, it takes a significant amount of randomness and a relatively large time horizon for this effect to take place. Even with $T = 10000$ and $\epsilon = 0.1$ we see that DEG still outperforms DG on the Heavy-Tail MAB instance as well as that TS only starts to become weakly dominant at $T = 10000$ for the Uniform MAB instance.

13 CONCLUSION

We considered a stylized duopoly setting where firms simultaneously learn from and compete for users. We analyzed two separate but connected variants of this model, one where consumers chose between firms based on their Bayesian expected reward and another where consumers chose between them based on their past performance. We showed in both cases that competition may not always induce firms to commit to better exploration algorithms, resulting in welfare losses for consumers. The primary mechanism in both was that consumers need to be incentivized to select a firm over its competitors, leading firms that engage in exploration to be starved of consumers before they make enough progress on their learning problem. We found that in order to incentivize "better" exploration strategies the key intuition is that a firm needs to have some "free" consumers that visit them without having to be incentivized to do so. We explored several economic mechanisms that can generate this. The first was by considering different variants of random utility models where stochastic choice gives firms enough free consumers to incentivize them to commit to "better" algorithms and we explored how this varied based on the nature of the random utility model. The second was by considering giving one firm a first-mover advantage where all consumers in the market would go to this firm without having to be incentivized to do so. While this first-mover advantage lead to the incumbent firm being incentivized to commit to a "better" algorithm it lead to a substantial market share going to this firm. We further analyzed to what extent this advantage

came from having a more defined reputation or more data from the incumbency period. We found that even a small amount of “data advantage" leads to substantial long-term market power due to the dynamics of competition.

REFERENCES

[1] Agarwal, A., Bird, S., Cozowicz, M., Dudik, M., Hoang, L., Langford, J., Li, L., Melamed, D., Oshri, G., Sen, S., and Slivkins, A. Multiworld testing: A system for experimentation, learning, and decision-making, 2016. A white paper, available at https://github.com/Microsoft/mwt-ds/raw/master/images/MWT-WhitePaper.pdf.

[2] Agarwal, A., Bird, S., Cozowicz, M., Hoang, L., Langford, J., Lee, S., Li, J., Melamed, D., Oshri, G., Ribas, O., Sen, S., and Slivkins, A. Making contextual decisions with low technical debt, 2017. Techical report at arxiv.org/abs/1606.03966.

[3] Aghion, P., Bloom, N., Blundell, R., Griffith, R., and Howitt, P. Competition and innovation: An inverted u relationship. *Quaterly J. of Economics 120*, 2 (2005), 701–728.

[4] Athey, S., and Schmutzler, A. Investment and market dominance. *RAND Journal of economics* (2001), 1–26.

[5] Athey, S., and Segal, I. An efficient dynamic mechanism. *Econometrica 81*, 6 (Nov. 2013), 2463–2485. A preliminary version has been available as a working paper since 2007.

[6] Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning 47*, 2-3 (2002), 235–256.

[7] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput. 32*, 1 (2002), 48–77. Preliminary version in *36th IEEE FOCS*, 1995.

[8] Azevedo, E., and Gottlieb, D. Perfect competition in markets with adverse selection. *Econometrica 85*, 1 (2017), 67–105.

[9] Babaioff, M., Kleinberg, R., and Slivkins, A. Truthful mechanisms with implicit payment computation. *J. ACM 62*, 2 (2015), 10. Subsumes the conference papers in *ACM EC 2010* and *ACM EC 2013*.

[10] Babaioff, M., Sharma, Y., and Slivkins, A. Characterizing truthful multi-armed bandit mechanisms. *SIAM J. on Computing 43*, 1 (2014), 194–230. Preliminary version in *10th ACM EC*, 2009.

[11] Bagwell, K. Informational product differentiation as a barrier to entry. *International Journal of Industrial Organization 8*, 2 (1990), 207–223.

[12] Bahar, G., Smorodinsky, R., and Tennenholtz, M. Economic recommendation systems. In *16th ACM EC* (2016).

[13] Bajari, P., Chernozhukov, V., Hortaçsu, A., and Suzuki, J. The impact of big data on firm performance: An empirical investigation. Tech. rep., National Bureau of Economic Research, 2018.

[14] Bergemann, D., and Välimäki, J. The dynamic pivot mechanism. *Econometrica 78*, 2 (2010), 771–789. Preliminary versions have been available since 2006.

[15] Bimpikis, K., Papanastasiou, Y., and Savva, N. Crowdsourcing exploration. *Management Science 64* (2018), 1477–1973.

[16] Bolton, P., and Harris, C. Strategic Experimentation. *Econometrica 67*, 2 (1999), 349–374.

[17] Bubeck, S., and Cesa-Bianchi, N. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning 5*, 1 (2012).

[18] Cabral, L. M., Riordan, M. H., et al. The learning curve, market dominance, and predatory pricing. *ECONOMETRICA-EVANSTON ILL- 62* (1994), 1115–1115.

[19] Che, Y.-K., and Hörner, J. Optimal design for social learning. *Quarterly Journal of Economics* (2018). Forthcoming. First published draft: 2013.

[20] Devanur, N., and Kakade, S. M. The price of truthfulness for pay-per-click auctions. In *10th ACM EC* (2009), pp. 99–106.

[21] Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. of Machine Learning Research (JMLR) 7* (2006), 1079–1105.

[22] Frazier, P., Kempe, D., Kleinberg, J. M., and Kleinberg, R. Incentivizing exploration. In *ACM EC* (2014), pp. 5–22.

[23] Ghosh, A., and Hummel, P. Learning and incentives in user-generated content: multi-armed bandits with endogenous arms. In *ITCS* (2013), pp. 233–246.

[24] Gittins, J., Glazebrook, K., and Weber, R. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, 2011.

[25] Gummadi, R., Johari, R., and Yu, J. Y. Mean field equilibria of multiarmed bandit games. In *13th ACM EC* (2012).

[26] Ho, C.-J., Slivkins, A., and Vaughan, J. W. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *J. of Artificial Intelligence Research 55* (2016), 317–359. Preliminary version appeared in *ACM EC 2014*.

[27] Hotelling, H. Stability in competition. *The Economic Journal 39*, 153 (1929), 41–57.

[28] Immorlica, N., Kalai, A. T., Lucier, B., Moitra, A., Postlewaite, A., and Tennenholtz, M. Dueling algorithms. In *43rd ACM STOC* (2011), pp. 215–224.

[29] Kakade, S. M., Lobel, I., and Nazerzadeh, H. Optimal dynamic mechanism design and the virtual-pivot mechanism. *Operations Research 61*, 4 (2013), 837–854.

[30] Keller, G., Rady, S., and Cripps, M. Strategic Experimentation with Exponential Bandits. *Econometrica 73*, 1 (2005), 39–68.

[31] Kleinberg, R. D., Waggoner, B., and Weyl, E. G. Descending price optimally coordinates search. Working paper,

2016. Preliminary version in *ACM EC 2016*.
[32] Kremer, I., Mansour, Y., and Perry, M. Implementing the "wisdom of the crowd". *J. of Political Economy 122* (2014), 988–1012. Preliminary version in *ACM EC 2014*.
[33] Lai, T. L., and Robbins, H. Asymptotically efficient Adaptive Allocation Rules. *Advances in Applied Mathematics 6* (1985), 4–22.
[34] Lambrecht, A., and Tucker, C. E. Can big data protect a firm from competition?
[35] Mansour, Y., Slivkins, A., and Syrgkanis, V. Bayesian incentive-compatible bandit exploration. In *15th ACM EC* (2015).
[36] Mansour, Y., Slivkins, A., Syrgkanis, V., and Wu, S. Bayesian exploration: Incentivizing exploration in bayesian games. Working paper, 2018. Available at `https://arxiv.org/abs/1602.07570`. Preliminary version in *ACM EC 2016*.
[37] Milgrom, P., and Stokey, N. Information, trade and common knowledge. *J. of Economic Theory 26*, 1 (1982), 17–27.
[38] Mitzenmacher, M., and Upfal, E. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.
[39] Perloff, J. M., and Salop, S. C. Equilibrium with product differentiation. *Review of Economic Studies LII* (1985), 107–120.
[40] Rothschild, M., and Stiglitz, J. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quaterly J. of Economics 90*, 4 (1976), 629–649.
[41] Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning 11*, 1 (2018), 1–96.
[42] Rysman, M. The economics of two-sided markets. *J. of Economic Perspectives 23*, 3 (2009), 125–144.
[43] Schmalensee, R. Product differentiation advantages of pioneering brands. *The American Economic Review 72*, 3 (1982), 349–365.
[44] Schumpeter, J. *Capitalism, Socialism and Democracy*. Harper & Brothers, 1942.
[45] Singla, A., and Krause, A. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *22nd WWW* (2013), pp. 1167–1178.
[46] Slivkins, A. Introduction to multi-armed bandits, 2018. A book draft, available at `http://research.microsoft.com/en-us/people/slivkins`. To be published with *Foundations and Trends in Machine Learning*.
[47] Veiga, A., and Weyl, G. Product design in selection markets. *Quarterly J. of Economics 131*, 2 (2016), 1007–1056.
[48] Vives, X. Innovation and competitive pressure. *J. of Industrial Economics 56*, 3 (2008).
[49] Weyl, G., and White, A. Let the right 'one' win: Policy lessons from the new economics of platforms. *Competition Policy International 12*, 2 (2014), 29–51.
[50] Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *J. Comput. Syst. Sci. 78*, 5 (2012), 1538–1556. Preliminary version in COLT 2009.
[51] Yue, Y., and Joachims, T. Interactively optimizing information retrieval systems as a dueling bandits problem. In *26th ICML* (2009), pp. 1201–1208.

Appendices

A BACKGROUND ON MULTI-ARMED BANDITS

This appendix provides some pertinent background on multi-armed bandits (*MAB*). We discuss `BIR` and monotonicity of several MAB algorithms, touching upon: `DynamicGreedy` and `StaticGreedy` (Section A.1), "naive" MAB algorithms that separate exploration and exploitation (Section A.2), and "smart" MAB algorithms that combine exploration and exploitation (Section A.3).

As we do throughout the paper, we focus on MAB with i.i.d. rewards and a Bayesian prior; we call it *Bayesian MAB* for brevity.

A.1 `DynamicGreedy` and `StaticGreedy`

We provide an example when `DynamicGreedy` and `StaticGreedy` have constant `BIR`, and prove monotonicity of `DynamicGreedy`. For the example, it suffices to consider *deterministic rewards* (for

each action a, the realized reward is always equal to the mean μ_a) and *independent priors* (according to the prior $\mathcal{P}_{\texttt{mean}}$, random variables $\mu_1, \ldots, \mu_K$ are mutually independent) each of *full support*.

The following claim is immediate from the definition of the CDF function

Claim A.1. *Assume independent priors. Let F_i be the CDF of the mean reward μ_i of action $a_i \in A$. Then, for any numbers $z_2 > z_1 > \mathbb{E}[\mu_2]$ we have $\Pr[\mu_1 \leq z_1 \text{ and } \mu_2 \geq z_2] = F_1(z_1)(1 - F_2(z_2))$.*

We can now draw an immediate corollary of the above claim

Corollary A.2. *Consider any problem instance of Bayesian MAB with two actions and independent priors which are full support. Then:*

(a) With constant probability, `StaticGreedy` *has a constant* `BIR` *for all steps.*

(b) Assuming deterministic rewards, with constant probability `DynamicGreedy` *has a constant* `BIR` *for all steps.*

Remark A.3. A similar result holds for rewards which are distributed as Bernoulli random variables. In this case we consider accumulative reward of an action as a random walk, and use a high probability variation of the law of iterated logarithms. (Details omitted.)

Next, we show that `DynamicGreedy` is monotone.

Lemma A.4. `DynamicGreedy` *is monotone, in the sense that* $\texttt{rew}(n)$ *is non-decreasing. Further,* $\texttt{rew}(n)$ *is strictly increasing for every time step n with* $\Pr[a_n \neq a_{n+1}] > 0$.

Proof. We prove by induction on n that $\texttt{rew}(n) \leq \texttt{rew}(n+1)$ for `DynamicGreedy`. Let a_n be the random variable recommended at time t, then $\mathbb{E}[\mu_{a_n}|\mathcal{I}_n] = \texttt{rew}(n)$. We can rewrite this as:

$$\texttt{rew}(n) = \mathop{\mathbb{E}}_{\mathcal{I}_n}[\mathop{\mathbb{E}}_{r_n}[\mu_{a_n}|r_n, \mathcal{I}_n]] = \mathop{\mathbb{E}}_{\mathcal{I}_{n+1}}[\mu_{a_n}|\mathcal{I}_{n+1}]$$

since $\mathcal{I}_{n+1} = (\mathcal{I}_n, r_n)$. At time $n+1$ `DynamicGreedy` will select an action a_{n+1} such that:

$$\texttt{rew}(n+1) = \mathbb{E}[\mu_{a_{n+1}}|\mathcal{I}_{n+1}] \geq \mathbb{E}[\mu_{a_n}|\mathcal{I}_n] = \texttt{rew}(n)$$

which proves the monotonicity. In cases that $\Pr[a_n \neq a_{n+1}] > 0]$ we have a strict inequality, since with some probability we select a better action then the realization of a_n. □

A.2 "Naive" MAB algorithms that separate exploration and exploitation

MAB algorithm `ExplorExploit` (m) initially explores each action with m agents and for the remaining $T - |A|m$ agents recommends the action with the highest observed average. In the explore phase it assigns a random permutation of the mK recommendations.

Lemma A.5. *The* `ExplorExploit` $(T^{2/3}\log|A|/\delta)$ *algorithm has, with probability $1-\delta$, for any $n \geq |A|T^{2/3}$ we have* `BIR` $(n) = O(T^{-1/3})$. *In addition,* `ExplorExploit` (m) *is monotone.*

Proof. In the explore phase we we approximate for each action $a \in A$, the value of μ_a by $\hat{\mu}_a$. Using the standard Chernoff bounds we have that with probability $1-\delta$, for every action $a \in A$ we have $|\mu_a - \hat{\mu}_a| \leq T^{-1/3}$.

Let $a^* = \arg\max_a \mu_a$ and a^{ee} the action that `ExplorExploit` selects in the explore phase after the first $|A|T^{2/3}$ agents. Since $\hat{\mu}_{a^*} \leq \hat{\mu}_{a^{ee}}$, this implies that $\mu_{a^*} - \mu_{a^{ee}} = O(T^{-1/3})$.

To show that `ExplorExploit` (m) is monotone, we need to show only that $\texttt{rew}(mK) \leq \texttt{rew}(mK+1)$. This follows since for any $t < mK$ we have $\texttt{rew}(t) = \texttt{rew}(t+1)$, since the recommended action is uniformly distributed for each time t. Also, for any $t \geq mK+1$ we have $\texttt{rew}(t) = \texttt{rew}(t+1)$ since we are recommending the same exploration action. The proof that $\texttt{rew}(mK) \leq \texttt{rew}(mK+1)$ is the same as for `DynamicGreedy` in Lemma A.4. □

We can also have a a phased version which we call PhasedExplorExploit (m_t), where time is partition in to phases. In phase t we have m_t agents and a random subset of K explore the actions (each action explored by a single agent) and the other agents exploit. (This implies that we need that $m_t \geq K$ for all t. We also assume that m_t is monotone in t.)

Lemma A.6. *Consider the case that $K = 2$ and the rewards of the actions are Bernoulli r.v. with parameter μ_i and $\Delta = \mu_1 - \mu_2$. Algorithm* PhasedExplorExploit (m_t) *is monotone and for $m_t = \sqrt{t}$ it has* $\mathtt{BIR}(n) = O(n^{-1/3} + e^{-O(\Delta^2 n^{2/3})})$.

Proof. We first show that it is monotone. Recall that $\mu_1 > \mu_2$. Let $S_i = \sum_{j=1}^{t} r_{i,j}$ be the sum of the rewards of action i up to phase t. We need to show that $\Pr[S_1 > S_2] + (1/2)\Pr[S_1 = S_2]$ is monotonically increasing in t. Consider the random variable $Z = S_1 - S_2$. At each phase it increases by $+1$ with probability $\mu_1(1-\mu_2)$, decreases by -1 with probability $(1-\mu_1)\mu_2$ and otherwise does not change.

Consider the values of Z up to phase t. We really care only about the probability that is shifted from positive to negative and vice versa.

First, consider the probability that $Z = 0$. We can partition it to $S_1 = S_2 = r$ events, and let $p(r, r)$ be the probability of this event. For each such event, we have $p(r, r)\mu_1$ moved to $Z = +1$ and $p(r, r)\mu_2$ moved to $Z = -1$. Since $\mu_1 > \mu_2$ we have that $p(r, r)\mu_1 \geq p(r, r)\mu_2$ (note that $p(r, r)$ might be zero, so we do not have a strict inequality).

Second, consider the probability that $Z = +1$ or $Z = -1$. We can partition it to $S_1 = r + 1; S_2 = r$ and $S_1 = r; S_2 = r + 1$ events, and let $p(r + 1, r)$ and $p(r, r + 1)$ be the probabilities of those events. It is not hard to see that $p(r + 1, r)\mu_2 = p(r, r + 1)\mu_1$. This implies that the probability mass moved from $Z = +1$ to $Z = 0$ is identical to that moved from $Z = -1$ to $Z = 0$.

We have showed that $\Pr[S_1 > S_2] + (1/2)\Pr[S_1 = S_2]$ and therefore the expected valued of the exploit action is non-decreasing. Since we have that the size of the phases are increasing, the BIR is strictly increasing between phases and identical within each phase.

We now analyze the BIR regret. Note that agent n is in phase $O(n^{2/3})$ and the length of his phase is $O(n^{1/3})$. The BIR has two parts. The first is due to the exploration, which is at most $O(n^{-1/3})$. The second is due to the probability that we exploit the wrong action. This happens with probability $\Pr[S_1 < S_2] + (1/2)\Pr[S_1 = S_2]$ which we can bound using a Chernoff bound by $e^{-O(\Delta^2 n^{2/3})}$, since we explored each action $O(n^{2/3})$ times. □

Remark A.7. Actually we have a tradeoff depending on the parameter m_t between the regret due to exploration and exploitation. (Note that the monotonicity is always guarantee assuming m_t is monotone.) If we can set that $m_t = 2^t$ then at time n we have $2/n$ probability of an exploit action. For the explore action we are in phase $\log n$ so the probability of a sub-optimal explore action is $n^{-O(\Delta^{-2})}$. This should give us $\mathtt{BIR}(n) = O(n^{-O(\Delta^{-2})})$.

A.3 "Smart" MAB algorithms that combine exploration and exploitation

MAB algorithm SuccesiveEliminationReset works as follows. It keeps a set of surviving actions $A_s \subseteq A$, where initially $A_s = A$. The agents are partition into phases, where each phase is a random permutation of the non-eliminated actions. Let $\hat{\mu}_{i,t}$ be the average of the rewards of action i up to phase t and $\hat{\mu}^* = \max_i \hat{\mu}_{i,t}$. We eliminate action i at the end of phase t, i.e., delete it from A_s, if $\hat{\mu}^*_t - \hat{\mu}_{i,t} > \log(T/\delta)/\sqrt{t}$. In SuccesiveEliminationReset we simply reset the algorithm with $A = A_s - A_{e,t}$, where $A_{e,t}$ is the set of eliminated actions after phase t. Namely, we restart $\hat{\mu}_{i,t}$ and ignore the old rewards before the elimination.

Lemma A.8. *The algorithm* SuccesiveEliminationReset, *has, with probability* $1 - \delta$, $\mathtt{BIR}(n) = O(\log(T/\delta)/\sqrt{n/K})$.

Proof. Let the best action be $a^* = \arg\max_a \mu_a$. With probability $1-\delta$ at any time n we have that for any action $i \in A_s$ that $|\hat{\mu}_i - \mu_i| \leq \log(T/\delta)/\sqrt{n/K}$, and $a^* \in A_s$. This implies that any action a such $\mu_{a^*} - \mu_a > 3\log(T/\delta)/\sqrt{n/K}$ is eliminated. Therefore, any action in A_s has BIR (n) of at most $6\log(T/\delta)/\sqrt{n/K}$. □

Lemma A.9. *Assume that if $\mu_i \geq \mu_j$ then the rewards r_i stochastically dominates the rewards r_j. Then,* SuccesiveEliminationReset *is monotone*

Proof. Consider the first time T an action is eliminated, and let $T = \tau$ be a realized value of T. Then, clearly for $n < \tau$ we have $\texttt{rew}(n) = \texttt{rew}(1)$.

Consider two actions $a_1, a_2 \in A$, such that $\mu_{a_1} \geq \mu_{a_2}$. At time $T = \tau$, the probability that a_1 is eliminated is smaller than the probability that a_2 is eliminated. This follows since $\hat{\mu}_{a_1}$ stochastically dominates $\hat{\mu}_{a_2}$, which implies that for any threshold θ we have $\Pr[\hat{\mu}_{a_1} \geq \theta] \geq \Pr[\hat{\mu}_{a_2} \geq \theta]$.

After the elimination we consider the expected reward of the eliminated action $\sum_{i \in A} \mu_i q_i$, where q_i is the probability that action i was eliminated in time $T = \tau$. We have that $q_i \leq q_{i+1}$, from the probabilities of elimination.

The sum $\sum_{i \in A} \mu_i q_i$ with $q_i \leq q_{i+1}$ and $\sum_i q_i = 1$ is maximized by setting $q_i = 1/|A|$. (We can see that if there are $q_i \neq 1/|A|$, then there are two $q_i < q_{i+1}$, and one can see that setting both to $(q_i + q_{i+1})/2$ increases the value.) Therefore we have that the $\texttt{rew}(\tau) \geq \texttt{rew}(\tau - 1)$.

Now we can continue by induction. For the induction, we can show the property for *any* remaining set of at most $k-1$ actions. The main issue is that SuccesiveEliminationReset restarts from scratch, so we can use induction. □

B NON-DEGENERACY VIA A RANDOM PERTURBATION

We show that Assumption (5) holds almost surely under a small random perturbation of the prior. We focus on problem instances with 0-1 rewards, and assume that the prior $\mathcal{P}_{\texttt{mean}}$ is independent across arms and has a finite support.[13] Consider the probability vector in the prior for arm a:

$$\vec{p}_a = \left(\Pr[\mu_a = \nu] : \ \nu \in \texttt{support}(\mu_a) \right).$$

We apply a small random perturbation independently to each such vector:

$$\vec{p}_a \leftarrow \vec{p}_a + \vec{q}_a, \quad \text{where} \quad \vec{q}_a \sim \mathcal{N}_a. \tag{31}$$

Here $\mathcal{N}_a$ is the noise distribution for arm a: a distribution over real-valued, zero-sum vectors of dimension $d_a = |\texttt{support}(\mu_a)|$. We need the noise distribution to satisfy the following property:

$$\forall x \in [-1,1]^{d_a} \setminus \{0\} \qquad \Pr_{q \sim \mathcal{N}_a} \left[x \cdot (\vec{p}_a + q) \neq 0 \right] = 1. \tag{32}$$

Theorem B.1. *Consider an instance of MAB with 0-1 rewards. Assume that the prior $\mathcal{P}_{\texttt{mean}}$ is independent across arms, and each mean reward μ_a has a finite support that does not include* 0 *or* 1. *Assume that noise distributions $\mathcal{N}_a$ satisfy property* (32). *If random perturbation* (31) *is applied independently to each arm a, then Eq. 5 holds almost surely for each history h.*

Remark B.2. As a generic example of a noise distribution which satisfies Property (32), consider the uniform distribution $\mathcal{N}$ over the bounded convex set

$$Q = \left\{ q \in \mathbb{R}^{d_a} \mid q \cdot \vec{1} = 0 \text{ and } \|q\|_2 \leq \epsilon \right\},$$

[13] The assumption of 0-1 rewards is for clarity. Our results hold under a more general assumption that for each arm a, rewards can only take finitely many values, and each of these values is possible (with positive probability) for every possible value of the mean reward μ_a.

where $\vec{1}$ denotes the all-1 vector. If $x = a\vec{1}$ for some non-zero value of a, then (32) holds because

$$x \cdot (p+q) = x \cdot p = a \neq 0.$$

Otherwise, denote $p = \vec{p}_a$ and observe that $x \cdot (p+q) = 0$ only if $x \cdot q = c \triangleq x \cdot (-p)$. Since $x \neq \vec{1}$, the intersection $Q \cap \{x \cdot q = c\}$ either is empty or has measure 0 in Q, which implies $\Pr_q[x \cdot (p+q) \neq 0] = 1$.

To prove Theorem B.1, it suffices to focus on two arms, and perturb one of them. Since realized rewards have finite support, there are only finitely many possible histories. Therefore, it suffices to focus on a fixed history h.

Lemma B.3. *Consider an instance of MAB with 0-1 rewards. Assume that the prior $\mathcal{P}_{\texttt{mean}}$ is independent across arms, and that* $\texttt{support}(\mu_1)$ *is finite and does not include* 0 *or* 1. *Fix history h. Suppose random perturbation* (31) *is applied to arm* 1, *with noise distribution $\mathcal{N}_1$ that satisfies* (32). *Then $\mathbb{E}[\mu_1 \mid h] \neq \mathbb{E}[\mu_2 \mid h]$ almost surely.*

Proof. Note that $\mathbb{E}[\mu_a \mid h]$ does not depend on the algorithm which produced this history. Therefore, for the sake of the analysis, we can assume w.l.o.g. that this history has been generated by a particular algorithm, as long as this algorithm can can produce this history with non-zero probability. Let us consider the algorithm that deterministically chooses same actions as h.

Let $S = \texttt{support}(\mu_1)$. Then:

$$\mathbb{E}[\mu_1 \mid h] = \sum_{v \in S} v \cdot \Pr[\mu_1 = v \mid h] = \sum_{v \in S} v \cdot \Pr[h \mid \mu_1 = v] \cdot \Pr[\mu_1 = v] \;/\; \Pr[h],$$
$$\Pr[h] = \sum_{v \in S} \Pr[h \mid \mu_1 = v] \cdot \Pr[\mu_1 = v].$$

Therefore, $\mathbb{E}[\mu_1 \mid h] = \mathbb{E}[\mu_2 \mid h]$ if and only if

$$\sum_{v \in S} (v - C) \cdot \Pr[h \mid \mu_1 = v] \cdot \Pr[\mu_1 = v] = 0, \quad \text{where} \quad C = \mathbb{E}[\mu_2 \mid h].$$

Since $\mathbb{E}[\mu_2 \mid h]$ and $\Pr[h \mid \mu_1 = v]$ do not depend on the probability vector $\vec{p}_1$, we conclude that

$$\mathbb{E}[\mu_1 \mid h] = \mathbb{E}[\mu_2 \mid h] \quad \Leftrightarrow \quad x \cdot \vec{p}_1 = 0,$$

where vector

$$x := (\ (v - C) \cdot \Pr[h \mid \mu_1 = v] :\ v \in S\ ) \in [-1, 1]^{d_1}$$

does not depend on $\vec{p}_1$.

Thus, it suffices to prove that $x \cdot \vec{p}_1 \neq 0$ almost surely under the perturbation. In a formula:

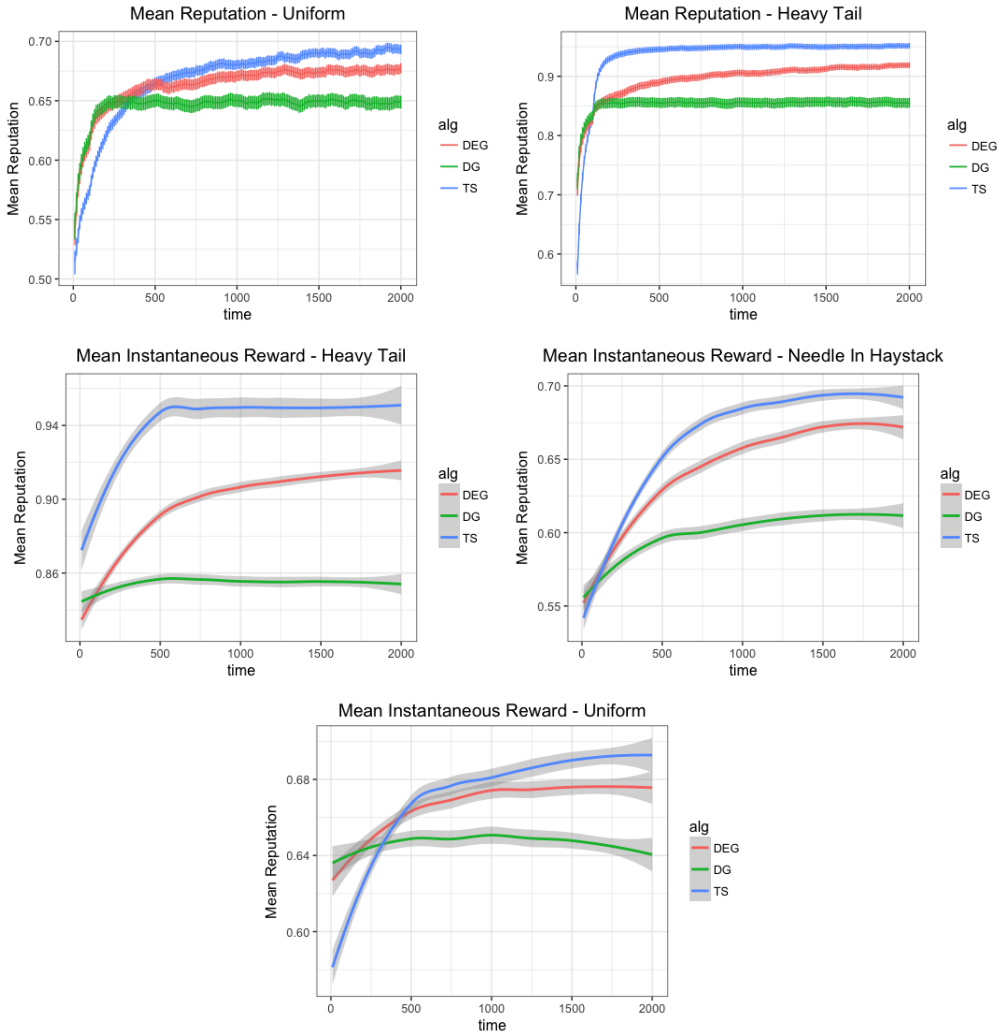
$$\Pr_{q \sim \mathcal{N}_1}\left[x \cdot (\vec{p}_1 + q) \neq 0\right] = 1 \tag{33}$$

Note that $\Pr[h \mid \mu_1 = v] > 0$ for all $v \in S$, because $0, 1 \notin S$. It follows that at most one coordinate of x can be zero. So (33) follows from property (32). □

We provide plots and tables for our experiments, which were omitted from the main text due to page constraints. In all cases, the plots and tables here are in line with those in the main text, and lead to similar qualitative conclusions.

C PLOTS FOR "PERFORMANCE IN ISOLATION"

We present additional plots for Section 8. First, we provide mean reputation trajectories for Uniform and Heavy-Tail MAB instances. Second, we provide trajectories for instantaneous mean rewards, for all three MAB instances.[14] In all plots, the shaded area represents 95% confidence interval.



D TEMPORARY MONOPOLY

We present additional experiments on temporary monopoly from Section 9, across various MAB instances and various values of the incumbent advantage parameter X.

Each experiment is presented as a table with the same semantics as in the main text. Namely, each cell in the table describes the duopoly game between the entrant's algorithm (the row) and the incumbent's algorithm (the column). The cell specifies the entrant's market share (fraction of

[14] These trajectories are smoothed via a non-parametric regression. More concretely, we use this option in `ggplot`: https://ggplot2.tidyverse.org/reference/geom_smooth.html.

rounds in which it was chosen) for the rounds in which he was present. We give the average (in bold) and the 95% confidence interval. NB: smaller average is better for the incumbent.

Heavy-Tail MAB Instance

	TS	DEG	DG
TS	**0.054** ±0.01	**0.16** ±0.02	**0.18** ±0.02
DEG	**0.33** ±0.03	**0.31** ±0.02	**0.26** ±0.02
DG	**0.39** ±0.03	**0.41** ±0.03	**0.33** ±0.02

Table 7. Temporary Monopoly: Heavy Tail, $X = 50$

	TS	DEG	DG
TS	**0.003** ±0.003	**0.083** ±0.02	**0.17** ±0.02
DEG	**0.045** ±0.01	**0.25** ±0.02	**0.23** ±0.02
DG	**0.12** ±0.02	**0.36** ±0.03	**0.3** ±0.02

Table 8. Temporary Monopoly: Heavy Tail, $X = 200$

	TS	DEG	DG
TS	**0.0017** ±0.002	**0.059** ±0.01	**0.16** ±0.02
DEG	**0.029** ±0.007	**0.23** ±0.02	**0.23** ±0.02
DG	**0.097** ±0.02	**0.34** ±0.03	**0.29** ±0.02

Table 9. Temporary Monopoly: Heavy Tail, $X = 300$

	TS	DEG	DG
TS	**0.002** ±0.003	**0.043** ±0.01	**0.16** ±0.02
DEG	**0.03** ±0.007	**0.21** ±0.02	**0.24** ±0.02
DG	**0.091** ±0.01	**0.32** ±0.03	**0.3** ±0.02

Table 10. Temporary Monopoly: Heavy Tail, $X = 500$

Needle-In-Haystack MAB Instance

	TS	DEG	DG
TS	**0.34** ±0.03	**0.4** ±0.03	**0.48** ±0.03
DEG	**0.22** ±0.02	**0.34** ±0.03	**0.42** ±0.03
DG	**0.18** ±0.02	**0.28** ±0.02	**0.37** ±0.03

Table 11. Temporary Monopoly: Needle-in-Haystack, $X = 50$

	TS	DEG	DG
TS	**0.17** ±0.02	**0.31** ±0.03	**0.41** ±0.03
DEG	**0.13** ±0.02	**0.26** ±0.02	**0.36** ±0.03
DG	**0.093** ±0.02	**0.23** ±0.02	**0.33** ±0.03

Table 12. Temporary Monopoly: Needle-in-Haystack, $X = 200$

	TS	DEG	DG
TS	**0.1** ±0.02	**0.28** ±0.03	**0.39** ±0.03
DEG	**0.089** ±0.02	**0.23** ±0.02	**0.36** ±0.03
DG	**0.05** ±0.01	**0.21** ±0.02	**0.33** ±0.03

Table 13. Temporary Monopoly: Needle-in-Haystack, $X = 300$

	TS	DEG	DG
TS	**0.053** ±0.01	**0.23** ±0.02	**0.37** ±0.03
DEG	**0.051** ±0.01	**0.2** ±0.02	**0.33** ±0.03
DG	**0.031** ±0.009	**0.18** ±0.02	**0.31** ±0.02

Table 14. Temporary Monopoly: Needle-in-Haystack, $X = 500$

Uniform MAB Instance

	TS	DEG	DG
TS	**0.27** ±0.03	**0.21** ±0.02	**0.26** ±0.02
DEG	**0.39** ±0.03	**0.3** ±0.03	**0.34** ±0.03
DG	**0.39** ±0.03	**0.31** ±0.02	**0.33** ±0.02

Table 15. Temporary Monopoly: Uniform, $X = 50$

	TS	DEG	DG
TS	**0.12** ±0.02	**0.16** ±0.02	**0.2** ±0.02
DEG	**0.25** ±0.02	**0.24** ±0.02	**0.29** ±0.02
DG	**0.23** ±0.02	**0.24** ±0.02	**0.29** ±0.02

Table 16. Temporary Monopoly: Uniform, $X = 200$

	TS	DEG	DG
TS	**0.094** ±0.02	**0.15** ±0.02	**0.2** ±0.02
DEG	**0.2** ±0.02	**0.23** ±0.02	**0.29** ±0.02
DG	**0.21** ±0.02	**0.23** ±0.02	**0.29** ±0.02

Table 17. Temporary Monopoly: Uniform, $X = 300$

	TS	DEG	DG
TS	**0.061** ±0.01	**0.12** ±0.02	**0.2** ±0.02
DEG	**0.17** ±0.02	**0.21** ±0.02	**0.29** ±0.02
DG	**0.18** ±0.02	**0.22** ±0.02	**0.29** ±0.02

Table 18. Temporary Monopoly: Uniform, $X = 500$

E REPUTATION VS. DATA ADVANTAGE

This section presents all experiments on data vs. reputation advantage (Section 10).

Each experiment is presented as a table with the same semantics as in the main text. Namely, each cell in the table describes the duopoly game between the entrant's algorithm (the **row**) and the incumbent's algorithm (the **column**). The cell specifies the entrant's market share for the rounds in which hit was present: the average (in bold) and the 95% confidence interval. NB: smaller average is better for the incumbent.

	TS	DEG	DG
TS	**0.0096** ± 0.006	**0.11** ± 0.02	**0.18** ± 0.02
DEG	**0.073** ± 0.01	**0.29** ± 0.02	**0.25** ± 0.02
DG	**0.15** ± 0.02	**0.39** ± 0.03	**0.33** ± 0.02

Table 19. Data Advantage: Heavy Tail, $X = 200$

	TS	DEG	DG
TS	**0.021** ± 0.009	**0.16** ± 0.02	**0.21** ± 0.02
DEG	**0.26** ± 0.03	**0.3** ± 0.02	**0.26** ± 0.02
DG	**0.34** ± 0.03	**0.4** ± 0.03	**0.33** ± 0.02

Table 20. Reputation Advantage: Heavy Tail, $X = 200$

	TS	DEG	DG
TS	**0.25** ± 0.03	**0.36** ± 0.03	**0.45** ± 0.03
DEG	**0.21** ± 0.02	**0.32** ± 0.03	**0.41** ± 0.03
DG	**0.18** ± 0.02	**0.29** ± 0.03	**0.4** ± 0.03

Table 21. Data Advantage: Needle-in-Haystack, $X = 200$

	TS	DEG	DG
TS	**0.35** ± 0.03	**0.43** ± 0.03	**0.52** ± 0.03
DEG	**0.26** ± 0.03	**0.36** ± 0.03	**0.43** ± 0.03
DG	**0.19** ± 0.02	**0.3** ± 0.02	**0.36** ± 0.02

Table 22. Reputation Advantage: Needle-in-Haystack, $X = 200$

	TS	DEG	DG
TS	**0.27** ± 0.03	**0.23** ± 0.02	**0.27** ± 0.02
DEG	**0.4** ± 0.03	**0.3** ± 0.02	**0.32** ± 0.02
DG	**0.36** ± 0.03	**0.29** ± 0.02	**0.3** ± 0.02

Table 23. Reputation Advantage: Uniform, $X = 200$

	TS	DEG	DG
TS	**0.2** ± 0.02	**0.22** ± 0.02	**0.27** ± 0.03
DEG	**0.33** ± 0.03	**0.32** ± 0.03	**0.35** ± 0.03
DG	**0.32** ± 0.03	**0.31** ± 0.03	**0.35** ± 0.03

Table 24. Data Advantage: Uniform, $X = 200$

	TS	DEG	DG
TS	**0.0017** ±0.002	**0.06** ±0.01	**0.18** ±0.02
DEG	**0.04** ±0.009	**0.24** ±0.02	**0.25** ±0.02
DG	**0.12** ±0.02	**0.35** ±0.03	**0.33** ±0.02

Table 25. Data Advantage: Heavy-Tail, $X = 500$

	TS	DEG	DG
TS	**0.022** ±0.009	**0.13** ±0.02	**0.21** ±0.02
DEG	**0.26** ±0.03	**0.29** ±0.02	**0.28** ±0.02
DG	**0.33** ±0.03	**0.39** ±0.03	**0.34** ±0.02

Table 26. Reputation Advantage: Heavy-Tail, $X = 500$

	TS	DEG	DG
TS	**0.098** ±0.02	**0.27** ±0.03	**0.41** ±0.03
DEG	**0.093** ±0.02	**0.24** ±0.02	**0.38** ±0.03
DG	**0.064** ±0.01	**0.22** ±0.02	**0.37** ±0.03

Table 27. Data Advantage: Needle-in-Haystack, $X = 500$

	TS	DEG	DG
TS	**0.29** ±0.03	**0.44** ±0.03	**0.52** ±0.03
DEG	**0.19** ±0.02	**0.35** ±0.03	**0.42** ±0.03
DG	**0.15** ±0.02	**0.27** ±0.02	**0.35** ±0.02

Table 28. Reputation Advantage: Needle-in-Haystack, $X = 500$

	TS	DEG	DG
TS	**0.14** ±0.02	**0.18** ±0.02	**0.26** ±0.03
DEG	**0.26** ±0.02	**0.26** ±0.02	**0.34** ±0.03
DG	**0.25** ±0.02	**0.27** ±0.02	**0.34** ±0.03

Table 29. Data Advantage: Uniform, $X = 500$

	TS	DEG	DG
TS	**0.24** ±0.02	**0.2** ±0.02	**0.26** ±0.02
DEG	**0.37** ±0.03	**0.29** ±0.02	**0.31** ±0.02
DG	**0.35** ±0.03	**0.27** ±0.02	**0.3** ±0.02

Table 30. Reputation Advantage: Uniform, $X = 500$

F MEAN REPUTATION VS. RELATIVE REPUTATION

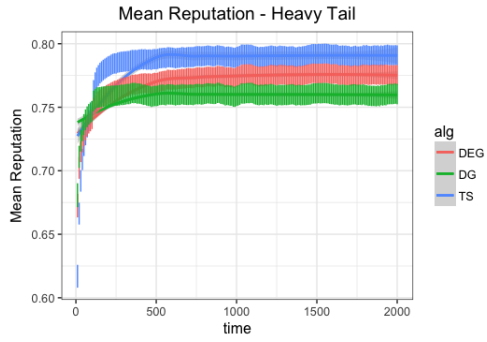
We present the experiments omitted from Section 11. Namely, experiments on the Heavy-Tail MAB instance with $K = 3$ arms, both for "performance in isolation" and the permanent duopoly game. We find that DEG > DG according to the mean reputation trajectory but that DG > DEG according to the relative reputation trajectory *and* in the competition game. As discussed in Section 11, the same results also hold for $K = 10$ for the warm starts that we consider.

The result of the permanent duopoly experiment for this instance is shown in Table 31.

	Heavy-Tail		
	$T_0 = 20$	$T_0 = 250$	$T_0 = 500$
TS vs. DG	**0.4** ±0.02 EoG 770 (0)	**0.59** ±0.01 EoG 2700 (2979.5)	**0.6** ±0.01 EoG 2700 (3018)
TS vs. DEG	**0.46** ±0.02 EoG 830 (0)	**0.73** ±0.01 EoG 2500 (2576.5)	**0.72** ±0.01 EoG 2700 (2862)
DG vs. DEG	**0.61** ±0.01 EoG 1400 (556)	**0.61** ±0.01 EoG 2400 (2538.5)	**0.6** ±0.01 EoG 2400 (2587.5)

Table 31. Duopoly Experiment: Heavy-Tail, $K = 3$, $T = 5000$.
Each cell describes a game between two algorithms, call them Alg1 vs. Alg2, for a particular value of the warm start T_0. Line 1 in the cell is the market share of Alg 1: the average (in bold) and the 95% confidence band. Line 2 specifies the "effective end of game" (EoG): the average and the median (in brackets).

The mean reputation trajectories for algorithms' performance in isolation:



Finally, the relative reputation trajectory of DEG vs. DG: