Competing Bandits: The Perils of Exploration under Competition*

Guy Aridor† Yishay Mansour‡ Aleksandrs Slivkins§ Zhiwei Steven Wu¶

August 2020

Learning from interactions with users is ubiquitous in modern customer-facing platforms, from product recommendations to web search to content selection to fine-tuning user interfaces. Many platforms purposefully implement *exploration*: making potentially suboptimal choices for the sake of acquiring new information. Online platforms routinely deploy A/B tests, and are increasingly adopting more sophisticated exploration methodologies based on *multi-armed bandits*, a standard and well-studied framework for exploration and making decisions under uncertainty (Gittins et al., 2011; Bubeck and Cesa-Bianchi, 2012; Slivkins, 2019; Lattimore and Szepesvári, 2020).

In this paper, we initiate a study of the interplay between *exploration* and *competition*. Platforms that engage in exploration typically need to compete against one another; most importantly, they compete for users. This creates a tension: while exploration may be essential for improving the service tomorrow, it may degrade quality and make users leave *today*, in which case there will be less users to learn from. This may further degrade the platform's performance relative to competitors who keep learning and improving from *their* users, and so forth. Taken to the extreme, such dynamics may create a "death spiral" effect when the vast majority of customers eventually switch to competitors. Users 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 systems.

The main high-level question is: **How does competition incentivize 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 under competition? Does increased competition lead to higher consumer welfare? How significant are "data feedback loops" — when more data leads to more users, which leads to even more data, etc. — and how they relate to the anti-trust considerations? To answer these questions, we study complex interactions between platforms' learning dynamics and users' self-interested behavior. The choice of a particular technology (exploration algorithm) is no longer an abstract, static choice with a predetermined outcome for the platform. Instead, we model the algorithms explicitly, and investigate how they play out in competition over an extended period of time. We offer a mix of theoretical results and numerical simulations, in a range of models.

Model: competition game. We consider 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.

(i) 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.

*This is an extended abstract of Aridor et al. (2020), which in turn subsumes conference papers Mansour et al. (2018) and Aridor et al. (2019).

†Columbia University, Department of Economics. Email: g.aridor@columbia.edu

‡Google and Tel Aviv University, Department of Computer Science. Email: mansour.yishay@gmail.com

§Microsoft Research New York City. Email: slivkins@microsoft.com

¶Carnegie-Mellon University, Pittsburgh, PA. Email: zstevenwu@cmu.edu. Partially done as an intern and a postdoc at MSR-NYC.

(ii) Each principal faces a 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. The reward distributions are initially unknown.

(iii) Principals simultaneously announce their bandit algorithms before round 1, and cannot change them afterwards. Each principal's objective is to maximize its market share (the fraction of users choosing this principal). Each principal only observes agents that chose this principal.

We investigate several model variants, as we vary agents' decision rule and/or posit a first-mover advantage.

Technology: multi-armed bandit algorithms. To compare between bandit algorithms, we build on prevalent intuition in the literature. We focus on standard notions of regret, and distinguish between three classes of bandit algorithms: ones that never explicitly explore (*greedy algorithms*), ones that explore without looking at the data (*exploration-separating algorithms*), and ones where exploration gradually zooms in on the best arm (*adaptive-exploration algorithms*). In isolation, *i.e.,* in the absence of competition, these classes are fairly well-understood. Greedy algorithms are terrible for a wide variety of problem instances, precisely because they never explore. Exploration-separated algorithms learn at a reasonable but mediocre rate across all problem instances. Adaptive-exploration algorithms are optimal in the worst case, and exponentially improve for "easy" problem instances. Generally, "better" algorithms are better in the long run, but could be worse initially.

Theoretical results. We consider a basic Bayesian model, where agents have a common Bayesian prior on reward distributions, and know the principals' algorithms. For tractability, we posit that agents do not receive any information about the previous agents' choices and rewards. Each agent knows the round (s)he arrives in, computes the Bayesian-expected reward for each principal, and use these two numbers to decide which principal to choose.

Our results depend crucially on the agents' decision rule:

(i) The most obvious decision rule (`HardMax`) maximizes the current agent's Bayesian-expected reward. We find that `HardMax` is not conducive to adopting better algorithms: each principal's dominant strategy is to choose the greedy algorithm. Further, we show that `HardMax` is sensitive to tie-breaking: if the tie-breaking is probabilistically biased in favor of one principal, then this principal has a simple "winning strategy" no matter what.

(ii) We dilute the `HardMax` agents with a small fraction of "random agents" who choose a principal uniformly at random. We call this model `HardMax&Random`. We find that better algorithms help in a big way: a sufficiently better algorithm is guaranteed to win all non-random agents after an initial learning phase. 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 in competition, and a pure Nash equilibrium typically does not exist.

(iii) We further relax the decision rule so that the probability of choosing a given principal varies smoothly as a function of the difference between principals' Bayesian-expected rewards; we call it `SoftMax`. Then, 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 would attract slightly more).

Economic interpretation: the inverted-U relationship. Interpreting the adoption of better algorithms as "innovation", our findings can be framed in terms of the *inverted-U relationship* between competition and innovation.[1]

Our decision rules differ in terms of rationality: from fully rational decisions with `HardMax` to relaxed rationality with `HardMax&Random` to an even more relaxed rationality with `SoftMax`. The same distinctions also control the severity of competition between the principals: from cut-throat competition with `HardMax` to a more relaxed competition with `HardMax&Random`, to an even more relaxed competition 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. While agents' rationality and severity of competition are often modeled separately in the literature, it is not unusual to have them modeled with the same "knob" (*e.g.,* Gabaix et al., 2016).

We identify the inverted-U relationship driven by the rationality/competitiveness distinctions outlined above: from `HardMax` to `HardMax&Random` to `SoftMax` to `Uniform`. We also find another, technically different, inverted-U

[1]This is a well-established concept in the economics literature, dating back to Schumpeter (1942), whereby too little or too much competition is bad for innovation, but intermediate levels of competition tend to be better (*e.g.,* Aghion et al., 2005; Vives, 2008).

relationship which zeroes in on the HardMax&Random model. We vary rationality/competitiveness inside this model, and track the marginal utility of switching to a better algorithm.

These inverted-U relationships are driven by different aspects in our model than the ones in the existing literature in economics. The latter focuses on the tradeoff between the R&D costs and the benefits that the improved technology provides in the competition. In our case, the barriers for innovations arise entirely from the reputational consequences of exploration in competition, even in the absence of R&D costs.

Numerical simulations. We consider a basic frequentist model. We posit that the agents observe signals about the principals' past performance, and base their decisions on these signals alone, without invoking any prior knowledge or beliefs. The performance signals are abstracted and aggregated as a scalar *reputation score* for each principal, modeled as a sliding window average of its rewards. Thus, agents' decision rule depends only on the two reputation scores. We refine and expand the theoretical results in several ways.

(i) We compare HardMax and HardMax&Random decision rules. We find that the greedy algorithm often wins under HardMax, with a strong evidence of the "death spiral" effect mentioned earlier. As predicted by the theory, better algorithms prevail if the expected number of "random" users is sufficiently large. However, this effect is negligible for smaller parameter values.

(ii) We investigate the first-mover advantage as a different channel to vary the intensity of competition: from the first-mover to simultaneous arrival to late-arriver. (We focus on the HardMax decision rule.) We find that the first-mover is incentivized to choose a more advanced exploration algorithm, whereas the late-arriver is often incentivized to choose the "greedy algorithm" (more so than under simultaneous arrival). Consumer welfare is higher under early/late arrival than under simultaneous entry. We frame these results in terms of an inverted-U relationship.

(iii) We decompose the first-mover advantage into two distinct effects: free data to learn from (*data advantage*), and a more definite, and possibly better reputation compared to an entrant (*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. Finally, we find an "amplification effect" of the data advantage: even a small amount thereof gets amplified under competition, causing a large difference in eventual market shares.

Economic interpretation: network effects of data. Our model speaks to policy discussions on regulating data-intensive digital platforms (Furman et al., 2019; Scott Morton et al., 2019), and particularly to the ongoing debate on the role of data in the digital economy. One fundamental question in this debate is whether data can serve a similar role as traditional "network effects", creating scenarios when only one firm can function in the market (Rysman, 2009; Jullien and Sand-Zantman, 2019). The death spiral/amplification effects mentioned above have a similar flavor as network effects: a relatively small amount of exploration (resp., data advantage) gets amplified under competition and causes the firm to be starved of users (resp., take over most of the market). A distinctive feature of our approach is that we explicitly model the learning problem of the firms and consider them deploying algorithms for solving this problem. Thus, we do not explicitly model the network effects, but they arise endogenously from our setup.

Our results highlight that understanding the performance of learning algorithms in isolation does not necessarily translate to understanding their impact in competition, precisely because competition leads to the endogenous generation of data observed by the firms. Approaches such as Lambrecht and Tucker (2015); Bajari et al. (2018); Varian (2018) argue that the diminishing returns to scale and scope of data in isolation mitigate such data feedback loops, but ignore the differences induced by learning in isolation versus under competition. Furthermore, explicitly incorporating the interaction between learning technology and data creation allows us to speak on how data advantages are characterized and amplified not only by data quantity, but also the increased data quality gathered by better learning algorithms.

Significance. Our theory takes a Bayesian approach and discovers several strong asymptotic results. The numerical simulations provide a more nuanced and "non-asymptotic" perspective. In essence, we look for substantial effects within relevant time scales. In fact, we start our investigation by determining what time scales are relevant in the context of our model. Our study has a dual purpose: (i) shed light on real-world implications of some typical scenarios, and (ii) investigate the space of models for describing the real world. As an example to clarify the latter point, consider the HardMax model with simultaneous entry. It is is not necessarily the most realistic model. However, our results elucidate the need for more refined models that allow for "free exploration" (*e.g.,* via random agents or early entry).

References

Philippe Aghion, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt. Competition and innovation: An inverted-u relationship. *The Quarterly Journal of Economics*, 120(2):701–728, 2005.

Guy Aridor, Aleksandrs Slivkins, and Steven Wu. The perils of exploration under competition: A computational modeling approach. In *20th ACM Conf. on Economics and Computation (ACM-EC)*, 2019.

Guy Aridor, Yishay Mansour, Aleksandrs Slivkins, and Steven Wu. Competing bandits: The perils of exploration under competition., 2020. Working paper. Subsumes conference papers in *ITCS 2018* and *ACM EC 2019*.

Patrick Bajari, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki. The impact of big data on firm performance: An empirical investigation. Technical report, National Bureau of Economic Research, 2018.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. Published with *Now Publishers* (Boston, MA, USA). Also available at `https://arxiv.org/abs/1204.5721`.

Jason Furman, Diane Coyle, Amelia Fletcher, Derek McAuley, and Philip Marsden. Unlocking digital competition. *Report of the digital competition expert panel*, 2019.

Xavier Gabaix, David Laibson, Deyuan Li, Hongyi Li, Sidney Resnick, and Casper G. de Vries. The impact of competition on prices with numerous firms. *J. of Economic Theory*, 165:1–24, 2016.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, Hoboken, NJ, USA, 2nd edition, 2011. The first edition, single-authored by John Gittins, has been published in 1989.

Bruno Jullien and Wilfried Sand-Zantman. The economics of platforms: A theory guide for competition policy. *TSE Digital Center Policy Papers series*, (1), 2019.

Anja Lambrecht and Catherine E Tucker. Can big data protect a firm from competition? 2015.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020. Versions available at `https://banditalgs.com/` since 2018.

Yishay Mansour, Aleksandrs Slivkins, and Steven Wu. Competing bandits: Learning under competition. In *9th Innovations in Theoretical Computer Science Conf. (ITCS)*, 2018.

Marc Rysman. The economics of two-sided markets. *J. of Economic Perspectives*, 23(3):125–144, 2009.

Joseph Schumpeter. *Capitalism, Socialism and Democracy*. Harper & Brothers, 1942.

Fiona Scott Morton, Pascal Bouvier, Ariel Ezrachi, Bruno Jullien, Roberta Katz, Gene Kimmelman, A Douglas Melamed, and Jamie Morgenstern. Committee for the study of digital platforms: Market structure and antitrust subcommittee report. *Chicago: Stigler Center for the Study of the Economy and the State, University of Chicago Booth School of Business*, 2019.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, November 2019. Published with *Now Publishers* (Boston, MA, USA). Also available at `https://arxiv.org/abs/1904.07272`.

Hal Varian. Artificial intelligence, economics, and industrial organization. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, 2018.

Xavier Vives. Innovation and competitive pressure. *J. of Industrial Economics*, 56(3), 2008.