Efficient Multi-Agent Experimentation and Multi-Choice Bandits

Alejandro Francetich*

September 29, 2018

Abstract

The first best in the multi-agent experimentation problem of Klein and Rady (2011) can be reinterpreted as a multi-choice bandit problem where a researcher experiments on up to two projects *simultaneously*. We exploit this analogy to characterize the optimal strategy: If the cost or the discount rate are low enough, after working on a single project unsuccessfully, the researcher takes on *both projects at once*. When the "bad project" can also yield successes, as in Keller and Rady (2010), we provide an example where the researcher delays taking on the second project, and we show that she can give up earlier in particular *if false positives are arbitrarily unlikely*.

Keywords: Experimentation, two-armed bandits, multi-choice bandits, negatively correlated arms, Poisson process

JEL Classification Numbers: D83, D90

*School of Business, University of Washington, Bothell. Email address: aletich@uw.edu. This paper was started while I was at the Decision Sciences Department at Bocconi University as a postdoctoral fellow. I am deeply indebted to David Kreps for continuously enlightening me, and to Alejandro Manelli, Pierpaolo Battigali, and Massimo Marinacci for their support and guidance. I am grateful to Sven Rady for his encouragement and insightful comments. The paper has benefited from feedback from Camelia Bejan, Juan Camilo Gomez, Amanda Friedenberg, Marco Ottaviani, attendants of the 2014 Canadian Economic Theory Conference and of the 25th International Game Theory Conference at Stony Brook, and seminar participants at UC Davis, UW Bothell, University of Sheffield, and Universidad Carlos III de Madrid, as well as anonymous referees. I gratefully acknowledge financial support from ERC advanced grant 324219. Any remaining errors and omissions are all mine.

1. Introduction

Imagine that a ship carrying treasure sinks near two islands; the treasure is known to be buried in one of the islands, but not in which one. An explorer can organize an expedition to either island, or parallel expeditions to both islands. It is more costly to explore both islands and there is only one treasure to be found, but simultaneous expeditions can cover more ground.

This problem shares similarities to academic and industrial research, as well as to recruiting. A researcher has a theoretical conjecture; she can work on a proof or a counterexample, or she can hire a research assistant to work on the counterexample *while* she works on the proof. A medical lab is conducting research on two different treatments for a disease; the staff can experiment on either treatment, or the director can hire additional researchers and have separate teams working on each treatment. A manager faces two job applicants with different profiles, and there is uncertainty about which one is a better match to the company; the manager can hire either applicant, or both for a probation period.

This novel variation of the multi-armed bandit problem is analogous to the social-planner's problem in multi-agent experimentation. The projects represent the agents, each of which has their own arm, and simultaneous experimentation is equivalent to multiple agents experimenting at once. In fact, in continuous time with Poisson discoveries and with constant marginal cost, the treasure-hunt problem is *isomorphic* to the social-planner's problem in Klein and Rady (2011, henceforth, KR). We exploit this isomorphism to characterize the optimal multi-choice research strategy.

Bergemann and Välimäki (2001) studies a multi-choice bandit with countably many ex-ante-identical arms where choosing additional arms is costless up to a fixed number; they show that their solution fails with only finitely many arms. Francetich and Kreps (2018a,b) allow for more than two arms and a more general correlation structure, thereby precluding tractable solutions. The papers investigate the performance of several decision heuristics.

2. Multi-Agent Experimentation as a Multi-Choice Bandit

Formally, a researcher or decision maker (henceforth, DM) can experiment on up to two projects, labelled 0 and 1. There is a cost $c > 0$ to undertaking each project. One and only one of the projects can produce successes; successes arrive for the "good project" according to a Poisson processes with arrival rate $\overline{\lambda} > c$. Rewards are normalized to 1. Payoffs are discounted at rate $\rho > 0$, and $\pi \in [0, 1]$ denotes the belief that project 0 is the fruitful one.

KR identify two parameter configurations, called *low* and (intermediate plus) *high stakes*. The case of low stakes is the case of *costly research*: $\rho(2c - \overline{\lambda}) > \overline{\lambda}(\overline{\lambda} - c)$; simultaneous research is not myopically profitable ($2c > \overline{\lambda}$), and the DM is too impatient to value information ($\rho > \overline{\lambda}(\overline{\lambda} - c)/(2c - \overline{\lambda})$). The case of high stakes is the case of

beneficial research: $\rho(2c-\overline{\lambda}) < \overline{\lambda}(\overline{\lambda}-c)$; either research is cheap ($c \leq \overline{\lambda}/2$) or the DM is sufficiently patient ($\rho < \overline{\lambda}(\overline{\lambda}-c)/(2c-\overline{\lambda})$).

The next result presents what Propositions 1 and 2 in KR teach us about our multi-choice bandit problem.

Proposition 1 (Optimal strategy)**.** *Define the cutoff beliefs:*

$$\overline{\pi}^1 := \frac{c\rho}{\overline{\lambda}(\overline{\lambda}+\rho-c)} \quad \textit{and} \quad \overline{\pi}^2 := \frac{\overline{\lambda}(\overline{\lambda}+\rho)-c\rho}{\overline{\lambda}(\overline{\lambda}+\rho+c)}.$$

The optimal strategy is as follows. If research is costly, work on project 0 if $\pi > \overline{\pi}^1$, on 1 if $\pi < 1-\overline{\pi}^1$, and otherwise on neither. Under beneficial research, work on 0 if $\pi > \overline{\pi}^2$, on 1 if $\pi < 1-\overline{\pi}^2$, and otherwise on both at once.

If the DM is sufficiently confident about a project, she begins working on it alone. While no successes occur, she becomes progressively pessimistic about this project and optimistic about the neglected one. Under costly research, however, her posterior does not move far enough to switch to the other project, and she eventually gives up. Under beneficial research, she takes on the second project without setting the "failing" one aside. Experimenting on both projects, our DM can observe the earliest next success, and she is not discouraged by lack of successes. The rationale for simultaneous research in KR is learning from my opponent: If we are both failing, I am not discouraged by my own failing; and if my opponent succeeds, I can switch to my safe arm faster.

Our DM experiments "more" when she is sufficiently uncertain. In Moscarini and Smith (2001), instead, experimentation intensifies as she becomes more confident. The difference is due to posteriors changing gradually over time, so experimentation is more costly when it takes longer for the posterior to reach decision thresholds.[1]

3. Ambiguous Successes

What if some of the treasure spills on the other island during the shipwreck? Now, finding gold does not guarantee that we are on the right island. In research, we may be able to prove partial lemmas even if the full proposition is false; a medical treatment may alleviate symptoms even if it proves ineffective in curing the disease; and a mismatched employee can still make valuable contributions to the company. This case is related to Keller and Rady (2010, henceforth, KR2), where the players have identical copies of a Poisson risky arm, which can either be good or bad; our researcher, instead, has one of each and her problem is figuring out which one is which.

We now have two opposing forces affecting experimentation. On the one hand, the opportunity cost of experimenting is lower; spending time on the wrong project is less costly. On the other hand, the informativeness of successes is also lower due to the

[1] I thank an anonymous referee for indicating that, with constant marginal costs and in discrete time, the experimentation pattern in Moscarini and Smith (2001) reverses.

presence of "false positives," and certainty is never reached. Thus, the productivity of the bad project can *hurt* the researcher and lead her to experiment *less*.

Let $\underline{\lambda} > 0$ be the arrival rate of the bad project. We assume that $\overline{\lambda} - c > 0 > \underline{\lambda} - c$, so that only the good project is worthwhile ex post. While a full characterization of the optimal strategy is beyond the scope of this article, we provide some comparative statics analysis and examples.

When the optimal strategy recommends experimenting on at most a single project, the problem is almost identical to Proposition 1 in KR2. In this case, the corresponding cutoff is:

$$\overline{\pi}^3 := \frac{(c-\underline{\lambda})\mu}{(\overline{\lambda}-c)(\mu+1)+(c-\underline{\lambda})\mu},$$

where μ is the positive root of the function $f(x) = \rho + \underline{\lambda} - \left(\overline{\lambda} - \underline{\lambda}\right) x - \underline{\lambda}\left(\underline{\lambda}/\overline{\lambda}\right)^x$. Our DM experiments on project 0 for beliefs above $\overline{\pi}^3$; on 1 for beliefs below $1 - \overline{\pi}^3$; and she gives up otherwise.[2]

Figure 1 depicts a situation where the DM *gives up faster* given no successes; waiting longer for an inconclusive success is not worthwhile. The next proposition shows that, if $\underline{\lambda}$ is sufficiently small, $\overline{\pi}^3 > \overline{\pi}^1$ if and only if $c > \rho$; see the online appendix for a proof.

Proposition 2 (Giving up faster)**.** *Write* $\overline{\pi}^3 = \overline{\pi}^3(\underline{\lambda})$*, and notice that* $\overline{\pi}^3(0) = \overline{\pi}^1$*. Then,* $\lim_{\underline{\lambda}\to 0} \overline{\pi}^{3\prime}(\underline{\lambda}) \propto c - \rho$.

For sufficiently small $\underline{\lambda}$, in particular *if false positives are arbitrarily unlikely*, the DM

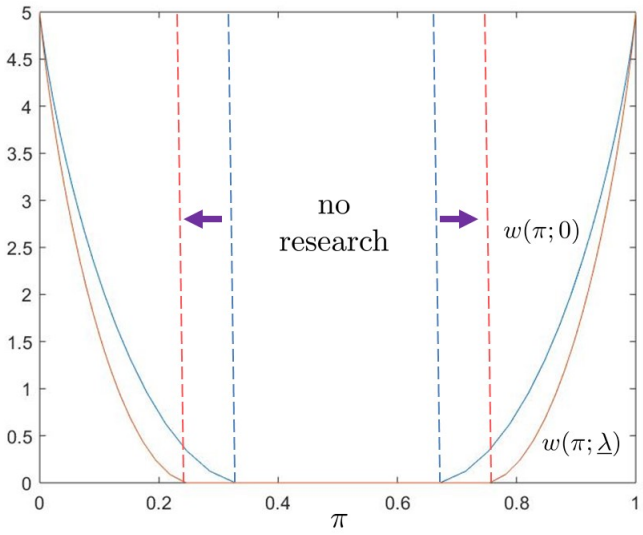

Figure 1: The red curve is the graph of the value function, $w(\pi, \underline{\lambda})$, when $\overline{\lambda} = 200$, $\underline{\lambda} = 50$, $\rho = 10$, and $c = 195$; the blue curve is the counterpart graph for $\underline{\lambda} = 0$, $w(\pi; 0)$.

[2]Thus, unlike KR2, we require that $\overline{\pi}^3 > \frac{1}{2}$. This condition is equivalent to a (partial) characterization of costly research with $\underline{\lambda} > 0$. See Proposition OA1 in the online appendix.

gives up faster conditional on no successes when she is sufficiently patient relative to the cost.

Let $\overline{\pi}^4 > \frac{1}{2}$ be the cutoff such that the DM focuses on project 0 for beliefs above $\overline{\pi}^4$, on 1 for beliefs below $1 - \overline{\pi}^4$, and otherwise works on both.[3] Figure 2 illustrates a case where $\overline{\pi}^4 < \overline{\pi}^2$, so the DM *delays taking on the second project*.[4] The myopic payoff is higher for the more-promising project on its own and makes up for the lower information value compared to simultaneous research. In KR, this example represents a situation where the social gain in value of information is smaller than the loss in myopic social surplus from both agents experimenting. Thus, the social planner delays assigning the second agent to her risky arm.

4. Conclusion

We exploit the analogy between multi-choice bandits and efficient multi-agent experimentation to shed light on multi-choice experimentation. An online appendix extends the analysis to the cases where the decision to neglect projects is irreversible (e.g. if projects are scooped); where joint research "destroys" information because the specific source of success cannot be identified; and where there is a third, independent but riskier project.

The analogy breaks down when successes in multi-agent experimentation are private.

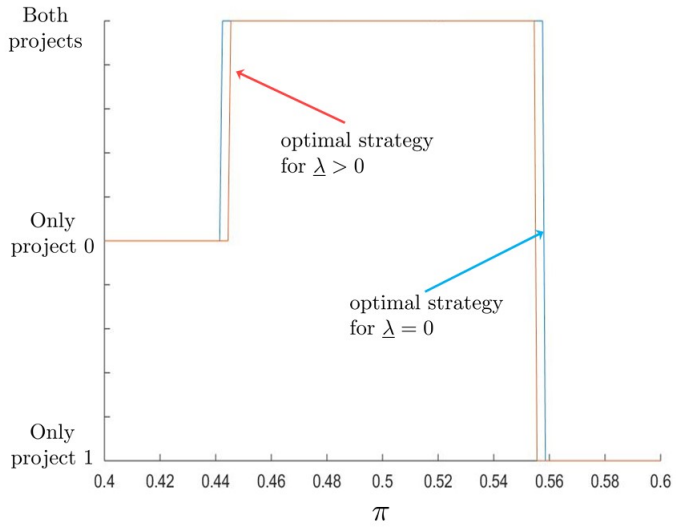

Figure 2: The red curve depicts the optimal strategy for $\overline{\lambda} = 90$, $\underline{\lambda} = 9$, $\rho = 20$, and $c = 60$; the blue curve depicts the optimal strategy in the benchmark $\underline{\lambda} = 0$.

[3] Now, the continuation value is endogenous. The online appendix describes the process to solve for $\overline{\pi}^4$. The equilibrium analysis in KR2 involves a similar construction, but one where only upward jumps can occur.

[4] Of course, we can also construct examples where $\overline{\pi}^4 > \overline{\pi}^2$, so the DM takes on the second project earlier now that both projects yield successes.

In this scenario, our researcher has more information than the social planner, who must rely on progress reports from the agents to learn about the state.

5. References

Bergemann, D and J. Välimäki (2001) "Stationary multi-choice bandit problems" *Journal of Economic Dynamics and Control* **25**, 1585–1594.

Francetich, A and D. Kreps (2018a) "Choosing a good toolkit: Bayes-rule based heuristics" Working paper.

Francetich, A and D. Kreps (2018b) "Choosing a good toolkit: Reinforcement learning" Working Paper.

Keller, G and S. Rady (2010) "Strategic experimentation with poisson bandits" *Theoretical Economics* **5**, 275–311.

Klein, N and S. Rady (2011) "Negatively correlated bandits" *Review of Economics Studies* **78**, 693–732.

Moscarini, G and L. Smith (2001) "The optimal level of experimentation" *Econometrica* **69**(6), 1629–1644.