A Short Survey on Bandit Experiments

Stanton Hudja and Daniel Woods

October 11th, 2024

"Bandit problems embody in essential form a conflict evident in all human action: information versus immediate payoff." - P. Whittle in Gittins (1989)

Introduction

The trade-off between information (exploration) and immediate payoff (exploitation) is prevalent in many decisions. For example, consider a choice between your favorite restaurant and a newly opened restaurant. You could exploit your current information by going to your favorite restaurant, but you would be missing out on information on the new restaurant. On the other hand, you could explore the new restaurant, but you will be sacrificing a good meal in the process. Other examples of this type of trade-off include resource exploration, clinical trials, and innovation. Given the prevalence of this type of trade-off, it is crucial to understand how individuals make exploration decisions.

The incentives of an exploration versus exploitation trade-off are captured by a class of decision problems known as bandit problems. In a bandit problem, an individual faces a repeated choice between various options, where at least one option has an unknown reward distribution. Bandit problems present individuals with an exploration versus exploitation trade-off as an individual only obtains information on the option that they implemented.[1] This type of feedback often results in a choice between implementing an option that maximizes expected immediate reward and collecting information on a relatively unknown option. Thus, individuals have to weigh both the immediate reward and future informational benefits of choosing each option.

The exploration versus exploitation trade-off at the heart of the bandit makes it a staple of many disciplines. In economics, bandits have been used to study dynamic public goods problems (Keller et al., 2005), voting for reforms (Strulovici, 2010), long-term contracts (Halac et al., 2016), and innovation contests (Halac et al., 2017). In addition to economics, bandit problems are of interest to the computer science, psychology, and operations management disciplines.

While bandit problems are quite common in many disciplines, experimental economists have only recently started focusing on human behavior in bandit problems.

[1] This type of feedback is crucial as giving feedback on all possible decisions would remove the incentive to explore.

Bandit experiments currently serve a few uses in economics. First, they are used to study individuals' willingness to explore. These studies typically present individuals with one unknown option and one known option. These experiments try to uncover when individuals become pessimistic enough to give up on exploration and choose the known option. Second, they are used to study individuals' strategies in exploration problems. These studies present individuals with various unknown options and try to estimate the strategies that individuals use to explore these options. Lastly, they are used to test theory that is based in a bandit framework. For example, bandit experiments have been used to test whether subjects free-ride on others' costly exploration in strategic experimentation environments.

Environments

There are two main types of bandit environments that are currently considered in economic experiments. The first is a "one-armed bandit" where an individual is presented with two options. The first option is a known option that has a known reward distribution. This reward distribution returns a constant payoff with some probability. The second option is an unknown option that has an unknown reward distribution. While the reward distribution is unknown, individuals are informed of the prior probabilities of each possible reward distribution that the unknown option can have. An individual repeatedly chooses between these two options until the end of the bandit problem. A one-armed bandit often reduces to a stopping problem as the information in the problem does not change once the known option is chosen.

A common example of a one-armed bandit is the single-agent exponential bandit problem (Keller et al., 2005). In this problem, an individual repeatedly chooses between a safe and risky option. The safe option always returns the reward s. The risky option is an unknown option that is either good with probability p_0 or bad with probability $1 - p_0$. A good risky option returns a constant payoff h with probability λ each time it is chosen (and 0 otherwise). A bad risky option always returns 0. If p_0 is sufficiently high, relative to the other parameters, an individual should start with the risky option and continually update their current belief (p) that the risky option is good based on the observed outcomes. If an individual ever observes h from the risky action, they know that it is good ($p = 1$). If they continue to observe 0 from the risky option, then their belief p should continue to decrease until they become sufficiently pessimistic and decide to switch to the safe option. Once an individual switches to the safe option, they should stick with it as their beliefs are no longer changing.

The single-agent exponential bandit model makes predictions for individuals' willingness to explore. In this environment, an individual should have a cutoff belief that represents the most pessimistic belief with which they are willing to explore the risky action. Once an individual's belief decreases past the cutoff belief, they should choose the safe action forever. An individual's cutoff belief determines their willingness to explore,

which can be thought of as the time that someone is willing to stick with the risky action in the absence of any rewards from the risky action.

One-armed bandits represent the trade-offs that arise when an individual is only considering one exploration option. For example, consider an oil executive who is drilling a new site. The one-armed bandit is a natural representation of this environment as the executive has to determine how long they are willing to continue drilling in the absence of any oil. In this example, drilling is represented by the risky option and the opportunity cost of the resources put into drilling is represented by the safe option. There are many other examples of one-armed bandits such as adopting a new technology, trying a new medication, and implementing a new reform.

The second type of bandit is the "multi-armed bandit" where an individual is presented with multiple options that each have an unknown reward distribution. An individual is presented with the prior probabilities for each reward distribution that each unknown option can take. An individual in this problem repeatedly chooses between these options until the end of the problem. A multi-armed bandit, unlike the one-armed bandit, is not a stopping problem as an individual may repeatedly go back and forth between the options under optimal behavior. This arises as each choice provides information on the option chosen.

A common experimental setup for the multi-armed bandit is the Bernoulli multi-armed bandit problem (Robbins, 1952) where an individual is presented with N unknown options. Each option in this problem has an unknown probability of returning a fixed payoff (otherwise, it returns zero). For example, consider the two-armed indefinite horizon bandit problem in Hudja and Woods (2022) where an individual is presented with two bandit options that each have an unknown reward rate. Individuals are told that each option has a reward rate that is uniformly drawn from {0.000, 0.001, ..., 0.999, 1.000}. Individuals are then told that after each choice there is a four percent probability that the bandit problem will end. This type of bandit problem can be solved using Gittins Indices (see Gittins (1989) for an explanation).

Multi-armed bandits represent the trade-offs that arise when an individual is sampling from many options that each have information value. For example, consider an individual who is shopping for vegan food for the first time. This individual must decide between multiple brands of vegan food that the individual is not familiar with. Multi-armed bandits can also be used to represent decisions with heterogeneous information on the options. For example, a multi-armed bandit could capture the incentives of an individual shopping for cleaning products. An individual is naturally going to have some information on some of the products, but there is still something that can be learned from using each option. Other examples of multi-armed bandits arise in online dynamic A/B testing, financial portfolio design, and job search.

While these two types of bandit environments are the most common, there are many other types that could be implemented in an experiment to study human behavior.[2] Experiments could incorporate features of both of the two aforementioned bandit problems. For example, bandit experiments could have a known option and multiple unknown options. Additionally, while experimentalists focus on bandits with analytical or numerical solutions, experimentalists could analyze other bandit environments. For example, the computer science literature considers other bandits such as the non-stationary 'restless' bandit. In this problem, the underlying reward distribution can change over time.

Experiments

While bandit problems have been studied for almost a century, there has been relatively few bandit experiments. The first experiment to implement a bandit problem, to the best of our knowledge, was conducted in Horowitz (1973). Horowitz (1973) analyzed whether subjects behave optimally in a two-armed bandit and finds that subjects tend to behave suboptimally. Over the next three decades, there were a few more experimental bandit studies (Meyer and Shi, 1995; Banks et al., 1997; Anderson, 2001; Gans et al., 2007). While these early papers pioneered experimental work on bandits, we will not focus on them as their experimental designs often differed from the current best practices in experimental economics.

Experimentalists have used the one-armed bandit to uncover how long individuals are willing to explore an unknown option. For example, Banks et al. (1997) analyze how subjects' willingness to explore in a one-armed bandit changes as the discount rate changes and as the information value of exploration changes. They do not find an effect of either factor on exploration. However, these results should be taken carefully as their experimental design makes it difficult to accurately identify subjects' willingness to explore. This difficulty makes it harder for them to identify if subjects are responding to their treatment changes. Hudja and Woods (2024) address this issue by analyzing behavior in a near-continuous time one-armed bandit, which allows for a much more accurate identification of subjects' willingness to explore. They find that individuals do respond to exploration incentives and that individuals tend to be less willing to explore than predicted by theory. They suggest that under-exploration may be driven by incorrect beliefs about exploration.

One-armed bandit experiments have also been used to address other questions. For example, Hoelzemann and Klein (2021) focus on strategic experimentation and investigate whether subjects free-ride off of other subjects' costly exploration. They find

[2] Banovetz and Oprea (2023) and Hu et al. (2013) implement bandits that somewhat differ from these two types.

strong evidence of free-riding. One-armed bandit experiments have also been used to analyze behavior in innovation contests. One-armed bandits are a natural fit for innovation contests as they can model an environment where an innovation may be possible and an individual needs to decide when to stop exerting effort. Deck and Kimbrough (2017) analyze different types of innovation contests in the lab. They find that contests where individuals are not informed of other contestants' successful innovation attempts outperform contests where individuals are informed of all contestants' successful innovation attempts. Hudja (2021) analyzes the effect of contest size on innovation contests and finds that larger innovation contests tend to result in a higher rate of obtaining an innovation.

Recently, there have been some experimental studies using multi-armed bandits. These studies have focused on the strategies that individuals use for exploration. For example, Gans et al. (2007) analyze the strategies that individuals use in multi-armed bandits. They find that hot-hand strategies tend to fit subject behavior the best out of simple models of discrete choice. Hudja and Woods (2022) build upon this work and reintroduce the multi-armed bandit in a manner that is consistent with best experimental economic practices. They estimate the strategies suggested in Gans et al. (2007), probabilistic strategies from the computer science literature, and new strategies that they developed. They find that a plurality of subjects are best fit by a biased reinforcement learning model, which builds on reinforcement learning by allowing subjects to be biased towards the last option they chose. Hudja et al. (2023) analyze how individuals respond to being blocked from implementing an option in a multi-armed bandit. They find that some subjects exhibit an aversion to being blocked.

Design Choices

The previous section highlights the heterogeneity in bandit experiments. In this section, we discuss different possible design choices and how they may interact with a bandit experiment.

Bandit experiments generally have two types of time horizons: indefinite horizons and finite horizons. Both of these time horizons have implications for experimental design and developing experimental predictions. In an indefinite horizon, future decisions are discounted by a factor δ. An indefinite horizon is generally implemented in an experiment by using a random termination probability: the current decision has a 1-δ chance of ending the bandit problem. Indefinite horizons present subjects with an environment where the expected number of future decisions does not change as the problem continues. This allows the experimenter to use Gittins Indices to make predictions for the experiment in the case of a multi-armed bandit. In a finite horizon, subjects know that there are a fixed number of decisions that they will make in the bandit problem. This allows optimal behavior to be computationally solved using backward induction as long as there are not

too many decisions or too many possible bandit options. One thing to consider about finite horizons is that experimentation incentives generally decrease over the course of a finite horizon bandit problem as the expected number of future decisions is decreasing.

Another important design choice is the prior information provided to subjects on the bandit options. In most theoretical bandit models, individuals have a fully specified prior for each of the bandit options. However, some past studies have chosen to either provide inaccurate initial information or to not disclose relevant prior information on the options. Both of these design choices complicate the researcher's ability to uncover how subjects respond to exploration incentives. In the former case, an individual may realize that the initial information is incorrect, which leads to a mental model that is hidden from the researcher. In the latter case, an individual has a mental prior that the researcher is unaware of.

Future Work

While there has been a recent uptick in bandit experiments, there is still a lot of work to be done. In this subsection, we discuss some areas for future work.

One area of future work is uncovering how individuals value experimentation. In one-armed bandit experiments, individuals tend to under-explore, which suggests that individuals may undervalue exploration in these environments. However, individuals tend to over-explore in multi-armed bandits as individuals tend to choose options that are myopically dominated more often than predicted. This disconnect between these two environments suggests a few possibilities. One possibility is that individuals' valuation of exploration depends on the environment they are placed in. Another possibility is that individuals undervalue exploration, but other factors are leading to artificially high levels of exploration in multi-armed bandits. More work needs to be done to better evaluate these possibilities and uncover individuals' attitudes towards exploration.

Another area of future work is better understanding individuals' randomness in exploration. Hudja and Woods (2022) document evidence of random behavior in multi-armed bandits and find that the best fitting strategies allow for "intelligent" randomness in subject behavior. In their paper, the best fitting strategies suggest that individuals behave more randomly when options are close in expected reward, which leads to more exploration in these cases. It is important for future work to study this "intelligent" randomness as it would suggest a fundamentally different type of exploration than suggested by optimal theory.

References

Anderson, C., 2001. Behavioral models of strategies in multi-armed bandit problems. *Doctoral Thesis*.

Anderson, C., 2012. Ambiguity aversion in multi-armed bandit problems. *Theory and Decision*, 72, pp. 15-33.

Banks, J., Olson, M. and Porter, D., 1997. An experimental analysis of the bandit problem. *Economic Theory*, 10(1), pp. 55-77.

Banovetz, J. and Oprea, R., 2023. Complexity and procedural choice. *American Economic Journal: Microeconomics*, 15(2), pp. 384-413.

Deck, C. and Kimbrough, E., 2017. Experimenting with contests for experimentation. *Southern Economic Journal*, 84(2), pp. 391-406.

Gans, N., Knox, G., and Croson, R., 2007. Simple models of discrete choice and their performance in bandit experiments. *Manufacturing & Service Operations Management*, 9(4), pp. 383-408.

Gittins, J.C., 1989. Multi-armed bandit allocation indices. Wiley and Sons.

Halac, M., Kartik, N., and Liu, Q., 2016. Optimal contracts for experimentation. *The Review of Economic Studies*, 83(3), pp. 1040-1091.

Halac, M., Kartik, N., and Liu, Q., 2017. Contests for experimentation. *Journal of Political Economy*, 125(5), pp. 1523-1569.

Hoelzemann, J. and Klein, N., 2021. Bandits in the lab. *Quantitative Economics*, 12(3), pp. 1021-1051.

Horowitz, A., 1973. Experimental study of the two-armed bandit problem. *Doctoral Thesis*.

Hu, Y., Kayaba, Y., and Shum, M., 2013. Nonparametric learning rules from bandit experiments: The eyes have it! *Games and Economic Behavior*, 81, pp. 215-231.

Hudja, S., 2021. Is experimentation invariant to group size? A laboratory analysis of innovation contests. *Journal of Behavioral and Experimental Economics*, 91, 101660.

Hudja, S. and Woods, D., 2022. Strategies in the multi-armed bandit. *Unpublished Manuscript*.

Hudja, S. and Woods, D., 2024. Exploration versus exploitation: a laboratory test of the single-agent exponential bandit model. *Economic Inquiry*, 62(1), pp. 267-286.

Hudja, S., Woods, D. and Gately, J.B., 2023. Forced experimentation. *Unpublished Manuscript*.

Keller, G., Rady, S., and Cripps, M., 2005. Strategic experimentation with exponential bandits. *Econometrica*, 73(1), pp. 39-68.

Meyer, R. and Shi, Y., 1995. Sequential choice under ambiguity: intuitive solutions to the armed-bandit problem. *Management Science*, 41(5), pp. 817-834.

Robbins, H., 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5), pp. 527–535.

Strulovici, B., 2010. Learning while voting: determinants of collective experimentation. *Econometrica,* 78(3), pp. 933-971.