Nonstationary Multiarm Bandits

Research Brief for the Public Affairs and Policy Lab

Austin Jang
Faculty: Ryan T. Moore

September 30, 2019

The primary concern of policymakers and social scientists is often determining which of several possible interventions is the best. Which e-mail message will yield the most campaign donations? Which program design will be most useful in assisting entrepreneurs in a developing country? One of the best research designs to answer these and similar questions is the multi-arm bandit (MAB) experiment. Unlike traditional randomized experiments, where the allocation of interventions (arms) is assigned statically at the beginning, MAB experiments allocate interventions adaptively throughout. This adaptive allocation setup offers two related advantages over traditional experimental setups. First, they spend fewer resources on suboptimal treatment arms, and second, they are faster at identifying the best intervention.[1]

However, the best intervention may change over time. A campaign message that worked today may be less effective tomorrow, and completely ineffective the day after that. We call this volatility *non-stationarity*. Non-stationarity poses a problem to MAB experiments because the allocation algorithm will allocate more trials to an arm that was optimal earlier in the experiment, even if that arm has since deteriorated in relative performance. Fortunately, several designs have emerged to account for this, called *non-stationary MABs*. These designs fall into two primary camps – those that systematically *discount* prior information when allocating interventions,[2] and those that try to *detect* changes in the arm quality.[3] The goal of our research was to compare and contrast the performance of various trial allocation algorithms. We look at both differences between non-stationary MABs and stationary experimental setups as well as differences within non-stationary MABs. We come to two primary conclusions. First, when choosing between non-stationary algorithms and stationary setups, non-stationary MABs clearly outperform traditional set ups when the experiment faces volatility. Second, when choosing between discounting and detection, the optimal choice is dependent on the environmental context. Below we discuss our methodology, preliminary results, and future steps.

[1] Offer-Westort, Coppock and Green (2018) discuss some limitations and conditionality of these benefits.

[2] Gupta, Granmo and Agrawala (2011); Raj and Kalyani (2017); Garivier and Moulines (2008); Burtini, Loeppky and Lawrence (2015)

[3] Hartland et al. (2006), DaCosta et al. (2008), Mellor and Shapiro (2013)

1 Methodology: Simulating Experiments

To explore our research question, we used simulated experiments to compare the optimality of various trial-allocation setups. To create our simulations, we augmented code from Google's MAB package.[4] The original code was designed to test traditional MAB algorithms in two settings: the stationary environment and the random walk environment. We expanded this code by adding additional classes of non-stationarity as well as adding non-stationary MAB algorithms. Specifically, we added two new types of non-stationarity: linear drift and global switching. Moreover, we expanded upon the traditional Thompson Sampling algorithm, a prominent MAB algorithm in the stationary setting, by creating Thompson Sampling with Bayesian detection and discounted Thompson Sampling. In total, we tested the performance of four allocation algorithms in nineteen different environmental conditions. For robustness, we repeated each experiment 100 times. To assess the performance of each allocation algorithm, we measured each algorithm's relative regret. Relative refers to the difference between the expected reward of pulling the best arm and the actual arm pulled. Pulling the best arm produces a regret of 0.

2 Select Results

The figure below looks at a subset of our results, comparing the algorithm performance in four specifications of the global switching environment, where we alter the size of the switch. The global switching environment represents a sudden shift in the relative performance of each arm. An example of a real world analogue to this simulation is how the effectiveness of certain campaign messages can suddenly change after a scandal breaks out. For our global switching environment, the performance of the best and worst arm switch every twenty-five periods.The four trial allocation algorithms we look at are: equal weighting (the standard experimental setup), Thompson sampling (a stationary MAB), and discounting and detection (the two non-stationary MABs).

We make two observations. First, regardless of the switch size, discounting and detection strategies consistently outperform the traditional allocation strategies by a wide margin. This is demonstrated in the graph because the two non-stationary algorithms, Discount-TS and Detect-TS, have much lower cumulative regret. Importantly, this result held true regardless of the switch size. Furthermore, the non-stationary MABs outperformed the other two setups in all nineteen of our simulation environments. Thus, we advise that social scientists use non-stationary MABs over traditional setups, regardless of the type or degree of non-stationarity. Second, while the choice between stationary and non-stationary allocation algorithms is straightforward, the decision between non-stationary algorithms is much more nuanced. Unlike the comparison between stationary and non-stationary algorithms, no clear "winner" emerges between detection and discounting. For instance, while both algorithms see a boost in performance as the switch size rises, the detection strategy sees a much greater increase in performance.

This theme was consistent across all nineteen environments: detecting and discounting are both improved by similar factors, but they do not experience equivalent relative improve-

[4] Ouyang et al. (2017)

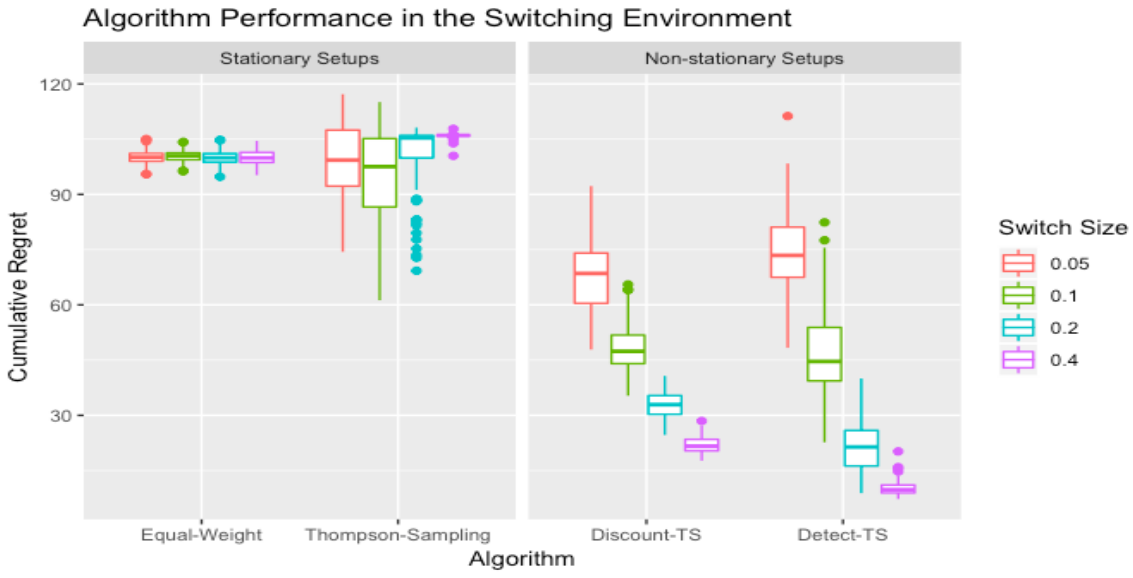

Figure 1: Non-stationary MABs demonstrate clear performance benefits compared to traditional alternatives.

ments. While we only included the subset of our results that looks at variation in switch size in the global switching environment, we saw similar trends with respect to the number of arms, the number of pulls per period, and other measures of non-stationarity in the random walk and linear drift environments.

3 Next Steps

The search for the best policy is a core concern for social scientists and policy makers, but experiments are constrained by time and resources. MAB experiments may offer a significant improvement on traditional experiments. However, the presence of non-stationarity can quickly degrade the performance of a traditional MAB, suggesting the important of non-stationary MABs. Our research has investigated the performance of several allocation algorithms to assess their relative performance in non-stationary environments. Our findings suggest that there is a clear choice between non-stationary and traditional algorithms given any degree of non-stationarity, but that the decision between detecting and discounting strategies is more nuanced.

We plan on continuing our research into MAB trial allocation. Specifically, we plan to investigate two new areas. First, we hope to examine the performance of additional algorithms, such as adaptive discounting.[5] Second, we hope to evaluate algorithms with respect to their ability to estimate causal effects. We are pursuing this in the sutumn of 2019 in an independent study. We are working on a draft of the research paper for submission to a peer-reviewed scholarly journal.

[5]Lu, Adams and Kantas (2019)

References

Burtini, Giuseppe, Jason Loeppky and Ramon Lawrence. 2015. Improving Online Marketing Experiments with Drifting Multi-armed Bandits. In *ICEIS (1)*. pp. 630–636.

DaCosta, Luis, Alvaro Fialho, Marc Schoenauer and Michèle Sebag. 2008. Adaptive operator selection with dynamic multi-armed bandits. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*. ACM pp. 913–920.

Garivier, Aurélien and Eric Moulines. 2008. "On upper-confidence bound policies for non-stationary bandit problems." *arXiv preprint arXiv:0805.3415* .

Gupta, Neha, Ole-Christoffer Granmo and Ashok Agrawala. 2011. Thompson sampling for dynamic multi-armed bandits. In *2011 10th International Conference on Machine Learning and Applications Workshops*. IEEE pp. 484–489.

Hartland, Cédric, Sylvain Gelly, Nicolas Baskiotis, Olivier Teytaud and Michele Sebag. 2006. "Multi-armed bandit, dynamic environments and meta-bandits.".

Lu, Xue, Niall Adams and Nikolas Kantas. 2019. "On adaptive estimation for dynamic Bernoulli bandits." *Foundations of Data Science* 1(2):197–225.

Mellor, Joseph and Jonathan Shapiro. 2013. Thompson sampling in switching environments with Bayesian online change detection. In *Artificial Intelligence and Statistics*. pp. 442–450.

Offer-Westort, Molly, Alexander Coppock and Donald P. Green. 2018. "Adaptive Experimental Design: Prospects and Applications in Political Science.".

Ouyang, Yunbo, Shuchao Bi, Zoey Chu and Chunqiu Zeng. 2017. "google/MAB.".
URL: *https://github.com/google/MAB*

Raj, Vishnu and Sheetal Kalyani. 2017. "Taming non-stationary bandits: A Bayesian approach." *arXiv preprint arXiv:1707.09727* .