Algorithms for Adaptive Experiments that Trade-off Statistical Analysis with Reward: Combining Uniform Random Assignment and Reward Maximization

Authors: Tong Li, Jacob Nogas, Haochen Song, Bingcheng Wang, Huayin Luo, Arghavan Modiri, Nina Deliu, Ben Prystawski, Sofia S. Villar, Audrey Durand, Anna Rafferty, Joseph J. Williams

Introduction

Randomized controlled trials conducted by allocating participants to arms at random is a traditional method to collect data in fields like online education and mobile health. In those trials, adaptive algorithms which change their allocation probability based on data collected are often used to increase user experience and avoid harmful actions. However, adaptive algorithms often collect data with measurement bias[1,2] . In particular, Thompson Sampling exhibits both lowered power (probability of correctly detecting a difference in arms) and higher false positive rates (FPR, probability of incorrectly concluding a difference exists)[3,4]. Given the well-documented challenges of drawing inferences from adaptively collected data[1,3,4,5], two general approaches have been taken to address this issue: developing methods to make correct inferences directly from data collected using existing adaptive algorithms, or changing the adaptive algorithm to collect better data. More closely related to this paper is the latter work that considers changing the adaptive algorithm to balance competing goals of minimizing regret and obtaining data that is useful for drawing conclusions.

This paper presents a novel algorithm, called TS-PostDiff that takes a Bayesian approach to mixing Thompson Sampling and Uniform Random: the probability a participant is assigned using Uniform Random allocation is the posterior probability that the difference between two arms is 'small' (below a certain threshold). This allows for more Uniform Random exploration when there is little or no reward to be gained and thus achieves better trade-off between FPR (False Positive Rate), Statistical Power, and Reward.

Algorithm

In TS-PostDiff, we define ϕ as the posterior probability (after t steps) that the difference in reward between the 2 arms is less than a threshold 'c', where 'c' is a pre-defined threshold parameter. Then, with probability ϕ, choose action uniformly at random and with probability 1 - ϕ, choose action according to an adaptive allocation strategy. An illustration of the algorithm is as follows:

Algorithm 1 TS-PostDiff algorithm pseudocode

Procedure: TS-PostDiff(α,β)

for $t = 1, 2, \ldots$ **do**
  for $k = 1, 2$ **do**
    Sample $p_k \sim$ Beta(α_k, β_k)
  end for
  if $|p_1 - p_2| < c$ **then**
    $x_t = u + 1$, with $u \sim$ Bern$(1/2)$ — {*Follow Uniform Random policy*}
  else
    for $k = 1, 2$ **do**
      Sample $p_k \sim$ Beta(α_k, β_k) — {*Resampling step*}
    end for
    $x_t = \arg\max_k p_k$ — {*Follow TS policy*}
  end if
  Apply x_t and observe the reward y_t
  Update posterior for selected arm
end for

TS-PostDiff was designed to balance between exploration and exploitation: (1) TS-PostDiff uses Uniform Random allocation when there is little evidence for a difference in arm means, reducing the chances of getting a false positive with no or minimal—cost to reward; (2) TS-PostDiff doing TS reward maximization when there seems to be a difference in arm means that is large relative to the parameter for small differences below which maximizing reward is not a high priority (maximizing reward while potentially reducing power).

In evaluating the TS-PostDiff algorithm, we tried to compare it with other algorithms that mix uniform random with MAB allocation to balance between the FPR and power. Epsilon/Top-Two Thompson Sampling $\epsilon/TTTS$[6] adds fixed amounts of UR allocation; it allocates with Uniform Random with probability ϵ and with TS with probability $1 - \epsilon$, and TS-Probability Clipping[5] defines a minimum amount of exploration, which ensures the algorithm always has a chance to explore the seemingly worse arm.

Results

We conducted 10 real-world educational experiments during 2021, where we collected data side by side using both UR (traditional RCT/experiment) and TS-PostDiff (adaptive experiment). The different arms tested alternative interventions to support students in doing better in online programming homework. We then used the effect sizes/arm differences and sample sizes of these real-world experiments to guide our choice of comprehensive simulations, where each algorithm (TS-postdiff, $\epsilon/TTTS$, TS-Probability Clipping, TS, and UR) was simulated over 10,000 experiments, and we compared their reward, false positive rate and power.

10,000 simulations
n = 1,171
FPR = 0.06

Algorithm	Reward	Power
TS-PostDiff (c = 0.12)	0.523(0.000)	0.458(0.005)
ϵ/TTTS (ϵ = 0.34)	0.525(0.000)	0.392(0.005)
TS-ProbClip (p_max = 0.83)	0.524(0.000)	0.365(0.005)

Table 1 compares 3 algorithms under the parameterization that have similar FPR in the null hypothesis setting. We see that, when effect size is 0.045, the rewards from all three algorithms are similar, but TS-PostDiff provides higher Power. This matches the design goal of the TS-PostDiff algorithm: when the true effect size is small, TS-PostDiff will explore more and put higher priority on statistical inference (in which case there won't be much loss in the reward).

$n = 785$							
Algorithm	Effect Size = 0.3			Effect Size = 0.1			Effect Size = 0
	Power	Reward	Prop Opt.	Power	Reward	Prop Opt.	Prop Sup.
Uniform Random	1.000	0.500	0.500	0.806	0.500	0.500	0.500
TS	0.980	0.642	0.974	0.564	0.536	0.860	0.703
TS-PostDiff ($c = 0.1$)	0.999	0.638	0.961	0.776	0.524	0.740	0.529
ϵ/TTTS ($\beta = 0.9375$, $\epsilon = 0.125$)	1.000	0.626	0.920	0.622	0.532	0.823	0.581
TS-ProbClip (prob_max = 0.9375)	0.997	0.628	0.925	0.527	0.534	0.843	0.690

Table 2: Comparison of Power, mean reward, proportion of optimal allocation, and proportion of superior allocation for Uniform Random, TS, TS-PostDiff ($c = 0.1$), ϵ/TTTS ($\beta = 0.9375/\epsilon = 0.125$), and TS-ProbClip ($prob\_max = 0.9375$). Results are shown for effect size 0.3, 0.1 and 0.0 in the first, second and third blocks respectively.

We further simulated a combination of settings where effect size varies from 0 to 0.1, 0.2, and 0.3; sample size goes from 100 to 1000; 40 different values of parameters for each of the three algorithms under their respective range. In our simulations, we consider a 2-armed Bernoulli Bandit setting with arm means p_1 and p_2. Due to limitation of space, we only present results under the setting: sample size equals to 785, effect size equals to 0.1 (p_1=0.55, p_2=0.45) and 0.3 (p_1=0.65, p_2=0.35). We choose c = 0.1 for TS-PostDiff, $\beta = 0.9375$ for $\epsilon/TTTS$, $prob_{max}$= 0.9375 for TS-Prob Clipping. These parameters all provide FPR equal to 0.08 when $p_1 = p_2 = 0.5$ and sample size is 785 (for fair comparison). Under those settings, we see again that TS-PostDiff achieves better Power/statistical inference when effect size is low, and better reward when effect size is large, compared to the other two algorithms.

Discussion

The research presents an innovative algorithm to adapt different sample sizes and effect sizes, with a good trade-off between reward, power and false positive rate. In comparison with other state of the art techniques, it maximizes reward when effect size is large while maintaining similar power, and it prioritizes and achieves higher power when effect size is small and there is little reward to be gained.

There are also several limitations of the current work, that point towards future directions.

TS-PostDiff does require choice of the parameter for what is a 'small' difference, so guidance is needed on choosing the parameter or tuning it based on data. We start with the ubiquitous and important 2-arm trials, and it would be valuable to explore 3 or more arms, such as generalizing TS-PostDiff.

It should be noted that some past work[2] formulates an explicit objective function for the trade-off between reward and measurement error in arm means. We hope the current paper can inform future work on how to more directly formalize an objective function for explicitly trading off between Reward, FPR and Power.

This is particularly valuable for the Thompson Sampling algorithm, as it normally has an internal bias due to its reward maximization nature. Such bias could potentially be further reduced for a dynamic threshold balancing the proportion of UR and TS.

References

1. [Nie et al., 2018] Xinkun Nie, Xiaoying Tian, Jonathan Taylor, and James Zou. Why adaptively collected data have negative bias and how to correct for it. In International Conference on Artificial Intelligence and Statistics, pages 1261–1269, 2018.

2. [Erraqabi et al., 2017] Akram Erraqabi, Alessandro Lazaric, Michal Valko, Emma Brunskill, and Yun-En Liu. Trading off rewards and errors in multi-armed bandits. In Artificial Intelligence and Statistics, pages 709–717, 2017.

3. [Rafferty et al., 2019] Anna Rafferty, Huiji Ying, and Joseph Williams. Statistical consequences of using multi-armed bandits to conduct adaptive educational experiments. JEDM—Journal of Educational Data Mining, 11(1):47– 79, 2019.

4. [Villar et al., 2015] Sofia S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. Statistical science: a review journal of the Institute of Mathematical Statistics, 30(2):199, 2015.

5. [Zhang et al., 2020] Kelly Zhang, Lucas Janson, and Susan Murphy. Inference for batched bandits. In Advances in Neural Information Processing Systems, volume 33, pages 9818–9829, 2020.

6. [Russo, 2016] Daniel Russo. Simple bayesian algorithms for best arm identification. In Conference on Learning Theory, pages 1417–1418, 2016.

bibliographysample-base