MULTI-ARMED BANDITS FOR BUDGET-CONSTRAINED DATA COLLECTION

APOORVA LAL

ABSTRACT. Survey response rates have declined dramatically over the last thirty years. Despite innovation in how we can conduct a survey—on WhatsApp, Facebook, MTurk, and elsewhere —the situation looks dire. One lever with which we can increase response rates is by increasing monetary incentives. But if we were to maximize the objective function of increasing response rates with money, we will end up bankrupting ourselves. This paper casts these as budget constrained online learning problems that are amenable to 'bandit'-based sequential learning algorithms. I propose methods to adaptively allocate monetary incentives and/or recruitment effort to maximise response rates and representativeness in surveys, and identify the best 'arm' under a budget constraint in pilot experiments. I provide simulation-based evidence that these algorithms improve upon current practice.

STANFORD UNIVERSITY, CA

E-mail address: apoorval@stanford.edu.

Date: March 28, 2022.

I thank Jens Hainmueller, Stefan Wager, Justin Grimmer, Avi Acharya, Dan Thompson, Gaurav Sood, Chunyi Zhao, Alex Cloud, Bryce McLaughlin, and Stanford methods lunch participants for comments and suggestions.

1. Introduction

Data collection is a key step in empirical work across the social sciences. In many observational settings, this involves surveying individual respondents about their preferences and circumstances, which, despite the proliferation of survey platforms, has been getting progressively harder as evidenced by declining survey response rates. This has dire consequences for policy-making, polling, and research, since non-representative surveys skew political and policy priorities towards the views expressed by respondents rather than the electorate. While extensive work has been done on reweighting methods to adjust for non-response[1], less attention has been paid to the design and allocation of survey incentives to increase response rates in the design stage so as to obviate the need for extensive reweighting in the analysis stage[2]. We focus on one specific source of heterogeneity in non-response rates across groups - differences in monetary willingness-to-accept (WTA) values - which can be learned using modern adaptive experimentation methods, and propose budget-constrained multi-armed bandits to learn and use these WTA values to increase response rates subject to budget-constraints and representativeness considerations. We provide simulation-based evidence that these algorithms improve upon current practice, and intend to evaluate their performance in field experiments in future work.

Bandit algorithms[3] are an important part of the extensive literature on sequential decision-making and Markov Decision Processes (MDPs) going back to foundational literature in the early-mid 20th century (Thompson, 1933; Wald, 1947; Robbins, 1952). While statistical decision theorists have long been interested in these problems (Gittins, 1979; Manski, 2004; Bergemann and Valimaki, 2006; Berry and Fristedt, 1985; Hirano and Porter, 2009;

[1] for a review of this work, see Caughey et al. (2020) and Hartman, Hazlett, and Sterbenz (2021)

[2] While monetary incentives have been shown to improve response rates in a variety of survey settings (Singer and Ye, 2013; Yan, Kalla, and Broockman, 2018; Dutz et al., 2021), the survey literature provides little practical advice with regard to calibrating these incentives. This calibration should ideally be performed possibly dynamically, as is proposed in the present paper

[3] whose colourful name originates from the analogy to an agent choosing among slot-machines ('one-armed bandits') with unknown reward probabilities and seeking to simultaneously learn the best 'arm' and pull it as often as possible to maximise payoffs.

Stoye, 2009; Hirano and Porter, 2020), bandit algorithms have only recently made inroads into applied research in the social sciences for their potential use in adaptive experimental design (Kasy and Sautmann, 2021; Offer-Westort, Coppock, and Green, 2021; Avivi et al., 2020; Nie, Brunskill, and Wager, 2020) and policy learning for empirical welfare maximisation (Kitagawa and Tetenov, 2018; Athey and Wager, 2021)[4]. These algorithms also underpin the burgeoning literature in reinforcement learning (Sutton and Barto, 2018), which propose methods for a wide class of decision problems ranging from playing board-games to driving cars. At their core, bandit algorithms aim to optimally combine 'exploration' (learning parameters of the data generating process) and 'exploitation' (making the best possible choice to maximise the payoff stream), and as such are adaptable to a wide range of applications in social science beyond the specific application developed in the current paper.

In this paper, we seek to adapt Multi-Armed Bandits (MABs) to the substantive task of survey design. We assume that the researcher has can set K different levels ('arms') of survey payments in order to incentivise respondents[5].For each respondent t, the researcher sets a compensation amount a, and observes whether the respondent completed the survey ($r_t = 1$) or not ($r_t = 0$), which we call the reward. The researcher is therefore interested in *exploration* - learning the expected probability of response for each payment arm μ_a, and *exploitation* - setting the payment arm to maximise the expected reward (total data collected). So far, this is a conventional MAB problem, and several algorithms exist to adaptively pull the right arm (i.e. set the appropriate payment level) to maximise the reward. The researcher can program a bandit algorithm to adaptively set the compensation amount to maximise response rates, or to maximise response rates subject to a representativeness constraint, and so on.

[4]Though these advances aren't without their challenges. Inference, in particular, becomes challenging because adaptive experiments introduce dependence over time and thus mechanically violate the IID structure that common inference procedures rely on. Hadad et al. (2021) propose an AIPW-based reweighting approach to construct valid confidence intervals in such settings.

[5]throughout, we work with a finite set of arms. While MABs with infinite number of arms are an emerging field in reinforcement learning and have been studied theoretically, they are far less tractable, and we therefore leave adaptation of continuous payment schemes to future work

A simple MAB characterisation abstracts from an important consideration in real-world applications of MABs, especially in the survey setting, where each arm costs a fixed amount, and the researcher has a finite budget B. Standard bandit algorithms seek only to maximise the expected reward in the long run, and do not account for whether arms have different 'prices'. Assuming people prefer more money to less money, it is entirely likely for response rates to be (weakly) increasing in compensation. In such a setting, bandit-algorithm may easily provide the trivial answer of 'pay everyone the largest possible amount' since the response probability is largest for these compensation levels, and consequently gather very little data before exhausting the budget. To address this problem, we propose budget-constrained multi-armed bandits, which add cost-considerations to well known budget-free bandit algorithms, Thompson sampling and Upper Confidence Bound (UCB), both of which loosely rely on choosing actions with the largest reward/cost ratio. Using a variety of simulation studies, we show that these budgeted bandits have greater cumulative reward (i.e. they collect more survey responses) than best-performing non-budget algorithms. Next, we propose adjustments to the budgeted algorithm that permits response-rate maximisation subject to representativeness concerns, wherein the researcher is interested in conducting a representative survey with different demographic groups that vary in their response rates to surveys. This is operationalised by dynamically adjusting arm-specific costs to account for representativeness gaps. We find that an amended version of Thompson sampling performs well along both axes - larger sample sizes as well as reasonably high representativeness - relative to random and stratified sampling (which target representativeness alone) and conventional bandit algorithms (which target sample size alone).

Our paper contributes to an emerging literature in the social sciences that seeks to improve experimental design and data-collection efforts using bandit algorithms (Offer-Westort, Coppock, and Green, 2021; Kasy and Sautmann, 2021; Avivi et al., 2020). Most of this literature has been concerned with adapting bandit algorithms to improve the estimation of the average treatment effect in experiments with multiple arms, while the latter is interested in a best-arm identification task. To our knowledge, the present paper proposes

the first bandit algorithm tailor-made to the task of survey-data collection with budget constraints and representativeness considerations, which is a standard problem that many applied researchers face. In doing so, we seek to open a closer dialogue between political methodology and the study of sequential decision making (which includes the general family of methods dealing with Markov decision processes, bandit algorithms, online optimisation, and reinforcement learning). Political methodology has historically been in active conversation with methods developed in statistics, econometrics, and psychology, but has thus far neglected to adapt highly practical engineering methods developed for sequential decision-making such as bandit algorithms. This may be because the broad class of problems that fall under sequential decision making have been studied across a wide variety of disciplines including pure and applied mathematics, operations research, economics, statistics, computer science, engineering, and as such have been discovered and rediscovered by professions with different preferences and terminology, which makes it challenging for applied researchers to delve into the technical literature on the abstract motivating problems in this literature in order to adapt them to their own ends. We hope that our study takes a first step towards illustrating the utility of these sequential decision-making methods for a wide variety of research tasks.

The paper is organised as follows: 2 provides an overview to bandit problems and introduces the proposed algorithms, 3 reports results from a simulation study, and 4 concludes.

2. Overview of bandit algorithms

2.1. Setup.

Consider a setting with binary rewards $r \in \{0, 1\}$, and K arms $a \in \{1, \dots K\} =: [K]$ with unknown probabilities of success $\mu_1, \dots, \mu_k \in [0, 1]$. So, each arm a is associated with an unknown Bernoulli distribution $\mathbb{P}_a$ with mean μ_a, and pulling the ath arm produces IID rewards r_a sampled from corresponding Bernoulli $\mathbb{P}_a$.

In a MAB problem, the agent's task is to maximise total reward $\mathbb{E}\left[\sum_{t=1}^{T} r_{at}\right]$. If the agent knew $\mu_1, \ldots, \mu_K$, the optimal action would be to always play the arm with the highest reward $a^* = \arg\max_{[K]} \mu_k$. However, the agent doesn't know μs, and therefore must incorporate learning μs into the problem. This is the *exploration versus exploitation* trade-off. Rewards are stochastic, so the agent focusses on maximising total expected reward.

$$\mathbb{E}\left[\text{Total Reward}\right] = \mathbb{E}\left[\sum_{t=1}^{T} r_{at}\right] = \sum_{t=1}^{T} \mathbb{E}\left[r_{at}\right] = \sum_{t=1}^{T} \mu_{x_t}$$

where $x_t \in \{a_1, a_2, \ldots, a_K\}$ $, t = 1, 2, \ldots, T$ are the sequence of arm-pulls. Payoff maximisation is equivalent to minimising expected $R(t)$. Lai and Robbins (1985) derive lower bounds on the regret for any 'consistent' algorithm must make on any given instance of the problem, and find that this is a logarithmic function of the number of pulls. Maximising total expected reward is equivalent to minimising cumulative expected regret

$$\begin{aligned}
\mathbb{E}\left[\text{Regret}\right] &= \mathbb{E}\left[\sum_{t=1}^{T} r_{a^*t} - r_{at}\right] \\
&= \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[r_{a^*t}\right]}_{\text{Payoff from always playing } a^*} - \sum_{t=1}^{T} \mathbb{E}\left[r_{at}\right] \\
&= T\mu^* - \sum_{t=1}^{T} \mu_{x_t}
\end{aligned}$$

Almost all MABs the incorporate the empirical mean of the rewards for each arm: $Q_a := \frac{\text{Sum of rewards received from arm } a}{\text{Number of times arm } a \text{ was pulled}}$. Q_a is unbiased for μ_a, and as such, is essential in both 'exploring' (learning $\boldsymbol{\mu}$s) and 'exploiting' (pulling the arm with the largest μ) and is a key component of a bigger class of solution methods known as Q-learning (Sutton and Barto, 2018). We provide an overview of canonical bandit algorithms in appendix A.1.

2.2. **Budgeted Bandits.**

In the standard analysis of bandit algorithms, the agent's goal is to maximise the expected cumulative reward from the sequence of pulls. However, this typically abstracts budget constraints that MABs may face in real-world applications. Pulling each arms may be associated with a fixed (Tran-Thanh et al., 2012) or random (Ding et al., 2013) cost, with the total available budget being set to some value B. Badanidiyuru, Kleinberg, and Slivkins (2018) provide a general framework for the analysis of such problems that combines bandit learning with stochastic integer programming, hence the name *bandits with knapsacks*.

In our substantive application, the budget constraint is particularly important. Since each 'arm' is a monetary reward for survey completion, we necessarily have fixed costs to pulling each arm, and a finite budget. So, under the reasonable assumption that larger payments are more likely to induce responses, we may have $\mu_1 \leq \dots \mu_K$ where $\{1, \dots, K\}$ are ordered by the monetary value of the arm. So, a conventional well-performing MAB, such as Thompson Sampling, might give us a trivial answer, which is to pay everyone the most (i.e. pull arm K with the maximum value). However, this may result in us receiving far too few responses.

2.2.1. **Knapsack-UCB.** To address problems with conventional MABs in budget-constrained settings, we first propose using the *Knapsack-based Upper confidence Bound Exploration and exploitation* (KUBE) algorithm proposed by Tran-Thanh et al. (2012) (henceforth TCRJ). The authors analyse a budget-limited MAB consisting of a machine with K arms, and a total budget of B. By pulling arm a, the agent has to pay c_a, and gets reward r_a. Since B is finite, the sequence of pulls is finite.

TCRJ cast the budgeted bandit an unbounded knapsack problem and propose two algorithms that $O(\log B)$ theoretical upper bound (Tran-Thanh et al., 2012, p1137-38). We concentrate on the computationally more tractable (albeit marginally less performant) fractional-KUBE algorithm. Fractional KUBE chooses the arm with the highest estimated upper-confidence-bound-to-cost ratio. Specifically, this involves pulling the following arm

$$A = \text{argmax}_{[K]} \frac{Q_a + \sqrt{\frac{2 \log t}{n_a}}}{c_a} \tag{2.1}$$

which is the UCB/Cost ratio for each arm, which makes it an intuitive budget-constrained adaptation of UCB.

Algorithm 1: Fractional KUBE Algorithm (adapted from Tran-Thanh et al. (2012, p. 1137))

Param: $B_t = B$ Starting Budget
Param: **Q** Vector of empirical mean of returns for each arm
Param: **N** Number of times each arm has been pulled
Param: **C** Vector of costs for each arm
while $B_t > \min_{[K]} c_a$: *pulling is feasible* **do**
  for $a = 1, \ldots, K$ **do**
    $(\text{UCB/Cost})_a := Q_a/c_a + \left(\sqrt{\frac{2 \log t}{n_a}}\right)/c_a$; // Update UCB/Cost ratio
  end for
  $A = \arg\max_{[K]} (\text{UCB/Cost})_a$; // Identify arm with highest UCB/Cost ratio
  r = BernoulliReward(A) ; // Pull arm a; Draw reward $r \in \{0, 1\}$
  $N_A = N_A + 1$; // Update number of pulls
  $Q_A = Q_A + \frac{1}{N_A}(r - Q_A)$; // Update Empirical Mean
  $B_{t+1} = B_t - c_A$; // Deduct cost of arm from budget
end while

2.2.2. **Cost-normalised Thompson sampling.** Thompson sampling (TS) Thompson (1933) is one of the earliest concepts in sequential learning. Interest in it was renewed through the analysis of its bounds and empirical results documenting its performance (Scott, 2010; Chapelle and Li, 2011; Agrawal and Goyal, 2012). Indeed, it has been shown to be asymptotically optimal for Bernoulli bandit problems, which is our use case. Thompson sampling involves specifying a prior distribution on each of the K bandits: $\pi(\mu_1), \ldots, \pi(\mu_K) =: \Pi_0$,

and updating the posterior Π_t sequentially and selecting the arm with the highest posterior probability of reward $A = \arg\max_{[K]} \pi(\mu_a|x_a)$. For Bernoulli outcomes, this is the most straight-forward, as it involves initialising success and failure counts (α, β), pulling arms, and updating them, as illustrated in 2.

Conventional Thompson sampling under-performs in budget-constrained settings. We propose minor adjustments that do not break the Beta-Binomial conjugacy that forms a large part of the appeal of TS in Bernoulli settings. Specifically, we propose amending the arm choice step alone in Thompson sampling from $\arg\max_{[K]} \mu_a$ to

$$A = \arg\max_{[K]} \frac{\widehat{\mu}_a}{\widetilde{c}_a} \quad \text{where} \quad \widetilde{c}_a = \frac{c_a}{\sum_K c_k}$$

where the denominator $\widetilde{c}_a$ is scaled cost that forces it on $[0, 1]$. This approach is related to the Budgeted-Thompson sampling algorithm (BTS) proposed by Xia et al. (2015), who propose pulling the arm with the highest reward/cost ratio when rewards and costs are random and distributed on $[0, 1]$. In our application, c_as are not random, but can be scaled to lie on the unit interval. Nevertheless, the regret bound in Xia et al. (2015) applies; the budgeted Thompson Sampling algorithm achieves a regret bound of $O(\ln B))$ where B is the budget.

In simulations, we also implement a version of the algorithm (based on the substantive setting of survey design) where incentives are provided conditional on completion of the survey, which mechanically means that c_as are only incurred if the reward r is 1.

2.2.3. **Targeting Representativeness.** To build representative samples, we propose altering Budgeted Thompson sampling to dynamically adjust costs to target representativeness. Intuitively, this permits researchers to dynamically emphasise exploration early on (by

Algorithm 2: Budgeted Thompson Sampling for Bernoulli Bandit

Parameter: $\mathbf{S}, \mathbf{F} = 0$ Success and failure counters for each arm
Param: $\mathbf{C}$ Vector of costs for each arm
while $B_t > \min_{[K]} c_a$: *(pulling is feasible)* **do**
  for a = *1, …, K* **do**
    Draw $\widehat{\mu}_a \sim \text{Beta}\,(S_a + 1, F_a + 1)$; // Draw from posterior
  end for
  $\widetilde{c}_{at} = c_{at} / \sum_K c_{kt}$; // Compute Normalised cost at time t
  $A = \arg\max_{[K]} \widehat{\mu}_a / \widetilde{c}_{at}$; // Identify arm with reward/cost ratio
  r = BernoulliReward(A) ; // Pull arm; draw reward $r \in \{0, 1\}$
  $S_A = S_A + r$; // Update Successes
  $F_A = F_A + (1 - r)$; // Update Failures
  $B_{t+1} = B_t - c_A$; // Deduct cost of arm from budget
end while

shrinking all costs towards common values) and target balance/representativeness afterwards by setting costs to be increasing in over-representation of a given group in the sample. This makes it more likely that the bandit will choose other arms that now appear 'cheaper' because they correspond with groups that are under-represented in the sample. Specifically, we set the cost vector $\mathbf{c}_{at}^g$ for stratum g to

$$\mathbf{c}_{at}^g = \left(1 + \underbrace{\left(\frac{B - b}{B} \right)}_{\text{Remaining budget share}} \psi^g \right)^{\gamma} \mathbf{c}_a$$

where $\psi^g := (\overline{x}_t - \widetilde{x})$ is current over-representation of group g in sample

where x is an indicator for a demographic characteristic g in the sample, $\widetilde{x}$ is the target share of group g in the final sample (for example, the population share of stratum g in the census), $\mathbf{c}_a$ is the initial set of costs. These costs are the same across groups initially, and then begin to grow for groups that are over-represented in the sample ($\psi^g > 0$). These departures from the original costs are scaled by how much of the budget has been exhausted: early in the process, when $\frac{B-b}{B} \approx 0$, group-specific deviations are approximately 0, while later on, the

algorithm prioritises representativeness more. $\gamma \geq 1$ is a positive parameter that controls the degree to which representativeness is prioritised relative to maximising rewards.

There are alternative approaches to incorporate representativeness into the objective function of the bandit algorithm. An information theoretic objective function extending the approach in Russo and Van Roy (2018) may be promising, but requires ex-ante targets on the precision of group-level means that may be unappealing in practice. Multi-objective Multi-armed bandits (MOMABs) (Hüyük and Tekin, 2021), where bandits are typically concerned with multi-valued rewards and may either 'satisfice' on one element of the reward or lexicographically order preferences where optimisation on one axis is prioritised over the other, may also be promising. However, since the secondary objective of representativeness is not a reward per se and is a property of the entire sample collected so far, we are unable to adapt MOMABs for our purposes. Adaptations of the above ideas to the task at hand is a promising avenue for future research.

3. Simulation Study

3.1. Survey bandit simulations.

3.1.1. Simulation Setup.

We conduct simulation studies motivated by the survey application. We simulate data with 10 arms where, for each arm

$$c_a \in \{2, 5, 10, 20\} \qquad \text{Uniform draw from set of costs} \tag{3.1}$$

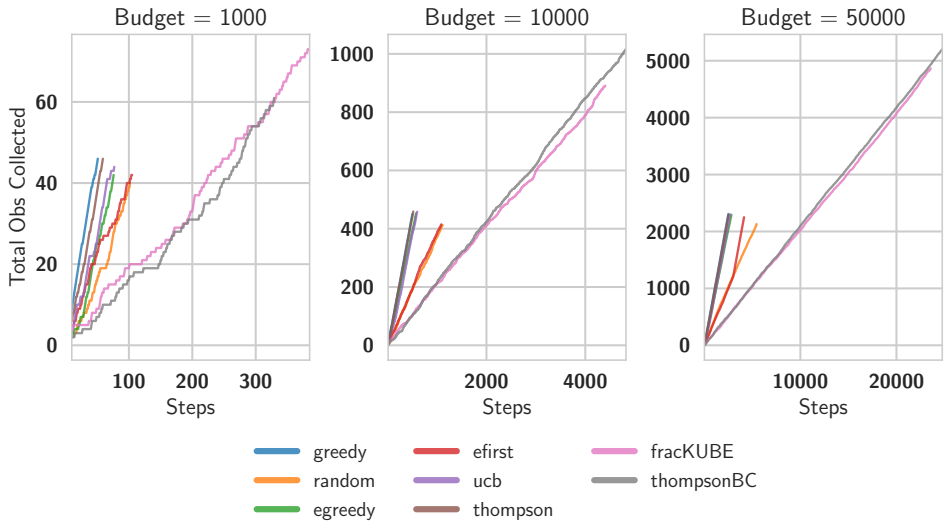
$$\mu_a \sim \text{Beta}\left(\alpha = \max\left(\frac{c_a}{5}, 1\right), \beta = 10/c_a\right) \tag{3.2}$$

Costs are drawn so that each arm costs between \$2 and \$20. The corresponding mean rewards are simulated from a beta distribution such that the reward probability $\mathbb{E}[\mu_a]$ is increasing in c_a, based on our substantive assumption that higher payments are more likely to elicit responses: for $c_a = 2, \mathbb{E}[\mu_a] = \frac{1}{5}$, while for $c_a = 20, \mathbb{E}[\mu_a] = \frac{8}{9}$. The cumulative reward in this setting is analogous to the total number of survey responses, since we model reward = 1 as a complete response.

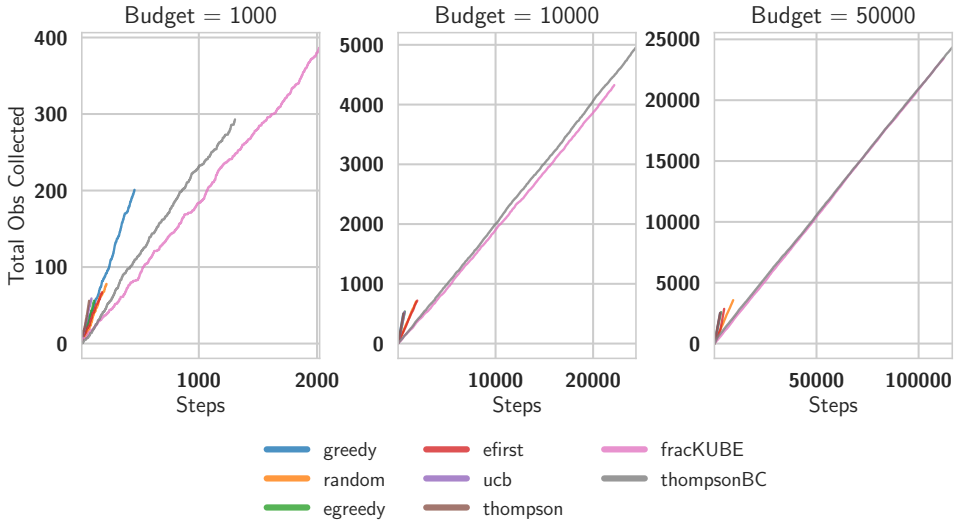
We also consider a case where costs c_a are only paid at the end of the survey (i.e. as a reward the respondent), which makes the budget last a lot longer, especially in the presence of multiple low-reward arms.

3.1.2. **Budget-Constrained Simulation.** In this section, we conduct simulations to mimic our substantive application of eliciting survey-responses under a budget constraint. We run our bandits on the same simulated data as above, but this time introduce a budget constraint B; each algorithm pulls its best arm until it exhausts the budget. We report the cumulative reward for three different budget values $B \in \{1000, 10000, 50000\}$ in fig 1a, cumulative rewards when costs are incurred contingent on response in fig 1b, and corresponding arm-pull sequences in figs 2a and 2b respectively.

The results are drastically different from the non-budget-constrained simulation. From fig 1a, we see that the budget-aware algorithms collect roughly twice as much data as the budget-agnostic ones, and this gap grows to nearly 8-12x more when costs are incurred conditional on success (1b). The reason for this, as we see in fig 2a, is that budget-agnostic algorithms (Thompson and $\varepsilon-$ first) exhaust their budget well before the budgeted algorithms by pulling expensive but not cost-efficient arms. Thompson sampling pulls the most expensive arm (with reward probability of 0.92 but cost of \$20) too frequently, while $\varepsilon-$ first mixes inefficiently across arms. In contrast, KUBE and budgeted Thompson ('thompsonBC') consistently pull the arm with the highest UCB/Cost ratio and therefore run for



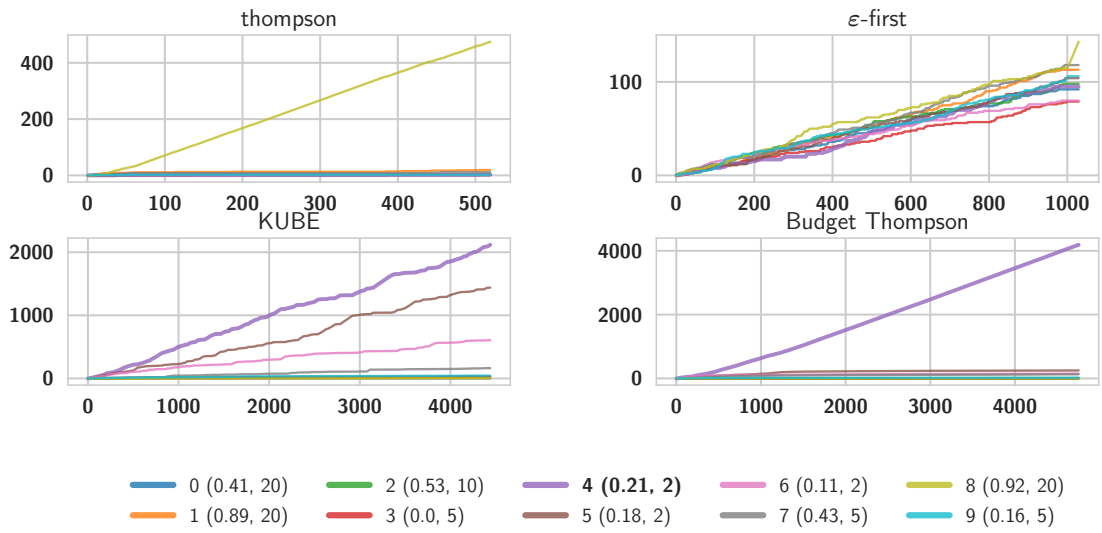
(A) Cumulative rewards for budgeted bandits. Maximum value of X-axis label indicates period at which the budget is exhausted by the best-performing bandit.



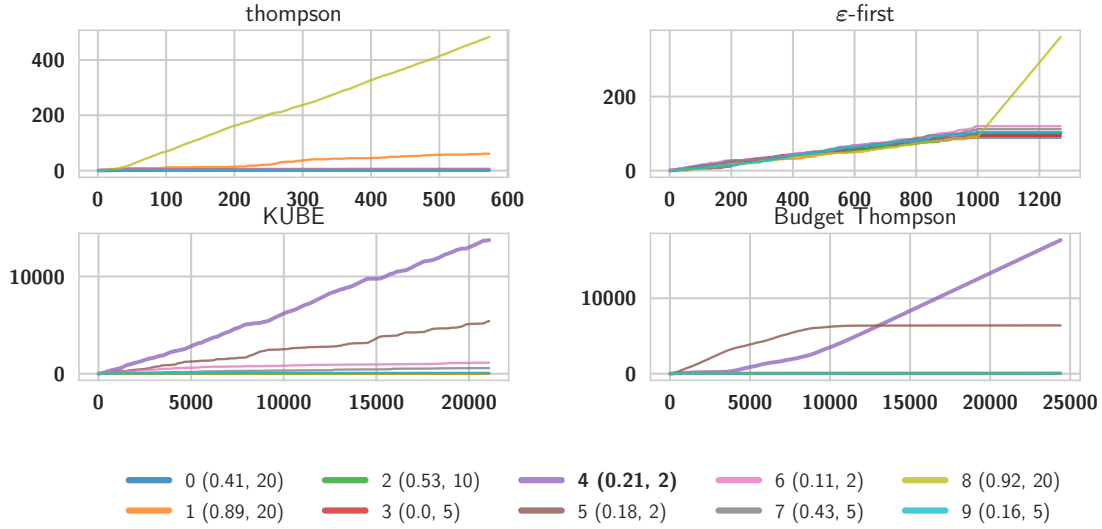
(B) Cumulative rewards for budgeted bandits where costs are incurred conditional on survey completion (i.e. iff reward = 1). Maximum value of X-axis label indicates period at which the budget is exhausted by each algorithm.

FIGURE 1. Cumulative rewards for each algorithm in budgeted simulations. Panels report runs with increasing budgets of $1000, $10000, and $ 50000 respectively.

many more steps before exhausting the budget, and consequently collect a much larger cumulative reward.



(A) Cumulative pull sequence for each arm for budgeted bandits. Maximum value of X-axis label indicates period at which the budget is exhausted by the best algorithm.



(B) Cumulative pull sequence for each arm for budgeted bandits where costs are incurred conditional on reward. Maximum value of X-axis label indicates period at which the budget is exhausted by each algorithm.

FIGURE 2. Cumulative number of pulls of each arm by each algorithm in budgeted simulations. The legend labels indicate arm index, followed by reward probability and cost, with the best arm (by reward probability - cost ratio) in bold.

3.1.3. **Budget-constrained Simulation targeting representativeness.** Next, we construct a simulation study to correspond with the cost-adjustment method for targeting representativeness outlined in section 2.2.3. We simulate data where there are two strata, E and F,

and members of group F respond to surveys at a substantially lower rate than E. We target equal shares of the two groups in our final sample because they are evenly distributed in the population.

We work with 5 payment levels $\$2, 5, 7, 10, 20$, which, combined with 2 groups, gives us 10 arms. Reward probabilities are increasing in payment, where μ_a^E is generated by eqn 3.2, and $\mu_a^F = (0.4, 0.5, 0.6, 0.7, 0.8) \cdot \mu_a^E$, so group F is particularly unlikely to respond for small payments, and this gap is decreasing in the magnitude of the payment.

We benchmark the performance of Budgeted Thompson sampling against random sampling and Thompson Sampling. The former is a 'pure-exploration' algorithm that assigns equal number of draws to each group (which approximates random sampling in surveys). The latter does not account for costs or representativeness at all. Budgeted Thompson sampling with different values of γ prioritises representativeness to varying degrees, with $\gamma = 0$ holding costs constant, while large values of γ prioritise representativeness more.

We report simulation results in figures 3, 4, and 5. From fig 3, we see that higher values of γ penalise over-representation of group E more by raising costs higher. In fig 4, we see that because group F responds to surveys at lower rates than E, conventional bandit algorithms like Thompson sampling or Budgeted-Thompson sampling (with $\gamma = 0$, which doesn't prioritise exploration at all) collect plenty of data, but the sample ends up with a almost exclusively group Es. Increasing γ, prioritises representativeness and therefore yields a sample closer to 50-50. For high values of γ, we get samples that are nearly as representative as full exploration (random sampling) and have larger sample sizes (based on x-axis values). Gains from Thompson are even larger when costs are conditional on reward in 5, where the total observations collected is significantly larger than with random sampling.

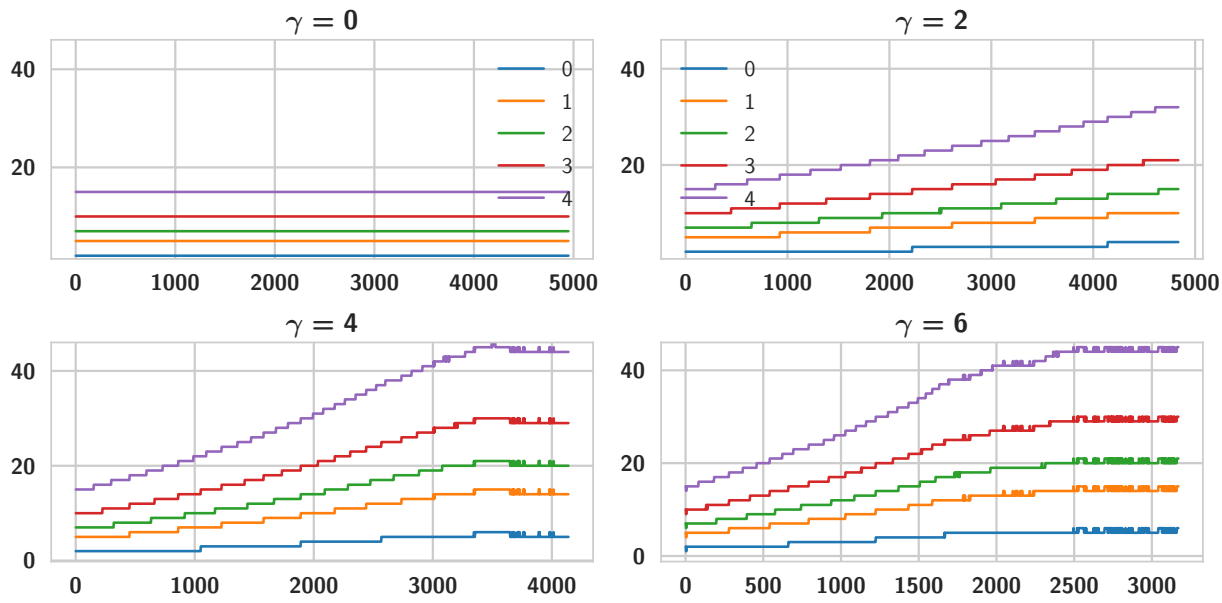


FIGURE 3. Costs for group E over time for different values of γ. $\gamma = 0$ holds costs at the original values, while higher values of γ raise costs more in response to over-representation of Es in the sample, thereby inducing the bandit algorithm to pick arms corresponding with group F.

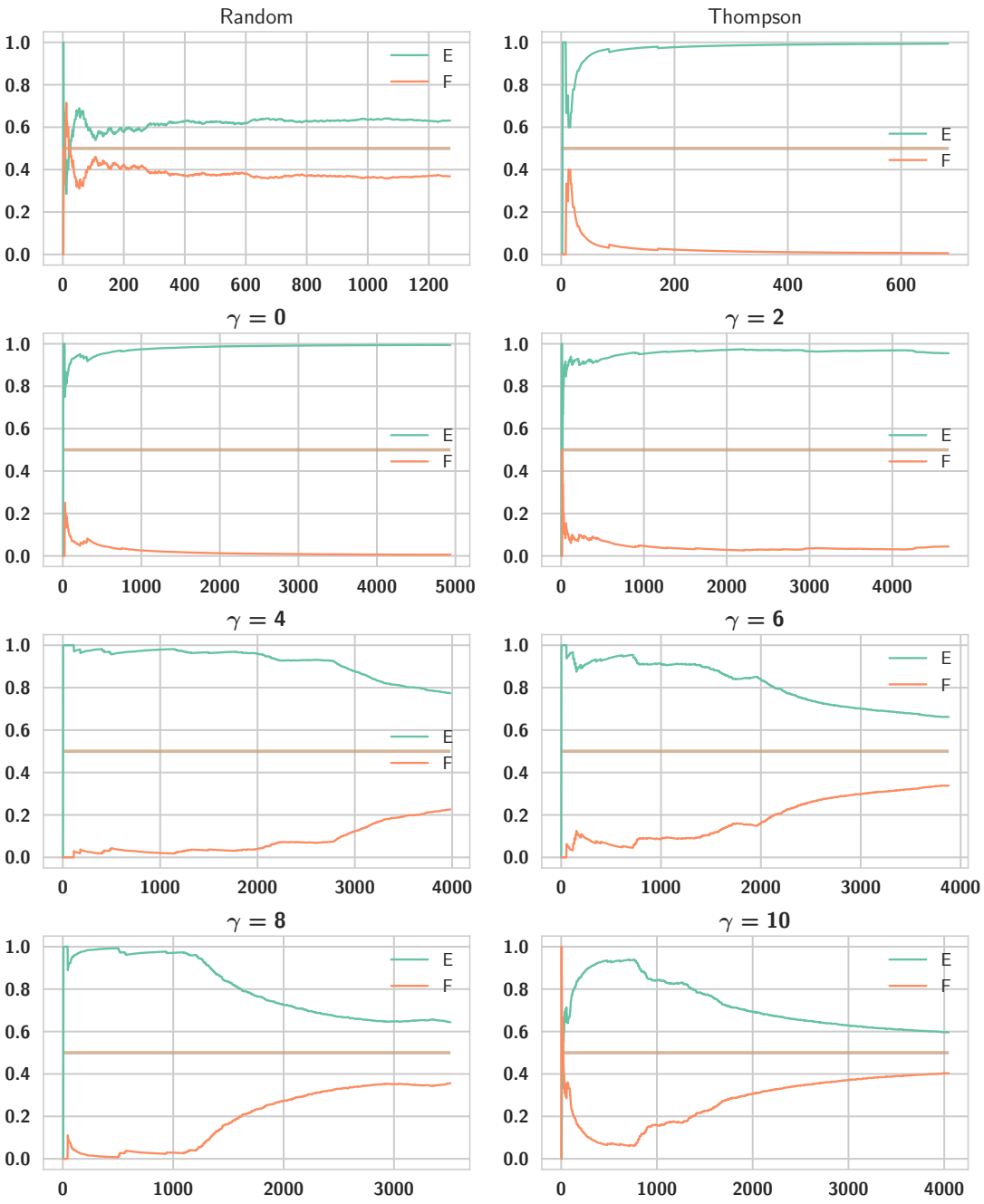


FIGURE 4. Shares of group E and F over time for different bandit algorithms, with the bottom four varying the degree of 'representativeness prioritisation' γ in 3.2. The x-axis values indicate the total number of observations collected by each method. We find that large values of γ produce larger samples than random sampling and comparable in their representativeness.

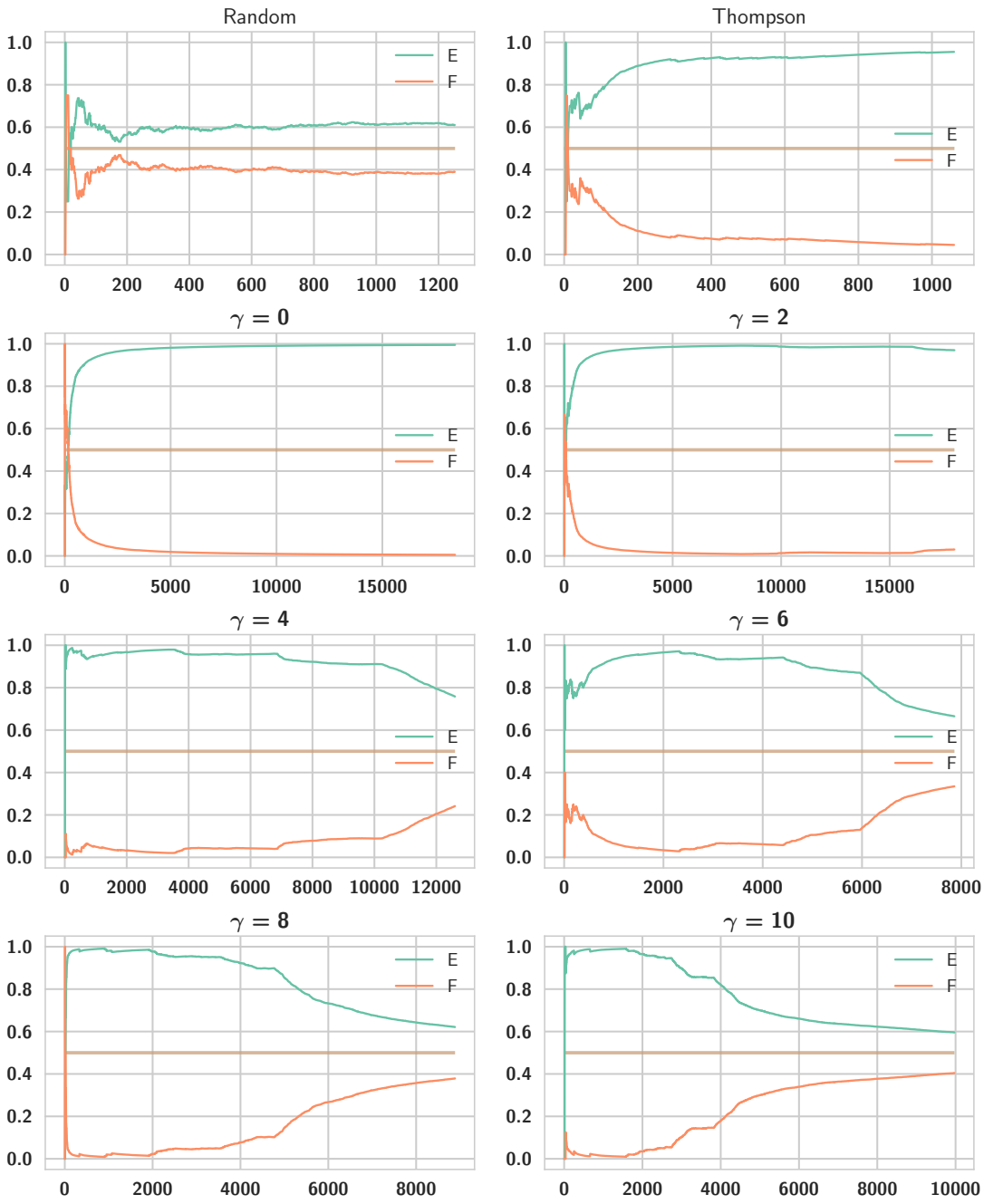


FIGURE 5. Shares of group E and F over time for different bandit algorithms with costs incurred conditional on reward, with the bottom four varying the degree of 'representativeness prioritisation' γ in 3.2. The x-axis values indicate the total number of observations collected by each method. We find that large values of γ produce larger samples than random sampling and comparable in their representativeness.

4. **Conclusion**

We propose a simple set of bandit algorithms to improve response rates in survey data collection more effectively, and demonstrate that these algorithms perform well relative to budget-less bandits and random sampling in simulation studies. Paying respondents in order to induce response is likely to improve response rates and representativeness relative to current practice of providing little to no incentive and relying on stratified sampling and weights to construct representative measures from highly non-representative and uneven response rates from different demographic groups. Simple adjustments to these bandit algorithms also yield larger and more representative samples than either random sampling or bandit algorithms.

References

Agrawal, Shipra and Navin Goyal (2012). "Analysis of thompson sampling for the multi-armed bandit problem". *Conference on learning theory*, pp. 39–1 (cit. on p. 8).

Athey, Susan and Stefan Wager (2021). "Policy learning with observational data". en. *Econometrica: journal of the Econometric Society* 89.1, pp. 133–161 (cit. on p. 3).

Avivi, Hadar et al. (2020). "Adaptive Correspondence Experiments". *AER, Papers and Proceedings* (cit. on pp. 3, 4).

Badanidiyuru, Ashwinkumar, Robert Kleinberg, and Aleksandrs Slivkins (Mar. 2018). "Bandits with Knapsacks". *Journal of the ACM* 65.3, pp. 1–55 (cit. on p. 7).

Bergemann, Dirk and Juuso Valimaki (Jan. 2006). "Bandit Problems" (cit. on p. 2).

Berry, Donald A and Bert Fristedt (1985). "Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability)". *London: Chapman and Hall* 5.71-87, pp. 7–7 (cit. on p. 2).

Caughey, Devin et al. (2020). *Target Estimation and Adjustment Weighting for Survey Nonresponse and Sampling Bias*. Cambridge University Press (cit. on p. 2).

Chapelle, Olivier and Lihong Li (2011). "An empirical evaluation of thompson sampling". *Advances in neural information processing systems* 24, pp. 2249–2257 (cit. on p. 8).

Ding, Wenkui et al. (2013). "Multi-armed bandit with budget constraint and variable costs". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 27. 1 (cit. on p. 7).

Dutz, Deniz et al. (2021). "Selection in Surveys". *NBER Working Paper* (cit. on p. 2).

Gittins, John C (1979). "Bandit processes and dynamic allocation indices". *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2, pp. 148–164 (cit. on p. 2).

Hadad, Vitor et al. (Apr. 2021). "Confidence intervals for policy evaluation in adaptive experiments". en. *Proceedings of the National Academy of Sciences of the United States of America* 118.15 (cit. on p. 3).

Hartman, Erin, Chad Hazlett, and Ciara Sterbenz (2021). "Kpop: A kernel balancing approach for reducing specification assumptions in survey weighting". *arXiv preprint arXiv:2107.08075* (cit. on p. 2).

Hirano, Keisuke and Jack R Porter (2009). "Asymptotics for statistical treatment rules". en. *Econometrica: journal of the Econometric Society* 77.5, pp. 1683–1701 (cit. on p. 2).

– (Jan. 2020). "Asymptotic analysis of statistical decision rules in econometrics". *Handbook of Econometrics*. Ed. by Steven N Durlauf et al. Vol. 7. Elsevier, pp. 283–354 (cit. on p. 3).

Hüyük, Alihan and Cem Tekin (May 2021). "Multi-objective multi-armed bandit with lexicographically ordered and satisficing objectives". *Machine learning* (cit. on p. 11).

Kasy, Maximilian and Anja Sautmann (2021). "Adaptive treatment assignment in experiments for policy choice". en. *Econometrica: journal of the Econometric Society* 89.1, pp. 113–132 (cit. on pp. 3, 4).

Kitagawa, Toru and Aleksey Tetenov (2018). "Who should be treated? empirical welfare maximization methods for treatment choice". *Econometrica* 86.2, pp. 591–616 (cit. on p. 3).

Lai, Tze Leung and Herbert Robbins (1985). "Asymptotically efficient adaptive allocation rules". *Advances in applied mathematics* 6.1, pp. 4–22 (cit. on p. 6).

Manski, Charles F (2004). "Statistical treatment rules for heterogeneous populations". *Econometrica* 72.4, pp. 1221–1246 (cit. on p. 2).

Nie, Xinkun, Emma Brunskill, and Stefan Wager (2020). "Learning when-to-treat policies". *Journal of the American Statistical Association*, pp. 1–18 (cit. on p. 3).

Offer-Westort, Molly, Alexander Coppock, and Donald P Green (2021). "Adaptive Experimental Design: Prospects and Applications in Political Science". *American Journal of Political Science* (cit. on pp. 3, 4).

Robbins, Herbert (Sept. 1952). "Some aspects of the sequential design of experiments". en. *Bulletin of the American Mathematical Society* 58.5, pp. 527–536 (cit. on p. 2).

Russo, Daniel and Benjamin Van Roy (Feb. 2018). "Learning to optimize via information-directed sampling". *Operations research* 66.1, pp. 230–252 (cit. on p. 11).

Scott, Steven L (Nov. 2010). "A modern Bayesian look at the multi-armed bandit". en. *Applied Stochastic Models in Business and Industry* 26.6, pp. 639–658 (cit. on p. 8).

Singer, Eleanor and Cong Ye (Jan. 2013). "The Use and Effects of Incentives in Surveys". *The Annals of the American Academy of Political and Social Science* 645.1, pp. 112–141 (cit. on p. 2).

Stoye, Jörg (2009). "Minimax regret treatment choice with finite samples". *Journal of Econometrics* 151.1, pp. 70–81 (cit. on p. 3).

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press (cit. on pp. 3, 6).

Thompson, William R (1933). "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". *Biometrika* 25.3/4, pp. 285–294 (cit. on pp. 2, 8).

Tran-Thanh, Long et al. (Apr. 2012). "Knapsack based optimal policies for budget-limited multi-armed bandits". arXiv: 1204.1909 [cs.AI] (cit. on pp. 7, 8).

Wald, Abraham (1947). "Foundations of a general theory of sequential decision functions". *Econometrica, Journal of the Econometric Society*, pp. 279–313 (cit. on p. 2).

Xia, Yingce et al. (2015). "Thompson sampling for budgeted multi-armed bandits". *arXiv preprint arXiv:1505.00146* (cit. on p. 9).

Yan, Alan, Joshua Kalla, and David Broockman (2018). "Increasing Response Rates and Representativeness of Online Panels Recruited by Mail:Evidence from Experiments in 12 Original Surveys". *Working paper* (cit. on p. 2).

Appendix A. Additional Material

A.1. Canonical Bandit Algorithms.

Here, we provide a brief overview of some standard approaches to solving the bandit problem. Most Bernoulli bandit algorithms take the form of algorithm 3.

Algorithm 3: General Algorithm for Bernoulli Bandit

Param: **Q** Vector of empirical mean of returns for each arm
Param: **N** Number of times each arm has been pulled
Param: **S** Number successes for each arm
Param: **F** Number failures for each arm
for $t = 1, \ldots, T$ **do**
  a = PickArm(**Q, N, S, F**) ; // Most Bandits only use **Q**,**N**
  r = BernoulliReward(a) ; // Pull arm a; get $r \in \{0, 1\}$
  $N_a = N_a + 1$; // Update number of pulls
  $Q_a = Q_a + \frac{1}{N_a}(r - Q_a)$; // Update Empirical Mean
  $S_a = S_a + r$; // Update Successes
  $F_a = F_a + (1 - r)$; // Update Failures
end for

A.1.1. **Random.** The most basic approach is to randomly pick arms in each period, which is a 'pure exploration' approach. This is obviously sub-optimal since we do not learn μs and adapt our decisions accordingly.

A.1.2. **Greedy.** At the other extreme end, an 'exploit-only' approach is to pull each arm m times, and thereafter pull the arm with the highest empirical average $A = \text{argmax}_{[K]}(Q_a)$. This approach is susceptible to getting stuck on sub-optimal arm thanks to insufficient exploration.

A.1.3. **ε-first.** An $\varepsilon-$first approach divides play into an initial explore phase, followed by an exploit phase. Specifically, it sets aside the first m rounds to learn the μs, and then pulls the best estimated arm $A = \text{argmax}_a(Q_a)$ repeatedly thereafter. As with greedy selection, this

leaves one open to estimation errors that might lead to one choosing a suboptimal action ad nauseam.

A.1.4. **ε-greedy.** An $\varepsilon-$ greedy approach is a noisy version of the Greedy approach, wherein we allow for an $\varepsilon-$ probability exploration. So, the arm with the highest empirical average is picked with probability $1-\epsilon$, and a random arm is picked with complementary probability ϵ.

The probability of selection for each arm a is

$$P\left(a_i\right) = \begin{cases} 1-\epsilon+\frac{\epsilon}{K} & \text{, if } a_i = \text{argmax}_{[K]}(Q_a) \\ \frac{\epsilon}{K} & \text{, otherwise} \end{cases}$$

A.1.5. **UCB.** This strategy is based on the 'Optimism in the Face of Uncertainty' principle and was proposed by Lai and Robbins (1985) and extended by Auer, Cesa-Bianchi, and Fischer (2002). We know Q_a is an unbiased estimate of μ_a. After N_a pulls of arm a we can be quantify how close Q_a is to μ_a using Hoeffding's Inequality, which yields the bound

$$\textbf{Pr}\left(|\mu_a - Q_a| \geq \epsilon\right) \leq 2\exp\left(-2N_a\epsilon^2\right)$$

Using a one sided version of this inequality we get:

$$\textbf{Pr}\left(\mu_a \geq Q_a + \epsilon\right) \leq \exp\left(-2N_a\epsilon^2\right)$$

So for arm a, whose average reward is Q_a after it has been pulled N_a times, μ_a exceeds the upper confidence bound(UCB) with probability $p = \exp\left(-2N_a\epsilon^2\right)$ We want the probability that μ_a exceeds UCB to decrease with t, the number of arm pulls so far. A common choice

(UCB1) is to use $p = t^{-4}$, which ensures that we select the optimal action in the limit as $t \to \infty$

$$\epsilon = \sqrt{\frac{-\log p}{2N_a}} = \sqrt{\frac{2\log t}{N_a}}$$

This implies the following arm choice

$$A = \text{argmax}_{[K]} \left[Q_a + \sqrt{\frac{2\log t}{N_a}} \right]$$

Intuitively, the second piece of the UCB shrinks as one learns μ_a more precisely (by pulling a more of then so $N_a \to \infty$).

A.1.6. **Thompson Sampling.** Thompson sampling (TS) Thompson (1933) is one of the earliest concepts in sequential learning. Interest in it was renewed through the analysis of its bounds and empirical results documenting its performance (Scott, 2010; Chapelle and Li, 2011; Agrawal and Goyal, 2012). Indeed, it has been shown to be asymptotically optimal for Bernoulli bandit problems, which is our use case.

Thompson sampling involves specifying a prior distribution on each of the K bandits: $\pi(\mu_1), \ldots, \pi(\mu_K)$, and sequentially selecting the arm with the highest posterior probability of reward and updating these posteriors.

$$A = \arg\max_{[K]} \pi(\mu_a | x_a)$$

A conventional choice for the Thompson sampling algorithm assumes a prior Beta$(1, 1) \equiv$ U $[0, 1]$ on μ_a for each arm. The Beta distribution is the conjugate prior for Bernoulli rewards because the posterior is also beta-distributed. Priors can be chosen to reflect substantive

knowledge on the part of the researcher: in the survey setting, since it is ex-ante known that response probabilities vary by group, priors can be chosen appropriately. The Bernoulli likelihood given s_a observed successes and f_a failures is

$$p(s_a, f_a \mid \mu_a) = \begin{pmatrix} s_a + f_a \\ s_a \end{pmatrix} \mu_a^{s_a}(1-\mu_a)^{f_a}$$

Then the posterior after observing data $\mathcal{D} := (s_a, f_a)$ is in the same family and can be written as

$$\begin{aligned} \pi(\mu_a|\mathcal{D}) &\propto \pi(\mu_a)\pi(\mathcal{D}|\mu_a) \\ &\propto \underbrace{\mu_a^{1-1}(1-\mu_a)^{1-1}}_{\text{Prior}} \overbrace{\mu_a^{s_a}(1-\mu_a)^{f_a}}^{\text{Likelihood}} \\ &\propto \mu_a^{1-1+s_a}(1-\mu_a)^{1-1+f_a} \end{aligned}$$

Thus the posterior distribution for μ_a is $\mu_a|\mathcal{D} \sim \text{Beta}\,(1+s_a, 1+f_a)$. This simply updates the α parameter in our posterior by the number of successes, and the β parameter by the number of failures. This algorithm is outlined in algorithm 4.

Algorithm 4: Thompson Sampling for Bernoulli Bandit

Parameter: $\mathbf{S}, \mathbf{F} = 0$ Success and failure counters for each arm)
for *t = 1, …, T* **do**
  for *a = 1, …, K* **do**
    Draw $\mu_a \sim \text{Beta}\,(S_a + 1, F_a + 1)$; // Draw from mean posterior
  end for
  a = $\arg\max_{[K]} \mu_a$; // Pull arm with highest draw for μ_a
  r = BernoulliReward(μ_a) ; // Draw reward $r \in \{0, 1\}$
  $S_a = S_a + r$; // Update Successes
  $F_a = F_a + (1 - r)$; // Update Failures
end for

This simple setup is well-understood and has good theoretical properties. The beta posterior becomes more and more concentrated around the empirical mean $S_a/(S_a+F_a)$ for each arm k as the number of plays increases.

A.2. **Additional Figures and tables.**

A.2.1. **Non-budget constrained simulation.** As a warmup, we compute a simple simulation, wherein we ignore arm-specific costs and consequently the budget constraint, and run each algorithm for a fixed number of periods. This is a standard MAB simulation where we use *cumulative reward* as our performance metric. We report the cumulative reward for three different horizons $T \in \{1000, 10000, 50000\}$ in figure A1, and the sequence of arm pulls in appendix figure A2. Consistent with Chapelle and Li (2011) and Russo et al. (2018), we find that Thompson-sampling is typically best-performing among the algorithms under consideration, closely followed by UCB and $\varepsilon-$first. For the longest time horizon ($\approx$ asymptopia), UCB and Thompson have almost identical cumulative rewards, followed by $\varepsilon-$ first.

A.2.1.1. Computation. Computation performed in Hunter (2007) and Van Der Walt, Colbert, and Varoquaux (2011).

References

Agrawal, Shipra and Navin Goyal (2012). "Analysis of thompson sampling for the multi-armed bandit problem". *Conference on learning theory*, pp. 39–1 (cit. on p. 25).

Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer (2002). "Finite-time analysis of the multiarmed bandit problem". *Machine learning* 47.2, pp. 235–256 (cit. on p. 24).

Chapelle, Olivier and Lihong Li (2011). "An empirical evaluation of thompson sampling". *Advances in neural information processing systems* 24, pp. 2249–2257 (cit. on pp. 25, 27).

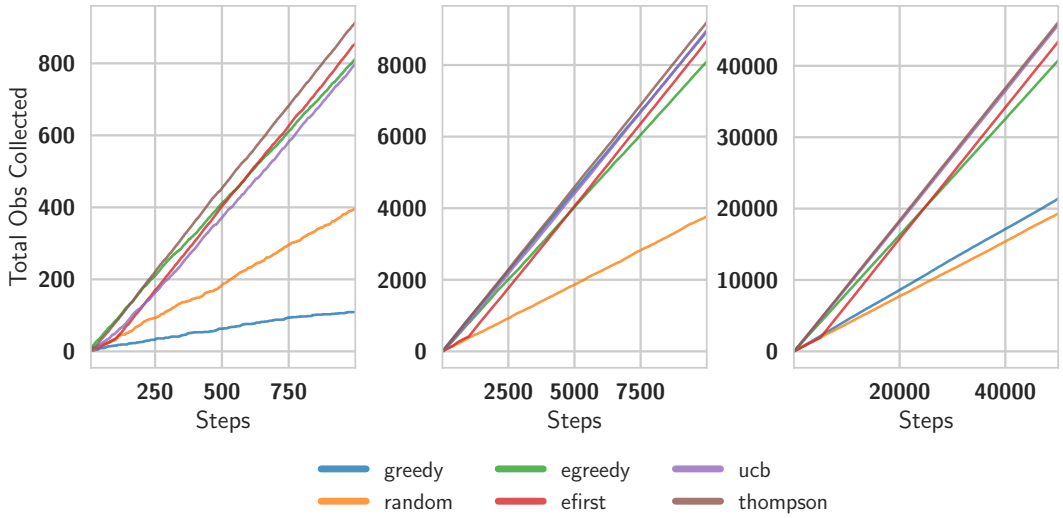


FIGURE A1. Cumulative rewards for each algorithm in budget-less simulations. Panels report runs for 1000, 10000, and 50000 run-horizons respectively

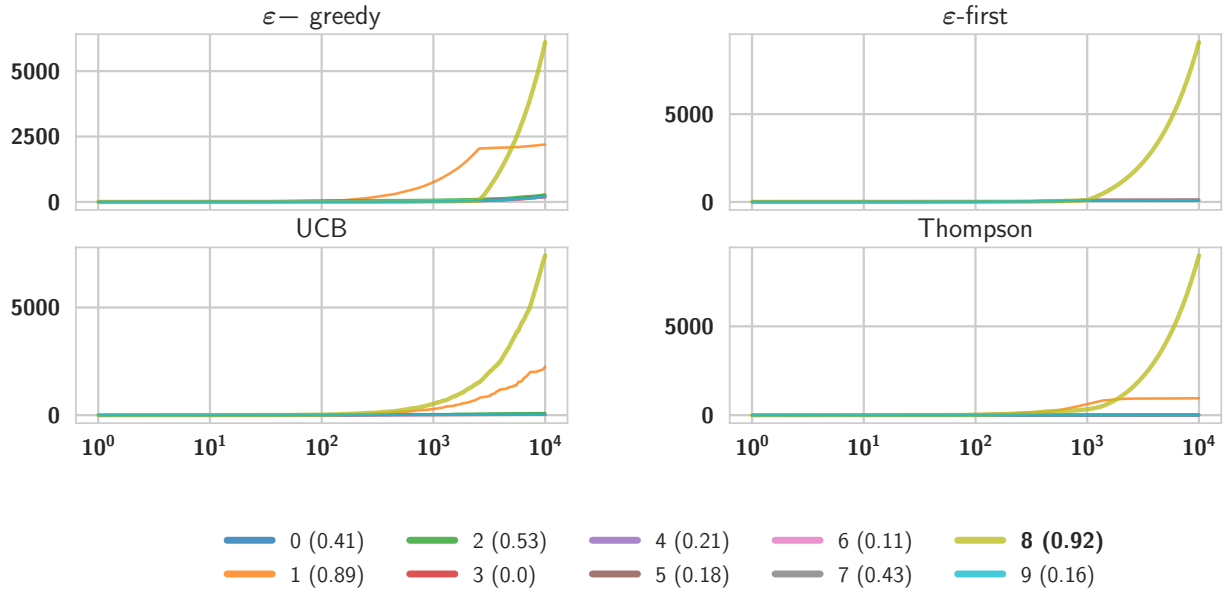


FIGURE A2. Cumulative number of pulls of each arm by each algorithm in budget-less simulations with 10,000 runs. The legend labels indicate arm index, followed by reward probability and cost, with the best arm (by reward probability) in bold.

Hunter, John D. (2007). “Matplotlib: A 2D Graphics Environment”. *Computing in science & engineering* 9.3, pp. 90–95 (cit. on p. 27).

Lai, Tze Leung and Herbert Robbins (1985). “Asymptotically efficient adaptive allocation rules”. *Advances in applied mathematics* 6.1, pp. 4–22 (cit. on p. 24).

Russo, Daniel et al. (2018). “A Tutorial on Thompson Sampling”. *Foundations and Trends® in Machine Learning* 11.1, pp. 1–96 (cit. on p. 27).

Scott, Steven L (Nov. 2010). “A modern Bayesian look at the multi-armed bandit”. en. *Applied Stochastic Models in Business and Industry* 26.6, pp. 639–658 (cit. on p. 25).

Thompson, William R (1933). “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples”. *Biometrika* 25.3/4, pp. 285–294 (cit. on p. 25).

Van Der Walt, Stefan, S Chris Colbert, and Gael Varoquaux (2011). “The NumPy array: a structure for efficient numerical computation”. *Computing in science & engineering* 13.2, pp. 22–30 (cit. on p. 27).