MULTI-ARMED BANDITS FOR BUDGET-CONSTRAINED DATA COLLECTION

Apoorva Lal, apoorval@stanford.edu

Motivation and contribution. Survey response rates have declined dramatically over the last thirty years. Despite innovation in how we can conduct a survey using various online mediums, response rates seldom exceed single digits and are particularly low among specific socio-economic groups, with dire consequences for policymaking, research, and polling. 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]. I 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. I provide simulation-based evidence that these algorithms improve upon current practice, and intend to evaluate their performance in field experiments in future work.

Methods. I assume that the researcher has can set K different levels ('arms') of survey payments in order to incentivise respondents[3], or K different experimental treatments with varying costs (e.g. Get-out-the-Vote treatments, which may vary in cost depending on the medium). 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 I call the reward. The researcher is therefore interested in *exploration* - learning the expected probability of response for each payment arm μ_a - but only inasmuch as it helps her in *exploitation* - setting the payment arm to maximise the expected reward (total data collected). Unlike with adaptive experiments, researcher isn't interested in precisely estimating the reward probabilities from each incentive level (their treatment effects), as it is well known that bandit algorithms perform poorly in terms of statistical power (Villar, Bowden, and Wason 2015), although amendments have been proposed (Kasy and Sautmann 2021; Offer-Westort, Coppock, and Green 2021).

A simple MAB characterisation abstracts from an important consideration in survey applications of MABs, where each arm costs a fixed amount c_a, 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 arm-specific costs. Under the plausible assumption that response rates are increasing in compensation[4], a conventional bandit-algorithm will therefore provide the trivial answer of 'pay everyone the largest possible amount' (pick $\arg\max_a c_a$) since the response probability is largest for these compensation levels, and consequently gather very little data before exhausting the budget. To accommodate the

[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), the survey literature provides little practical advice with regard to calibrating these incentives, possibly dynamically, as is proposed in the present paper

[3] Monetary incentives have been shown to improve response rates in a variety of survey settings (Singer and Ye 2013), and may be preferable to stratified sampling in settings plagued with non-response

[4] $\mu_k \leq \mu_{k+1}$ where the k arms are increasing in cost c_k

budget constraint, I amend well-known bandit algorithms - UCB and Thompson sampling - to maximise a cost-normalized upper confidence bound and posterior mean respectively, which results in them choosing actions with the largest reward/cost ratio. For UCB1, this implies choosing

$$A = \text{argmax}_{[K]} \frac{\widehat{\mu}_a + \sqrt{\frac{2 \log t}{n_a}}}{c_a}$$

where $\widehat{\mu}_a$ is the estimated mean for arm a, t is the total number of rounds played, and n_a is the total number of times arm a has been pulled; the maximand is therefore the upper-confidence bound normalized by cost. This approach was first proposed by Tran-Thanh et al. (2012), who call this approach the *fractional Knapsack-based Upper Confidence Bound Exploration and Exploitation* (KUBE) algorithm, and prove that it achieves logarithmic regret.

For Thompson sampling, we propose choosing the arm that maximises the ratio of the posterior reward probability[5] and normalized cost:

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

where θ_k is the posterior mean for the reward probability and $\widetilde{c}_a$ is a normalized cost that scales all arm costs to lie on the unit interval[6]. The Thompson sampling approach is particularly well-suited to this problem because, given uncertain and potentially long gaps between when surveys are dispatched and responses are received, researchers may use a 'batched' approach wherein posteriors distributions are updated at a fixed cadence (say, on a weekly basis). 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.

Finally, I 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. I propose an amended version of Thompson sampling that dynamically adjusts arm-specific costs to account for representativeness gaps. Specifically, I set the cost vector $\mathbf{c}_{at}^g$ to

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

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

[5] we use the simple Bernoulli bandit formulation in this setting. The reward is Bernoulli, and we use a beta-prior $\alpha_a = 1, \beta_a = 1$ for all arms so as to sample from a beta posterior

[6] this is one of many possible parametrisations of the optimization problem. In current work, I am developing an information-theoretic objective function for the bandit algorithm, as well as a complementary dynamic-programming based finite sample solution

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), and $\mathbf{c}_a$ 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. $\gamma \geq 0$ is a choice parameter that represents the tradeoff between sample-size and representativeness, where larger values correspond to greater weight on representativeness at the cost of sample size (reward).

Preliminary Results. I first conduct a simulation study with 10 arms where costs are drawn uniformly from $\{2, 5, 10, 20\}$, and the corresponding means are simulated from a beta distribution[7] 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[8]. The cumulative reward is akin to the total number of survey responses since I model reward $= 1$ as a complete response. I let each algorithm run until it exhausts its budget, and average across 1000 runs. I benchmark the budgeted bandit algorithms against a standard set of well-known bandit algorithms: greedy, which pulls each arm a set number of times to learn means and then proceeds to pull the arm with the highest mean afterwards; random, which is a pure exploration algorithm that approximates random sampling in surveys; ε-greedy, which plays random with probability ε and greedy with complementary probability; $\varepsilon-$first, which uses the first $\varepsilon-$share of the budget to explore, and subsequently exploits; UCB1, which pulls $\arg\max_{[K]} \widehat{\mu}_a + \sqrt{2 \log t / n_a}$, and Thompson sampling, which pulls $\arg\max_{[K]} \theta_k$. I report the results in figure 1(a), and find that budgeted bandits (fracKUBE and thompsonBC) have much greater cumulative reward (i.e. they collect more survey responses) than best-performing non-budget algorithms.
Next, I perform a simulation study with two strata with different response propensities to evaluate the performance of dynamic cost-adjustment. I simulate data where there are two strata, E and F, and F responds to surveys at a substantially lower rate than E. I target equal shares of the two groups in our final sample because they are evenly distributed in the population. Reward probabilities are increasing in payment, where μ_a^E is generated by as before, 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. I report sample shares and sample sizes in figure 1(b), and find that the dynamic-cost adjustment approach to representative samples performs well along both axes - we get larger sample sizes as well as representativeness - relative to random and stratified sampling (which targets representativeness alone) and simple Thompson sampling ($\gamma = 0$, which target rewards – sample size – alone), which produces a sample comprised entirely of group E, which is highly unrepresentative. $\gamma > 0$ trades off sample size for more representative samples and, for sufficiently high values of γ, produces more representative samples than random sampling but with much larger sample sizes.

[7] Reward are drawn from $\mu_a \sim \text{Beta}\left(\alpha = \max\left(\frac{c_a}{5}, 1\right), \beta = 10/c_a\right)$
[8] For $c_a = 2, \mathbb{E}[\mu_a] = \frac{1}{5}$, while for $c_a = 20, \mathbb{E}[\mu_a] = \frac{8}{9}$.

Conclusion. This project proposes 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

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

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*.

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. URL: https://www.econometricsociety.org/doi/10.3982/ECTA17527.

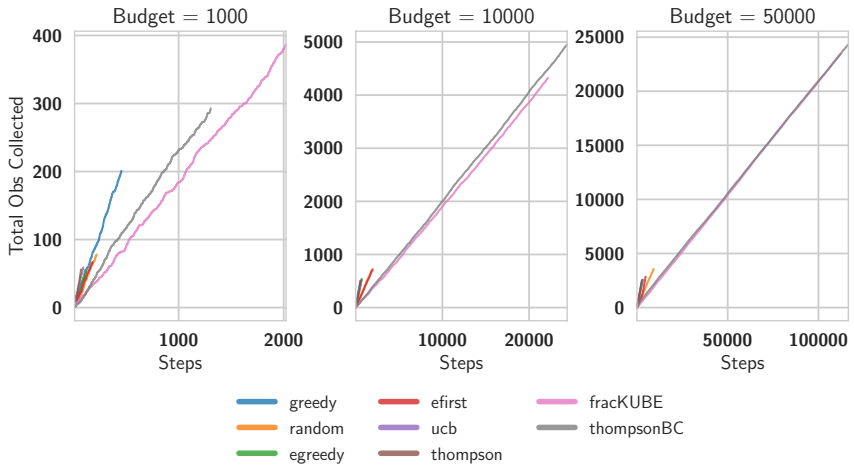
Offer-Westort, Molly, Alexander Coppock, and Donald P Green (2021). "Adaptive Experimental Design: Prospects and Applications in Political Science". *American Journal of Political Science*.

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. URL: https://doi.org/10.1177/0002716212458082.

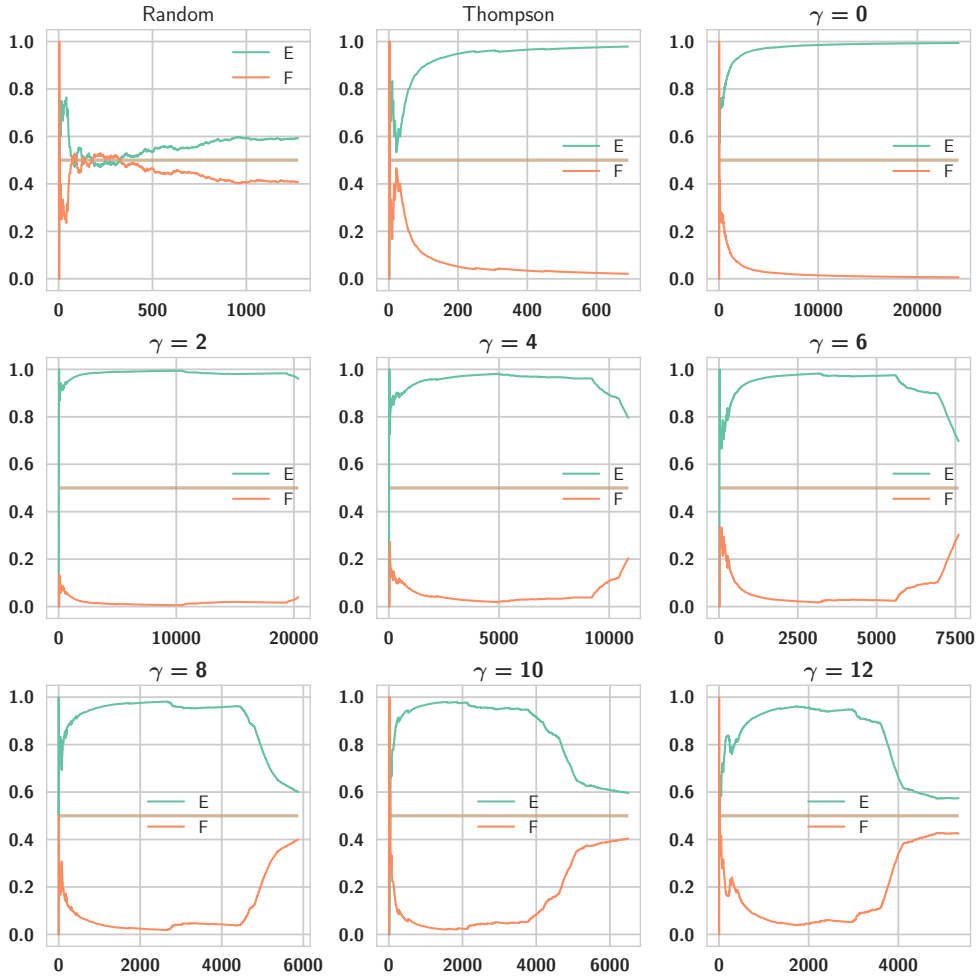
Tran-Thanh, Long et al. (Apr. 2012). "Knapsack based optimal policies for budget-limited multi-armed bandits". arXiv: 1204.1909 [cs.AI].

Villar, Sofía S, Jack Bowden, and James Wason (2015). "Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges". en. *Statistical science: a review journal of the Institute of Mathematical Statistics* 30.2, pp. 199–215. URL: http://dx.doi.org/10.1214/14-STS504.

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*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3136245.



(a) Cumulative rewards (total sample sizes) for budgeted bandits.



(b) Shares of group E and F over time for different bandit algorithms, with the bottom six varying the degree of 'representativeness prioritisation' γ. The x-axis values indicate the total number of observations collected by each method.