Bandit Algorithms for Data Collection

Apoorva Lal

October 31, 2021

Stanford University

Introduction

- Survey response rates have plummeted over the last thirty years, with dire consequences for policymaking, research, and polling.
  - Extensive work has been done on re-weighting and imputation methods to adjust for non-response *ex-post*
  - Less attention has been paid to the design and allocation of survey incentives *ex-ante* to increase response rates
- This paper: choosing incentive-levels in surveys as an online learning problem
  - 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
  - proposes budget-constrained multi-armed bandits to learn and use these WTA values to increase response rates subject to budget-constraints and representativeness considerations.

Setup

- binary rewards $r \in \{0, 1\}$, and $a \in \mathcal{A} := [K]$ 'arms' (treatment arms) with an unknown probability of success $\mu_1, \ldots, \mu_k \ \in [0, 1]$
- Pulling the ath arm produces reward r_a sampled from Bernoulli distribution $\mathbb{P}_a$ with mean μ_a. The agent's task is to maximise total reward $\mathbb{E}\left[\sum_{t=1}^{T} r_{at}\right]$.
- If we knew $\mu_1, \ldots, \mu_K$, the optimal action would simply be to always play the arm with the highest reward $a^* = \arg\max_{[K]} \mu_k$.
    - However, we don't, and therefore we need to incorporate learning μs into the problem. This is the *exploration versus exploitation* tradeoff.
    - Reward maximisation is equivalent to minimising regret $\mathbb{E}\left[\text{Regret}\right] = \mathbb{E}\left[\sum_{t=1}^{T} r_{a^*,t} - r_{at}\right]$
    - lower bounds on the regret for any 'consistent' algorithm is logarithmic in the number of pulls t (Lai and Robbins 1985)
- empirical mean $Q_a := \frac{\text{Sum of rewards received from arm } a}{\text{Number of times arm } a \text{ was pulled}}$ is unbiased for μ_a, so we update it every time arm a is pulled

Thompson Sampling

$$\overbrace{\pi\left(\mu_a \mid \mathcal{D}\right)}^{\text{Posterior}} \propto \pi\left(\mu_a\right)\pi\left(\mathcal{D} \mid \mu_a\right) \propto \underbrace{\mu_a^{1-1}\left(1-\mu_a\right)^{1-1}}_{\text{Prior}} \overbrace{\mu_a^{s_a}\left(1-\mu_a\right)^{f_a}}^{\text{Likelihood}}$$

$$\propto \mu_a^{1-1+s_a}\left(1-\mu_a\right)^{1-1+f_a} \propto \mu_a^{s_a}\left(1-\mu_a\right)^{f_a}$$

Parameter: $\mathbf{S}, \mathbf{F} = 0$ Success and failure counters for each arm)

for *t = 1, ..., T* **do**

  for *a = 1, ..., K* **do**

    Draw $\mu_a \sim \text{Beta}\left(S_a + 1, F_a + 1\right)$; `// 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

Adding arm-specific costs and budget constraints

- Vanilla bandits: agent's goal is to maximise the expected cumulative reward from the sequence of pulls at T ($\to\infty$).
- MABs may face budget constraints in real-world applications
  - Pulling each arms may be associated with a fixed (Tran-Thanh et al. 2012) (henceforth TCRJ) or random (Ding et al. 2013) cost
- In surveys, an 'arm' is a monetary reward for survey completion, we necessarily have fixed costs to pulling each arm, and a finite budget.
  - Can offer payment conditional on completion (pay c_a only if reward is 1)
- under the reasonable assumption that larger payments are more likely to induce responses, we may have $\mu_1 \leq \ldots \mu_K$ where $\{1, \ldots, K\}$ are ordered by the monetary value of the arm WLOG.
- A conventional MAB might give us a trivial answer: pay everyone the most (i.e.~pull arm K with the maximum value).

Knapsack-based UCB

- 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.
- Budget constrained UCB (Fractional KUBE): Pull the arm that maximises the UCB/cost ratio

$$A = \arg\max_{[K]} \frac{\overbrace{Q_a + \sqrt{\frac{2\log t}{n_a}}}^{\text{UCB}}}{\underbrace{c_a}_{\text{cost}}} \tag{1}$$

Budgeted Thompson sampling

- Draw from posterior reward probability, but max reward/cost ratio Pull

$$\arg\max_{[K]} \mu_a / \left(c_{a,t} / \sum_a c_a \right)$$

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 $\widehat{\mu}_a \sim$ 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

Representativeness through cost-adjustment

- Inducing representativeness by adjusting costs $c_{a,t}$
  - initially prioritise exploration (keep $\mathbf{c}$ low)
  - later balance / representativeness $c_a \propto (\overline{x}_n - \widetilde{x})$
- vary c_a dynamically to target balance (in experiments) or representativeness (in surveys)

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

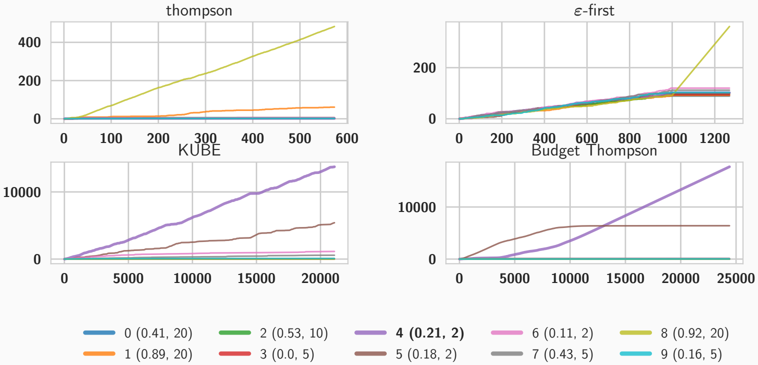
where $\psi^g := (\overline{x}_t - \widetilde{x})$ is current over-representation of group g in sample

- Alternative: incorporate representativeness directly into objective function and solve dynamic program

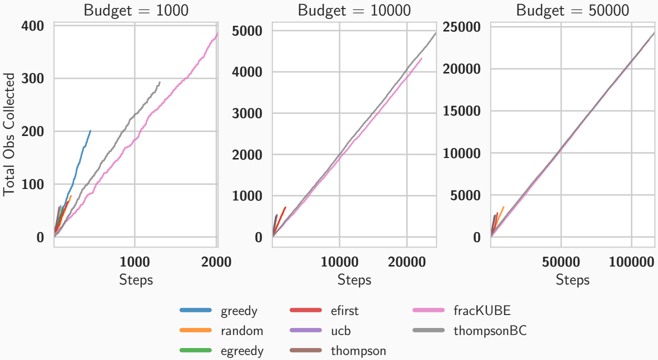
Setup

- Costs are drawn from a discrete uniform $c_a \sim \{2, 5, 10, 20\}$. 10 arms.
- The corresponding mean rewards are simulated $\mu_a \sim \text{Beta}\left(\alpha = \max(c_a/5, 1), \beta = 10/c_a\right)$
    - This ensures that the reward probability, $\mathbb{E}\left[\mu_a\right] = \frac{\alpha}{\alpha+\beta}$, is increasing in c_a, which is based on our substantive assumption that higher payments are more likely to elicit responses
        - For $c_a = 2, \mathbb{E}\left[\mu_a\right] = \frac{1}{5}$, while for $c_a = 20, \mathbb{E}\left[\mu_a\right] = \frac{4}{4.5} = \frac{8}{9}$.

Arm-pulls with budget constraints (cost conditional on reward)

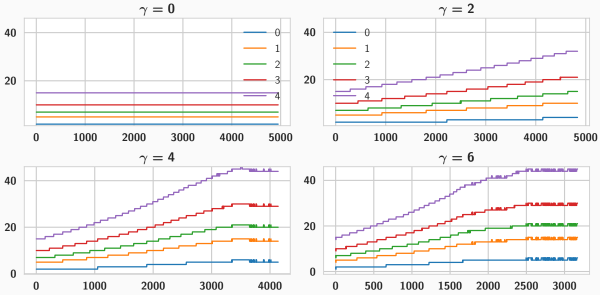


Cumulative Payoffs with budget constraints (cost conditional on reward)

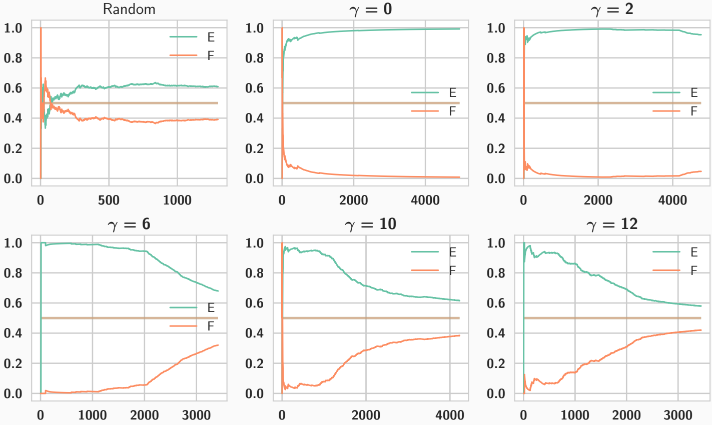


Altering costs to improve representativeness

- Two groups: E, F, with μ_a^E generated as before, and $\mu_a^F = (0.4, 0.5, 0.6, 0.7, 0.8) \cdot \mu_a^E$
- Target in survey: $50\%, 50\%$ groups E and F
- Costs for group E increase over time



Sample shares of groups in simulations



Conclusion

- Data collection using different strategies can be framed as a bandit problem
- However, conventional bandits abstract from arm-specific costs
    - severely limits their applicability in many social-scientific settings
- I propose cost-normalised UCB and Thompson, which work well in such settings
    - can be calibrated to prioritise representativeness later in data collection

Future work

- formal results for dynamic-cost adjustment setup
- formalism as a dynamic programming problem with information-theoretic objective function

Thanks!

Ding, Wenkui et al. (2013). “Multi-armed bandit with budget constraint and variable costs”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 27. 1.

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

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