Adaptive treatment assignment in experiments for policy choice

Maximilian Kasy   Anja Sautmann

October 1, 2019

Introduction

The goal of many experiments is to inform policy choices:

1. **Job search assistance** for refugees:
    - Treatments: Information, incentives, counseling, ...
    - Goal: Find a policy that helps as many refugees as possible to find a job.
2. **Clinical trials**:
    - Treatments: Alternative drugs, surgery, ...
    - Goal: Find the treatment that maximize the survival rate of patients.
3. Online **A/B testing**:
    - Treatments: Website layout, design, search filtering, ...
    - Goal: Find the design that maximizes purchases or clicks.
4. Testing **product design**:
    - Treatments: Various alternative designs of a product.
    - Goal: Find the best design in terms of user willingness to pay.

Example

- There are 3 treatments d.
- $d = 1$ is best, $d = 2$ is a close second, $d = 3$ is clearly worse. (But we don't know that beforehand.)
- You can potentially run the experiment in 2 waves.
- You have a fixed number of participants.
- After the experiment, you pick the best performing treatment for large scale implementation.

How should you design this experiment?

1. Conventional approach.
2. Bandit approach.
3. Our approach.

Conventional approach

Split the sample equally between the 3 treatments,
to get precise estimates for each treatment.

- After the experiment, it might still be hard to distinguish whether treatment 1 is best, or treatment 2.
- You might wish you had not wasted a third of your observations on treatment 3, which is clearly worse.

The conventional approach is

1. good if your goal is to get a precise estimate for each treatment.
2. not optimal if your goal is to figure out the best treatment.

Bandit approach

Run the experiment in **2 waves**
split the first wave equally between the 3 treatments.
Assign **everyone** in the second (last) wave to
the **best performing treatment** from the first wave.

- After the experiment, you have a lot of information on the d that performed best in wave 1, probably $d = 1$ or $d = 2$,
- but much less on the other one of these two.
- It would be better if you had split observations equally between 1 and 2.

The bandit approach is

1. good if your goal is to maximize the outcomes of participants.
2. not optimal if your goal is to pick the best policy.

Our approach

Run the experiment in **2 waves**
split the first wave equally between the 3 treatments.
Split the second wave between
the **two best performing** treatments from the first wave.

- After the experiment you have the maximum amount of information to pick the best policy.

Our approach is

1. good if your goal is to pick the best policy,
2. not optimal if your goal is to estimate the effect of all treatments, or to maximize the outcomes of participants.

Let θ^d denote the average outcome
that would prevail if everybody was assigned to treatment d.

What is the objective of your experiment?

1. Getting precise treatment effect estimators, powerful tests:

   $$\text{minimize} \sum_d (\hat{\theta}^d - \theta^d)^2$$

   ⇒ Standard experimental design recommendations.

2. Maximizing the outcomes of experimental participants:

   $$\text{maximize} \sum_i \theta^{D_i}$$

   ⇒ Multi-armed bandit problems.

3. Picking a welfare maximizing policy after the experiment:

   $$\text{maximize } \theta^{d^*},$$

   where d^* is chosen after the experiment.
   ⇒ This talk.

Preview of findings

- **Optimal** adaptive **designs** improve expected welfare.
- Features of optimal treatment assignment:
  - Shift toward better performing treatments over time.
  - But don't shift as much as for Bandit problems:
    We have no "exploitation" motive!
- Fully optimal assignment is computationally challenging in large samples.
- We propose a simple **exploration sampling** algorithm.
  - Prove theoretically that it is rate-optimal for our problem.
  - Show that it dominates alternatives in calibrated simulations.

Literature

- Adaptive designs in clinical trials:
  - Berry (2006).
- Bandit problems:
  - Gittins index (optimal solution to some bandit problems): Weber et al. (1992).
  - Regret bounds for bandit problems: Bubeck and Cesa-Bianchi (2012).
  - Thompson sampling: Russo et al. (2018).
- Reinforcement learning:
  - Ghavamzadeh et al. (2015),
  - Sutton and Barto (2018).
- Best arm identification:
  - Russo (2016).
    Key reference for our theory results.
- Empirical examples for our simulations:
  - Ashraf et al. (2010),
  - Bryan et al. (2014),
  - Cohen et al. (2015).

Setup and optimal treatment assignment

Exploration sampling

Theoretical analysis

Calibrated simulations

Implementation in the field

Covariates and targeting

Setup

- Waves $t = 1, \ldots, T$, sample sizes N_t.
- Treatment $D \in \{1, \ldots, k\}$, outcomes $Y \in \{0, 1\}$.
- Potential outcomes Y^d.
- Repeated cross-sections:
  $(Y^0_{it}, \ldots, Y^k_{it})$ are i.i.d. across both i and t.
- Average potential outcome:
  $$\theta^d = E[Y^d_{it}].$$
- Key choice variable:
  Number of units n^d_t assigned to $D = d$ in wave t.
- Outcomes:
  Number of units s^d_t having a "success" (outcome $Y = 1$).

Design objective and Bayesian prior

- **Policy objective** $\theta^d - c^d$.
  - where d is chosen after the experiment,
  - and c^d is the unit cost of implementing policy d.
- **Prior**
  - $\theta^d \sim Beta(\alpha_0^d, \beta_0^d)$, independent across d.
  - Posterior after period t: $\theta^d | \boldsymbol{m}_t, \boldsymbol{r}_t \sim Beta(\alpha_t^d, \beta_t^d)$
- **Posterior expected social welfare**
  as a function of d:

$$\begin{aligned} SW(d) &= E[\theta^d | \boldsymbol{m}_T, \boldsymbol{r}_T] - c^d \\ &= \frac{\alpha_T^d}{\alpha_T^d + \beta_T^d} - c^d. \end{aligned}$$

Optimal assignment: Dynamic optimization problem

- Solve for the optimal experimental design using backward induction.
- Denote by V_t the value function after completion of wave t.
- Starting at the end, we have

$$V_T(\boldsymbol{m}_T, \boldsymbol{r}_T) = \max_d \left(\frac{\alpha_0^d + r_T^d}{\alpha_0^d + \beta_0^d + m_T^d} - c^d \right).$$

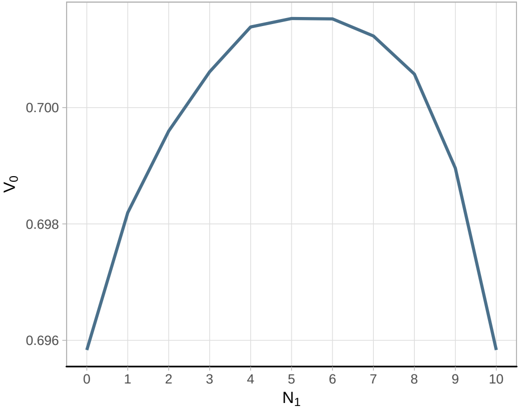
- Finite state and action space.
  $\Rightarrow$ Can, in principle, solve directly for optimal rule using dynamic programming: Complete enumeration of states and actions.

Simple examples

- Consider a small experiment
  with 2 waves, 3 treatment values (minimal interesting case).
- The following slides plot expected welfare
  as a function of:
  1. **Division of sample** size between waves, $N_1 + N_2 = 10$.
     $N_1 = 6$ is optimal.
  2. **Treatment assignment** in wave 2, given wave 1 outcomes.
     $N_1 = 6$ units in wave 1, $N_2 = 4$ units in wave 2.

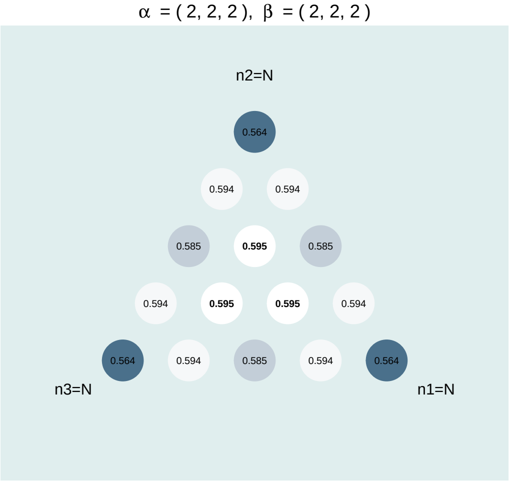
Dividing sample size between waves

- $N_1 + N_2 = 10$.
- Expected welfare as a function of N_1.
- Boundary points ≈ 1-wave experiment.
- $N_1 = 6$ (or 5) is optimal.



Expected welfare, depending on 2nd wave assignment

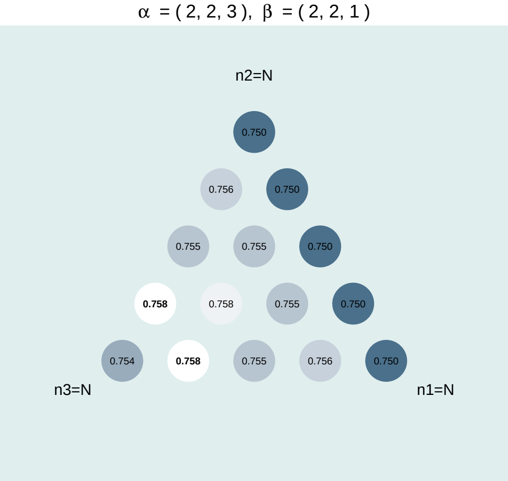
After one success, one failure for each treatment.



Light colors represent higher expected welfare.

Expected welfare, depending on 2nd wave assignment

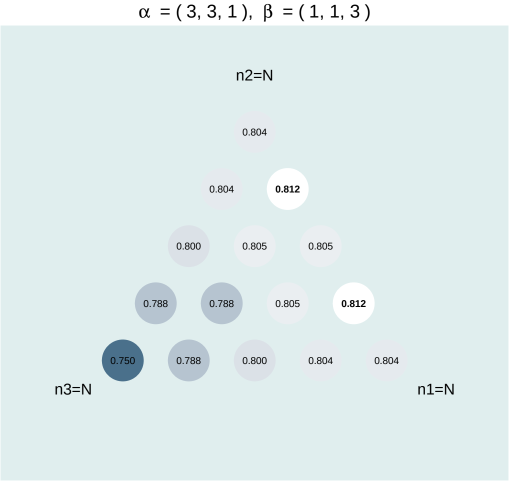
After one success in treatment 1 and 2, two successes in 3



Light colors represent higher expected welfare.

Expected welfare, depending on 2nd wave assignment

After one success in treatment 1 and 2, no successes in 3.



Light colors represent higher expected welfare.

Computational complexity

- Most efficient dynamic programming approach: "Full memoization."
  - **Time complexity**:

$$\sum_{t=1}^{T-1} O\left((M_t N_{t+1})^{2k-1}\right) + O(M_T^{2k-1} k).$$

  - **Memory complexity**:

$$\sum_{t=1}^{T} O\left(M_t^{2k-1}\right).$$

- Sketch of calculation:
  - State space at end of wave t (values for (m_t, r_t)): $O(M_t^{2k-1})$.
  - Action space (values for n_{t+1}): $O(N_{t+1}^{k-1})$.
  - Possible transitions: $O(N_{t+1}^{k})$.
  - Computation time for V_t at a given state, and given $V_{t+1}(\cdot)$: $O(N_{t+1}^{2k-1})$.
- $\Rightarrow$ Computationally impractical.
- Simpler alternatives?

Setup and optimal treatment assignment

Exploration sampling

Theoretical analysis

Calibrated simulations

Implementation in the field

Covariates and targeting

Thompson sampling

- **Thompson sampling**
  - Old proposal by Thompson (1933).
  - Popular in online experimentation.
- Assign each treatment with probability equal to the posterior probability that it is optimal.

$$p_t^d = P\left(d = \underset{d'}{\operatorname{argmax}} \, (\theta^{d'} - c^{d'}) | \boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1} \right).$$

- Easily implemented: Sample draws $\widehat{\boldsymbol{\theta}}_{it}$ from the posterior, assign

$$D_{it} = \underset{d}{\operatorname{argmax}} \left(\hat{\theta}_{it}^d - c^d \right).$$

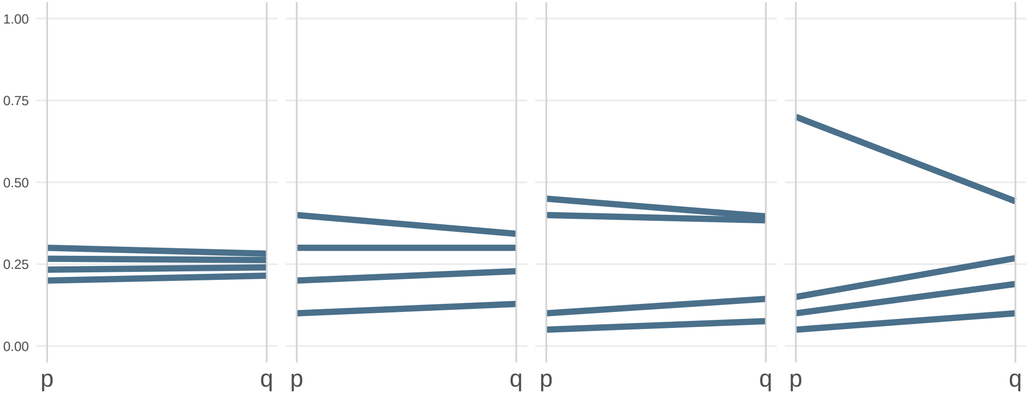
Exploration sampling

- Agrawal and Goyal (2012) proved that Thompson-sampling is rate-optimal for the multi-armed bandit problem.
- It is not for our policy choice problem!
- We propose two modifications:
  1. **Expected Thompson sampling**:
     Assign non-random shares p_t^d of each wave to treatment d.
  2. **Exploration sampling**:
     Assign shares q_t^d of each wave to treatment d, where

$$
\begin{aligned}
q_t^d &= S_t \cdot p_t^d \cdot (1 - p_t^d), \\
S_t &= \frac{1}{\sum_d p_t^d \cdot (1 - p_t^d)}.
\end{aligned}
$$

- These modifications
  1. yield rate-optimality (theorem coming up), and
  2. improve performance in our simulations.

Illustration of the mapping from Thompson to exploration sampling



Setup and optimal treatment assignment

Exploration sampling

Theoretical analysis

Calibrated simulations

Implementation in the field

Covariates and targeting

Theoretical analysis

Thompson sampling – results from the literature

- **In-sample regret** (bandit objective):
  $\sum_{t=1}^{T} \Delta^d$, where $\Delta^d = \max_{d'} \theta^{d'} - \theta^d$.
- Agrawal and Goyal (2012) (Theorem 2): For Thompson sampling,

$$\lim_{T \to \infty} E\left[\frac{\sum_{t=1}^{T} \Delta^d}{\log T} \right] \leq \left(\sum_{d \neq d^*} \frac{1}{(\Delta^d)^2} \right)^2 .$$

- Lai and Robbins (1985):
  No adaptive experimental design can do better than this $\log T$ rate.
- Thompson sampling only assigns a share of units of order $\log(M)/M$ to treatments other than the optimal treatment.

Results from the literature continued

- This is good for in-sample welfare, bad for learning:
  We stop learning about suboptimal treatments very quickly.
- Bubeck et al. (2011) Theorem 1 implies:
  Any algorithm that achieves $\log(M)/M$ rate for in-sample regret
  (such as Thompson sampling)
  can at most achieve **polynomial rate** for our objective Δ^{d^*}.
- By contrast (easy to show): Any algorithm that assigns shares
  converging to non-zero shares for each treatment
  achieves **exponential rate** for our objective.
- Our result (next slide): Exploration sampling achieves the
  (constrained) best exponential rate.

Exploration sampling

Proposition

Assume fixed wave size $N_t = N$.
As $T \to \infty$, exploration sampling satisfies:

1. *The share of observations assigned to the best treatment converges to $1/2$.*
2. *All the other treatments d are assigned to a share of the sample which converges to a non-random share $\bar{q}^d$. $\bar{q}^d$ is such that the posterior probability of d being optimal goes to 0 at the same exponential rate for all sub-optimal treatments.*
3. *No other assignment algorithm for which statement 1 holds has average regret going to 0 at a faster rate than exploration sampling.*

Sketch of proof

Our proof draws heavily on Russo (2016). Proof steps:

1. Each treatment is assigned infinitely often.
   $\Rightarrow p_T^d$ goes to 1 for the optimal treatment and to 0 for all other treatments.
2. Claim 1 then follows from the definition of exploration sampling.
3. Claim 2: Suppose p_t^d goes to 0 at a faster rate for some d.
   Then exploration sampling stops assigning this d.
   This allows the other treatments to "catch up."
4. Claim 3: Balancing the rate of convergence implies efficiency.
   This follows from an efficiency bound for best-arm-selection in Russo (2016).

Calibrated simulations

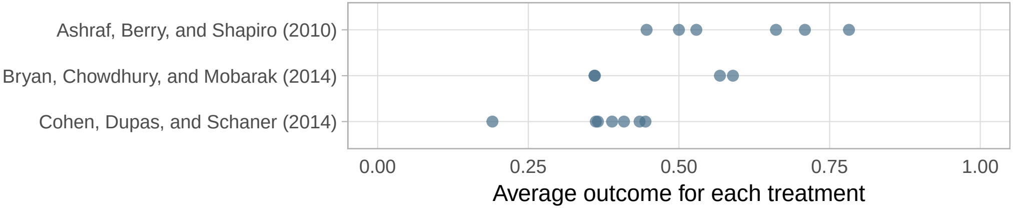
- Simulate data calibrated to estimates of 3 published experiments.
- Set $\boldsymbol{\theta}$ equal to observed average outcomes for each stratum and treatment.
- Total sample size same as original.

Ashraf, N., Berry, J., and Shapiro, J. M. (2010). Can higher prices stimulate product use? Evidence from a field experiment in Zambia.
American Economic Review, 100(5):2383–2413

Bryan, G., Chowdhury, S., and Mobarak, A. M. (2014). Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh.
Econometrica, 82(5):1671–1748

Cohen, J., Dupas, P., and Schaner, S. (2015). Price subsidies, diagnostic tests, and targeting of malaria treatment: evidence from a randomized controlled trial.
American Economic Review, 105(2):609–45

Calibrated parameter values



- Ashraf et al. (2010): 6 treatments, evenly spaced.
- Bryan et al. (2014): 2 close good treatments, 2 worse treatments (overlap in picture).
- Cohen et al. (2015): 7 treatments, closer than for first example.

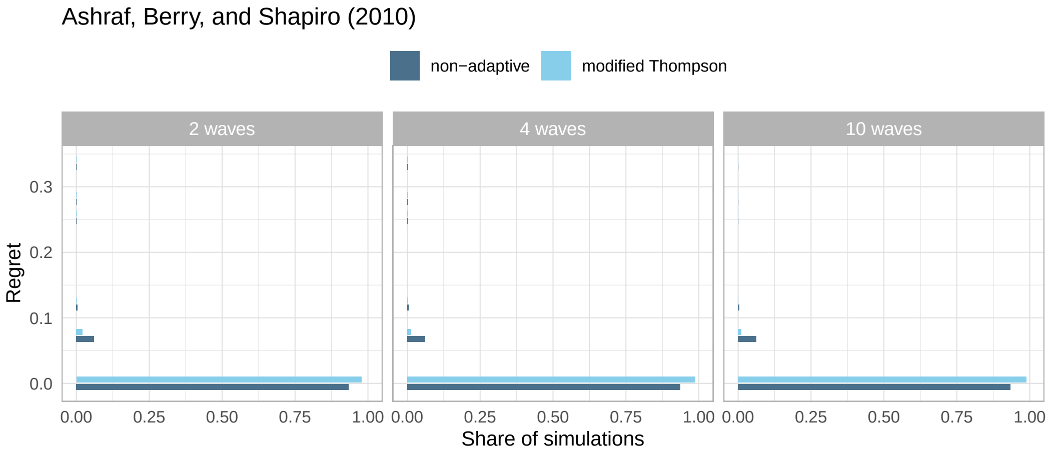
Summary of simulation findings

- With two waves, relative to non-adaptive assignment:
  - Thompson reduces average regret by 15-58 %,
  - exploration sampling by 21-67 %.
- Similar pattern for the probability of choosing the optimal treatment.
- Gains increase with the number of waves, given total sample size.
  - Up to 85% for exploration sampling with 10 waves for Ashraf et al. (2010).
- Gains largest for Ashraf et al. (2010),
  followed by Cohen et al. (2015),
  and smallest for Bryan et al. (2014).

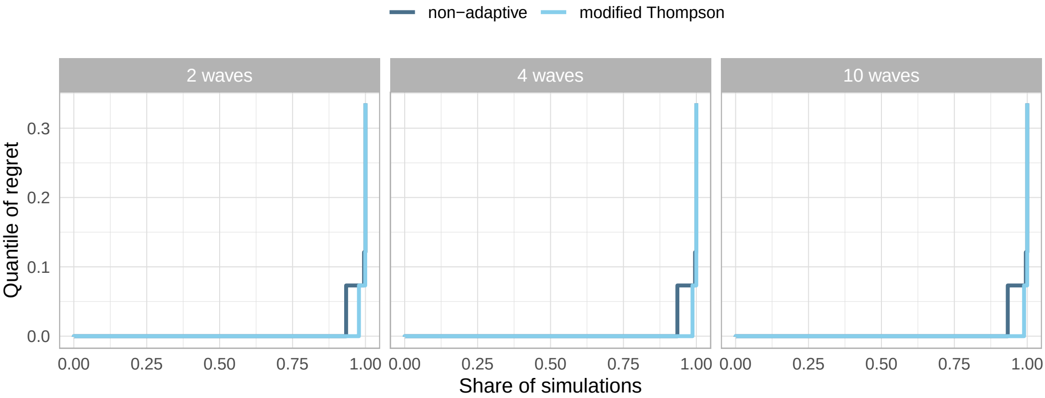
Plots of simulation results

- Compare exploration sampling to non-adaptive assignment.
- Full distribution of regret.
  (Difference between $\max_d \theta^d$ and θ^{d^*} for the d^* chosen after the experiment.)
- 2 representations:
  1. Histograms
     Share of simulations with any given value of regret.
  2. Quantile functions
     (Inverse of) integrated histogram.
- Histogram bar at 0 regret equals share optimal.
- Integrated difference between quantile functions is
  difference in average regret.
- Uniformly lower quantile function means
  1st-order dominated distribution of regret.

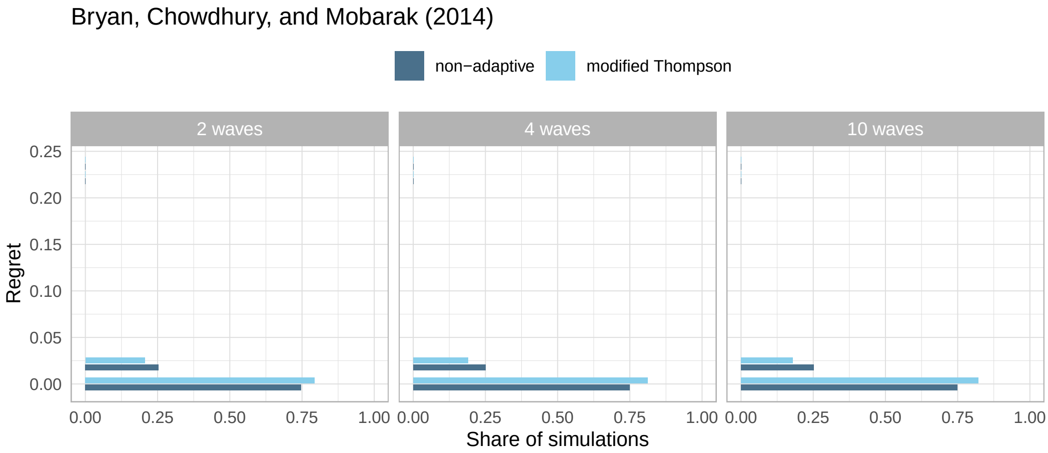
Policy Choice and Regret Distribution



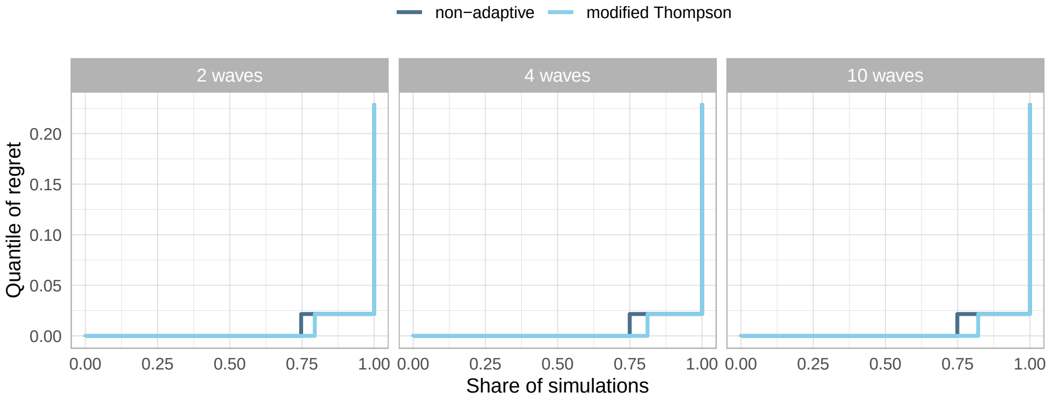
Policy Choice and Regret Distribution



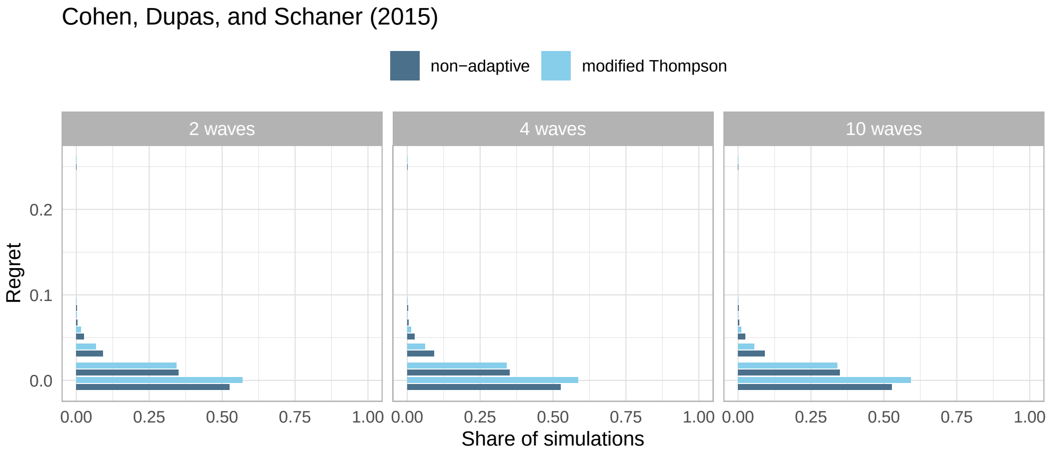
Policy Choice and Regret Distribution



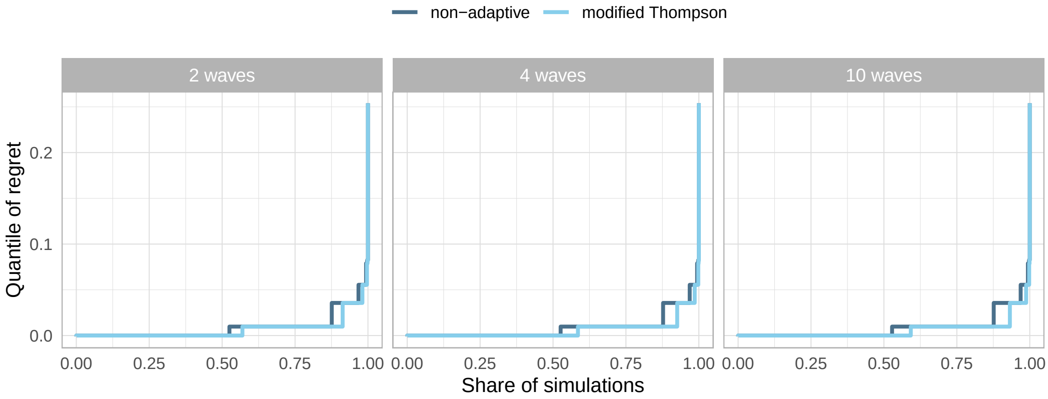
Policy Choice and Regret Distribution



Policy Choice and Regret Distribution

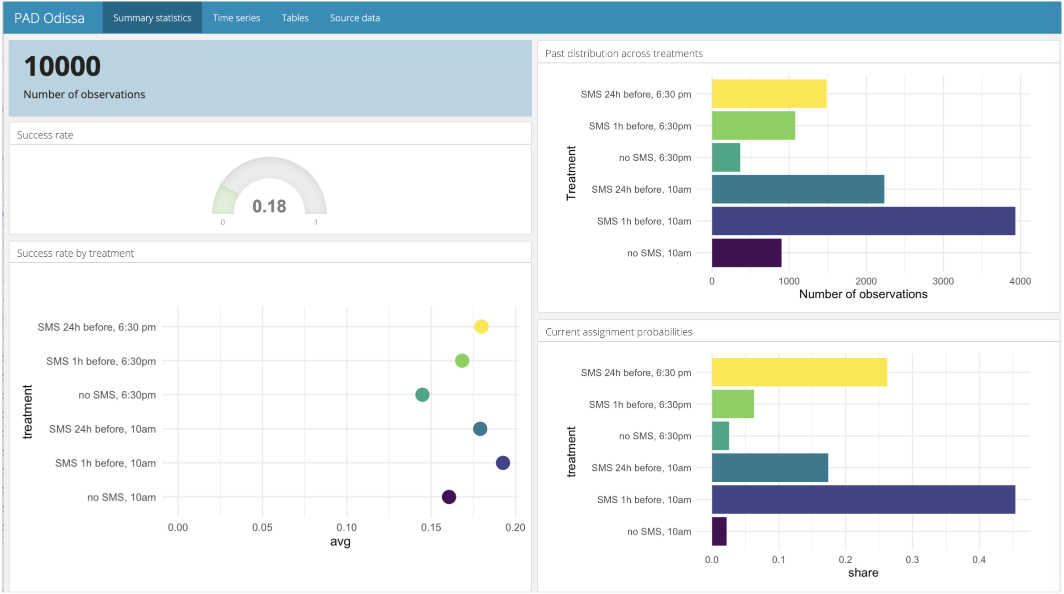


Policy Choice and Regret Distribution



Implementation in the field

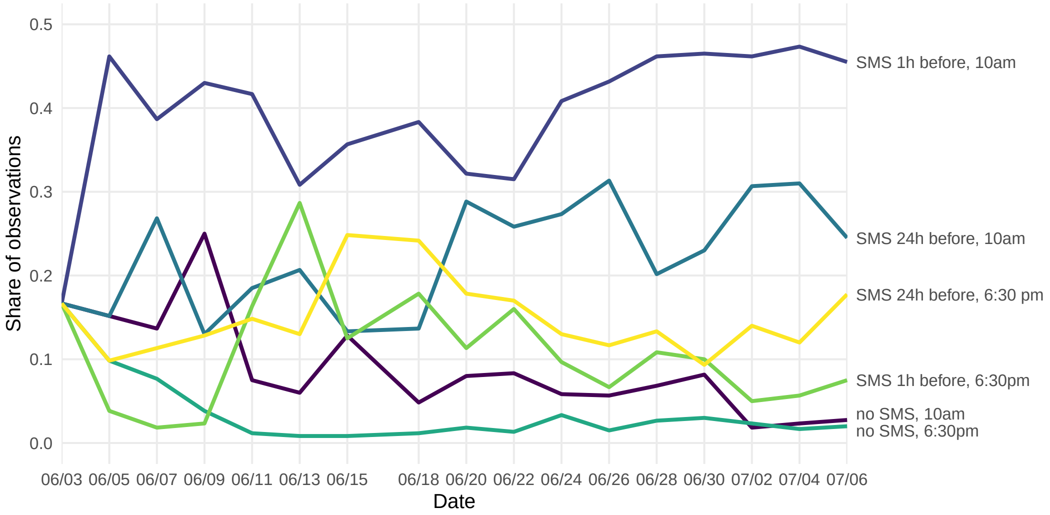
- NGO Precision Agriculture for Development (PAD) and Government of Odisha, India.
- Enrolling rice farmers into customized advice service by mobile phone.
- Waves of 600 farmers called through automated service; total of 10K calls.
- Outcome: did the respondent answer the enrollment questions?


PAD Odissa
Summary statistics
Time series
Tables
Source data
10000
Number of observations
Success rate
0.18
0
1
Success rate by treatment
SMS 24h before, 6:30 pm
SMS 1h before, 6:30pm
no SMS, 6:30pm
SMS 24h before, 10am
SMS 1h before, 10am
no SMS, 10am
treatment
0.00
0.05
0.10
0.15
0.20
avg
Past distribution across treatments
Treatment
0
1000
2000
3000
4000
Number of observations
Current assignment probabilities
0.0
0.1
0.2
0.3
0.4
share

Posterior parameters

Treatment	Mean	St. dev.	Probability optimal
No SMS, 10am call	0.161	0.012	0.009
SMS 1h before, 10am call	0.193	0.006	0.754
SMS 24h before, 10am call	0.179	0.008	0.073
No SMS, 6:30pm call	0.147	0.018	0.011
SMS 1h before, 6:30pm call	0.169	0.011	0.027
SMS 24h before, 6:30 pm call	0.180	0.010	0.126

Assignment shares over time



Setup and optimal treatment assignment

Exploration sampling

Theoretical analysis

Calibrated simulations

Implementation in the field

Covariates and targeting

Extension: Covariates and treatment targeting

- Suppose now that
  1. We additionally observe a (discrete) covariate X.
  2. The policy to be chosen can **target treatment** by X.
- How to adapt exploration sampling sampling to this setting?
- Solution: Hierarchical Bayes model,
  to optimally combine information across strata.
- Example of a **hierarchical Bayes** model:

$$\begin{aligned} Y^d|X=x,\theta^{dx},(\alpha_0^d,\beta_0^d) &\sim Ber(\theta^{dx}) \\ \theta^{dx}|(\alpha_0^d,\beta_0^d) &\sim Beta(\alpha_0^d,\beta_0^d) \\ (\alpha_0^d,\beta_0^d) &\sim \pi, \end{aligned}$$

- No closed form posterior, but can use Markov Chain Monte Carlo to sample from posterior.

MCMC sampling from the posterior

Combining Gibbs sampling & Metropolis-Hasting

- Iterate across replication draws ρ:
  1. **Gibbs** step: Given $\boldsymbol{\alpha}_{\rho-1}$ and $\boldsymbol{\beta}_{\rho-1}$,
     - draw $\theta^{dx} \sim Beta(\alpha^d_{\rho-1} + s^{dx}, \beta^d_{\rho-1} + m^{dx} - s^{dx})$.
  2. **Metropolis** step: Given $\boldsymbol{\beta}_{\rho-1}$ and $\boldsymbol{\theta}_\rho$,
     - draw $\alpha^d_\rho \sim$ (symmetric proposal distribution).
     - Accept if an independent uniform is less than the ratio of the posterior for the new draw, relative to the posterior for $\alpha^d_{\rho-1}$.
     - Otherwise set $\alpha^d_\rho = \alpha^d_{\rho-1}$.
  3. **Metropolis** step: Given $\boldsymbol{\theta}_\rho$ and $\boldsymbol{\alpha}_\rho$,
     - proceed as in 2, for β^d_ρ.
- This converges to a stationary distribution such that

$$P\left(d = \underset{d'}{\operatorname{argmax}}\ \theta^{d'x} | \boldsymbol{m}_t, \boldsymbol{r}_t\right) = \underset{R\to\infty}{\operatorname{plim}} \frac{1}{R} \sum_{\rho=1}^{R} \mathbf{1}\left(d = \underset{d'}{\operatorname{argmax}}\ \theta^{d'x}_\rho\right).$$

Conclusion

- Different objectives lead to different optimal designs:
    1. Treatment effect estimation / testing: Conventional designs.
    2. In-sample regret: Bandit algorithms.
    3. Post-experimental policy choice: This talk.
- If the experiment can be implemented in multiple waves, adaptive designs for policy choice
    1. significantly increase welfare,
    2. by focusing attention in later waves
       on the best performing policy options,
    3. but not as much as bandit algorithms.
- Implementation of our proposed procedure is easy and fast,
  and easily adapted to new settings:
    - Hierarchical priors,
    - non-binary outcomes...
- Interactive dashboard for treatment assignment:
  `https://maxkasy.shinyapps.io/exploration_sampling_dashboard/`

Thank you!