Adaptive Experiments for Policy Choice

Maximilian Kasy   Anja Sautmann

March 8, 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:

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

   $\Rightarrow$ Standard experimental design recommendations.

2. Maximizing the outcomes of experimental participants:

$$\text{maximize } \sum_{i} \theta^{D_i}$$

   $\Rightarrow$ Multi-armed bandit problems.

3. Picking a welfare maximizing policy after the experiment:

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

   where d^* is chosen after the experiment.
   $\Rightarrow$ 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 **modified Thompson** algorithm.
  - Show that it dominates alternatives in calibrated simulations.
  - Prove theoretically that it is rate-optimal for our problem.

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

- 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:
  $(Y_{it}^0, \ldots, Y_{it}^k)$ are i.i.d. across both i and t.
- Average potential outcome:

$$\theta^d = E[Y_{it}^d].$$

- Key choice variable:
  Number of units n_t^d assigned to $D = d$ in wave t.
- Outcomes:
  Number of units s_t^d having a "success" (outcome $Y = 1$).

Treatment assignment, outcomes, state space

- Treatment assignment in wave t: $\boldsymbol{n}_t = (n_t^1, \ldots, n_t^k)$.
- Outcomes of wave t: $\boldsymbol{s}_t = (s_t^1, \ldots, s_t^k)$.
- Cumulative versions:

$$M_t = \sum_{t' \leq t} N_{t'}, \qquad \boldsymbol{m}_t = \sum_{t' \leq t} \boldsymbol{n}_t, \qquad \boldsymbol{r}_t = \sum_{t' \leq t} \boldsymbol{s}_t.$$

- Relevant information for the experimenter in period $t+1$ is summarized by $\boldsymbol{m}_t$ and $\boldsymbol{r}_t$.
- Total trials for each treatment, total successes.

Design objective

- Policy objective $SW(d)$:
  Average outcome Y, net of the cost of treatment.
- Choose treatment d after the experiment is completed.
- Posterior expected social welfare:

$$SW(d) = E[\theta^d | \boldsymbol{m}_T, \boldsymbol{r}_T] - c^d,$$

where c^d is the unit cost of implementing policy d.

Bayesian prior and posterior

- By definition, $Y^d|\theta \sim Ber(\theta^d)$.
- Prior: $\theta^d \sim Beta(\alpha_0^d, \beta_0^d)$, independent across d.
- Posterior after period t:

$$\theta^d|\boldsymbol{m}_t, \boldsymbol{r}_t \sim Beta(\alpha_t^d, \beta_t^d)$$
$$\alpha_t^d = \alpha_0^d + r_t^d$$
$$\beta_t^d = \beta_0^d + m_t^d - r_t^d.$$

- In particular,

$$SW(d) = \frac{\alpha_0^d + r_T^d}{\alpha_0^d + \beta_0^d + m_T^d} - c^d.$$

Optimal assignment: Dynamic optimization problem

- Dynamic stochastic optimization problem:
  - States $(\boldsymbol{m}_t, \boldsymbol{r}_t)$,
  - actions $\boldsymbol{n}_t$.
- 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.

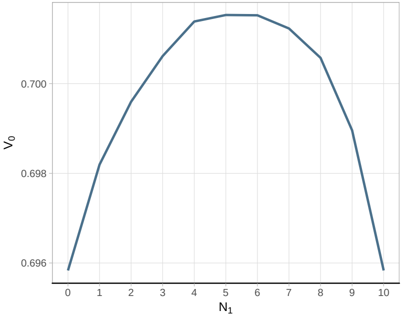
Simple examples

- 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.
- Keep in mind:

$$\alpha_1 = (1, 1, 1) + \boldsymbol{s}_1$$
$$\beta_1 = (1, 1, 1) + \boldsymbol{n}_1 - \boldsymbol{s}_1$$

Dividing sample size between waves

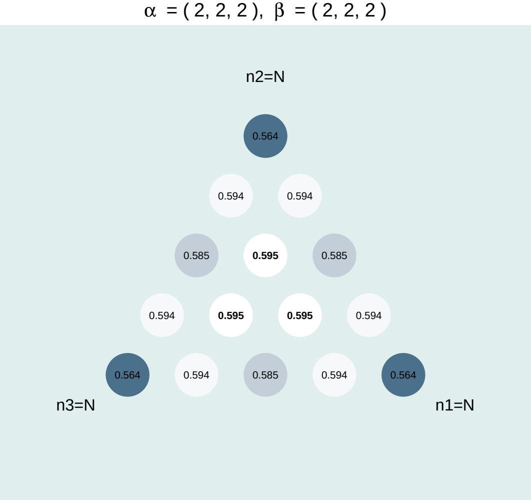
- $N_1 + N_2 = 10$.
- Expected welfare as a function of N_1.
- Boundary points $\approx$ 1-wave experiment.
- $N_1 = 6$ (or 5) is optimal.



- Next slides: Expected welfare as a function of wave 2 treatment assignment.

Expected welfare, depending on 2nd wave assignment

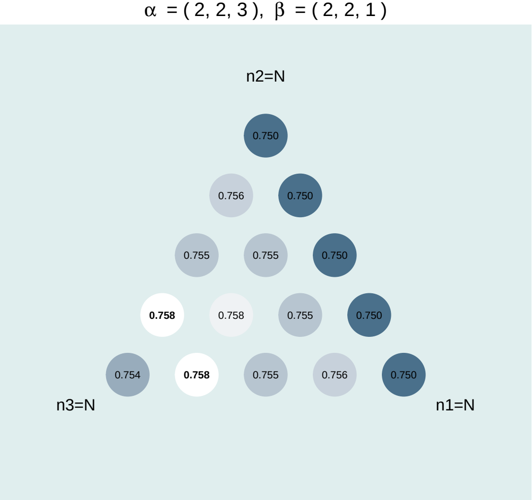
After one success, one failure for each treatment.



Light colors represent higher expected welfare.

Expected welfare, depending on 2nd wave assignment

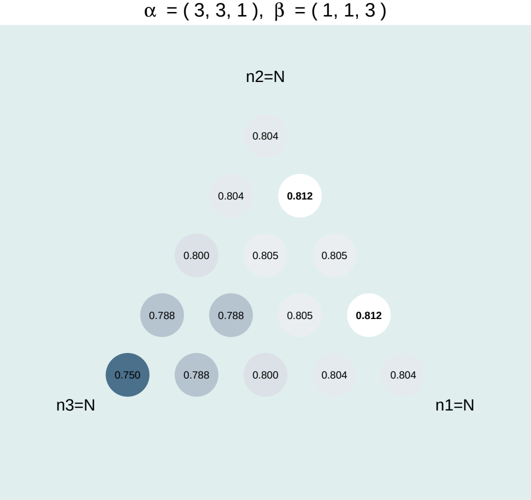
After one success in treatment 1 and 2, two successes in 3



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.

Thompson sampling

- Fully optimal solution is computationally impractical.
  Per wave, $O(N_t^{2k})$ combinations of actions and states.
  $\Rightarrow$ simpler alternatives?
- **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

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

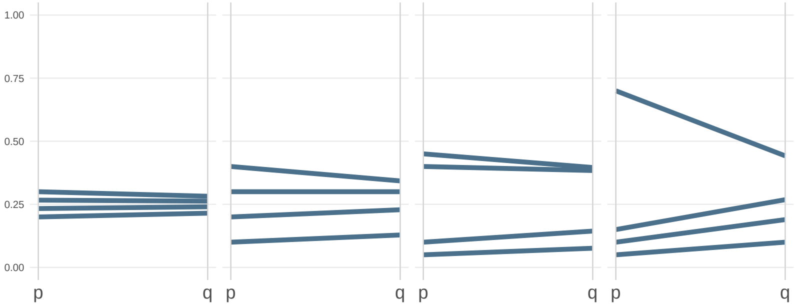
Modified Thompson sampling

- 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. **Modified Thompson sampling**:
     Assign shares q_t^d of each wave to treatment d, where

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

- These modifications
  1. Improve performance in our simulations.
  2. Will be theoretically motivated later in this talk.
     In particular, we will show (constrained) rate-optimality.

Illustration of the mapping from Thompson to modified Thompson



Calibrated simulations

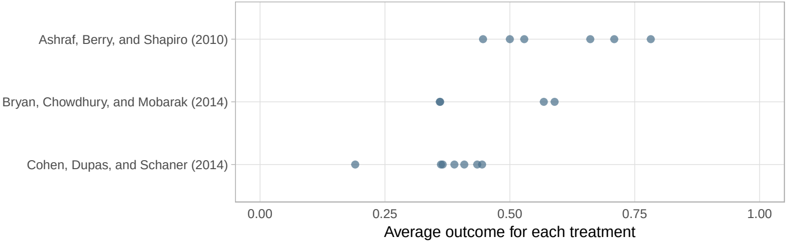
- Simulate data calibrated to estimates of 3 published experiments.
- Set θ equal to observed average outcomes for each stratum and treatment.
- Total sample size same as original.

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.

Coming up

- Compare 4 **assignment methods**:
  1. Non-adaptive (equal shares)
  2. Thompson
  3. Expected Thompson
  4. Modified Thompson
- Report 2 **statistics**:
  1. Average regret:
     Average difference, across simulations, between $\max_d \theta^d$ and θ^d for the d chosen after the experiment.
  2. Share optimal:
     Share of simulations for which the optimal d is chosen after the experiment (and thus regret equals 0).

Visual representations

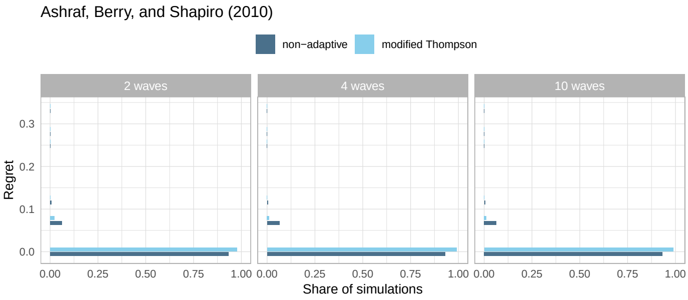
- Compare modified Thompson to non-adaptive assignment.
- Full distribution of regret.
- 2 representations:
    1. Histograms
       Share of simulations with any given value of regret.
    2. Quantile functions
       (Inverse of) integrated histogram.
- Histogram bar at 0 regret equals share optimal.
- Integrated difference between quantile functions is difference in average regret.
- Uniformly lower quantile function means 1st-order dominated distribution of regret.

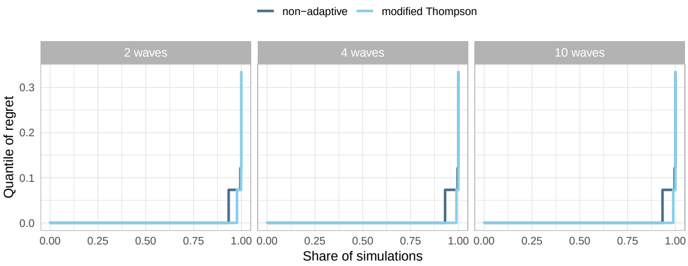
Regret and Share Optimal

Table: Ashraf, Berry, and Shapiro (2010)

Statistic	2 waves	4 waves	10 waves
Regret			
modified Thompson	0.002	0.001	0.001
expected Thompson	0.002	0.001	0.001
Thompson	0.002	0.001	0.001
non-adaptive	0.005	0.005	0.005
Share optimal			
modified Thompson	0.977	0.990	0.988
expected Thompson	0.970	0.981	0.983
Thompson	0.971	0.981	0.983
non-adaptive	0.933	0.930	0.932
Units per wave	502	251	100

Policy Choice and Regret Distribution



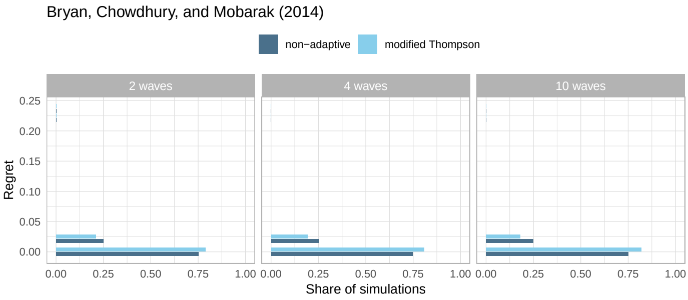


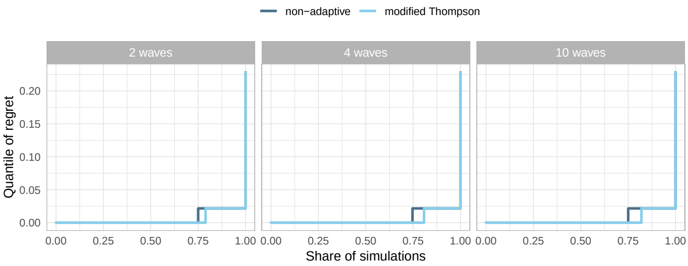
Regret and Share Optimal

Table: Bryan, Chowdhury, and Mobarak (2014)

Statistic	2 waves	4 waves	10 waves
Regret			
modified Thompson	0.005	0.004	0.004
expected Thompson	0.005	0.004	0.004
Thompson	0.005	0.004	0.004
non-adaptive	0.005	0.005	0.005
Share optimal			
modified Thompson	0.789	0.807	0.820
expected Thompson	0.784	0.800	0.804
Thompson	0.786	0.796	0.808
non-adaptive	0.750	0.747	0.750
Units per wave	935	467	187

Policy Choice and Regret Distribution



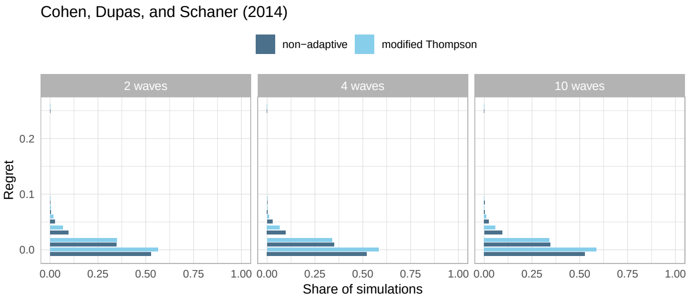


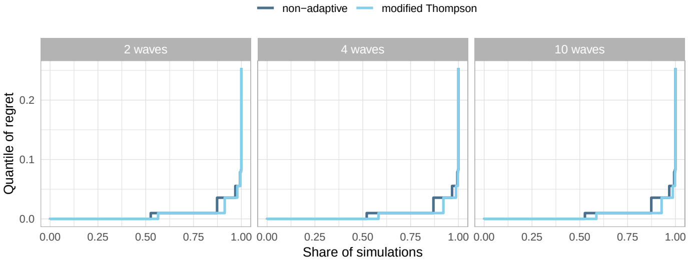
Regret and Share Optimal

Table: Cohen, Dupas, and Schaner (2014)

Statistic	2 waves	4 waves	10 waves
Regret			
modified Thompson	0.007	0.006	0.006
expected Thompson	0.007	0.006	0.006
Thompson	0.007	0.007	0.006
non-adaptive	0.009	0.009	0.009
Share optimal			
modified Thompson	0.565	0.582	0.587
expected Thompson	0.564	0.582	0.575
Thompson	0.562	0.581	0.590
non-adaptive	0.526	0.521	0.527
Units per wave	1080	540	216

Policy Choice and Regret Distribution





Theoretical analysis

Thompson sampling

- Literature: In-sample regret for bandit algorithms.
  - Agrawal and Goyal (2012) (Theorem 2):
    For Thompson sampling,

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

    where $\Delta^d = \max_{d'}\theta^{d'} - \theta^d$.
  - 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.
- This is good for in-sample welfare, bad for learning:
  - We stop learning about suboptimal treatments very quickly.
  - The posterior variance of θ^d for $d \neq d^*$ goes to zero at a rate no faster than $1/\log(M)$.

Modified Thompson sampling

Proposition

Assume fixed wave size $N_t = N$.
As $T \to \infty$, modified Thompson 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 modified Thompson 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 modified Thompson.
3. Claim 2: Suppose p_t^d goes to 0 at a faster rate for some d. Then modified Thompson 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)

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 modified Thompson 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

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

Inference

- For inference, we have to be careful with adaptive designs.
  1. **Standard inference** won't work:
     Sample means are biased, t-tests don't control size.
  2. But: **Bayesian** inference can ignore adaptiveness!
  3. **Randomization tests** can be modified to work.
- Example to get intuition for bias:
  - Flip a fair coin.
  - If head, flip again, else stop.
  - Probability dist: 50% tail-stop, 25% head-tail, 25% head-head.
  - Expected share of heads?

$$.5 \cdot 0 + .25 \cdot .5 + .25 \cdot 1 = .375 \neq .5.$$

- Randomization inference:
  - Strong null hypothesis: $Y_i^1 = \ldots = Y_i^k$.
  - Under null, easy to re-simulate treatment assignment.
  - Re-calculate test statistic each time.
  - Take $1 - \alpha$ quantile across simulations as critical value.

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...

Thank you!