Foundations of machine learning

Experiments for policy choice

Maximilian Kasy

Department of Economics, University of Oxford

Winter 2024

Outline

- Alternative objectives for the design of experiments.
- Exploration sampling as a modification of Thompson sampling.
- The oracle optimal allocation for the policy choice problem.
- Exploration sampling converges to the oracle optimal allocation.
- Simulations and empirical application.

Takeaways for this part of class

- Adaptive designs improve expected welfare.
- Features of the 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!
  - Asymptotically: Equalize power for comparisons
    of each suboptimal treatment to the optimal one.
- 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,
    because it equalizes power across suboptimal treatments.
  - Show that it dominates alternatives in calibrated simulations.

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 maximizes 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.

What is the objective of your experiment?

1. Getting precise treatment effect estimators, powerful tests:

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

$\Rightarrow$ Standard experimental design recommendations.

2. Maximizing the outcomes of experimental participants:

$$\max \sum_i \theta^{D_i}$$

$\Rightarrow$ Multi-armed bandit problems.

3. Picking a welfare maximizing policy after the experiment:

$$\max \theta^{d^*},$$

where d^* is chosen after the experiment.
$\Rightarrow$ This lecture.

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 $\boldsymbol{t}$: $\boldsymbol{n}_t = (n_t^1, \ldots, n_t^k)$.
- Outcomes of wave $\boldsymbol{t}$: $\boldsymbol{s}_t = (s_t^1, \ldots, s_t^k)$.
- 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 $\boldsymbol{t}+1$ is summarized by $\boldsymbol{m}_t$ and $\boldsymbol{r}_t$.
- Total trials for each treatment, total successes.

Design objective and Bayesian prior

- **Policy objective** $\theta^{d_T^*}$.
  - where $\boldsymbol{d}_T^*$ is chosen after the experiment.
- **Prior**
  - $\theta^d \sim Beta(\alpha_0^d, \beta_0^d)$, independent across $\boldsymbol{d}$.
  - Posterior after period $\boldsymbol{t}$: $\theta^d | \boldsymbol{m}_t, \boldsymbol{r}_t \sim Beta(\alpha_t^d, \beta_t^d)$

$$\alpha_t^d = \alpha_0^d + r_t^d$$
$$\beta_t^d = \beta_0^d + m_t^d - r_t^d.$$

- **Posterior expected social welfare**
  as a function of $\boldsymbol{d}$:

$$SW_T(d) = E[\theta^d | \boldsymbol{m}_T, \boldsymbol{r}_T],$$
$$= \frac{\alpha_T^d}{\alpha_T^d + \beta_T^d},$$
$$d_T^* \in \underset{d}{\operatorname{argmax}}\ SW_T(d).$$

Regret

- True optimal treatment: $d^{(1)} \in \arg\max_{d'} \theta^{d'}$.
- **Policy regret** when choosing treatment d:

$$\Delta^d = \theta^{d^{(1)}} - \theta^d.$$

- Maximizing expected social welfare is equivalent to minimizing the expected policy regret at T,

$$E[\Delta^d | \boldsymbol{m}_T, \boldsymbol{r}_T] = \theta^{d^{(1)}} - SW_T(d)$$

- **In-sample regret**: Objective considered in the bandit literature,

$$\frac{1}{M} \sum_{i,t} \Delta^{D_{it}}.$$

Different from policy regret $\Delta^{d_T^*}$!

Reminder: Thompson sampling

- **Thompson sampling**
  Assign each treatment with probability equal to the posterior probability that it is optimal.

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

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

$$D_{it} = \underset{d}{\operatorname{argmax}}\ \hat{\theta}_{it}^d.$$

- **Expected Thompson sampling**
  - Straightforward modification for the batched setting.
  - Assign non-random shares $\boldsymbol{p}_t^d$ of each wave to treatment $\boldsymbol{d}$.

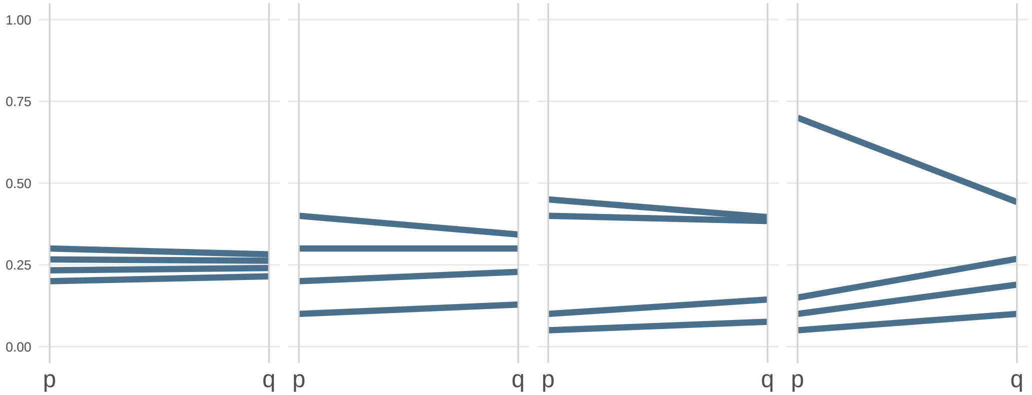
Exploration sampling

- 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 the following modification.
- **Exploration sampling**:
  Assign shares q_t^d of each wave to treatment d, where

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

- This modification
  1. yields rate-optimality (theorem coming up), and
  2. improves performance in our simulations.

Illustration of the mapping from Thompson to exploration sampling



The rate-optimal assignment: Lemma 1

Denote the estimated success rate of d at time T by $\hat{\theta}_T^d = \frac{1+r_T^d}{2+m_T^d}$.

The rate of convergence to zero of expected policy regret

$$\mathrm{R(T)} = \sum_d \Delta^d \cdot P\left(\operatorname*{argmax}_{d'} \hat{\theta}_T^{d'} = d\right)$$

is equal to the slowest rate of convergence Γ^d across $d \neq d^{(1)}$ for the probability of d being estimated to be better than $d^{(1)}$.

Lemma

- Assume that the optimal policy $d^{(1)}$ is unique. Suppose that for all d

$$\lim_{T\to\infty} -\frac{1}{NT} \log P\left(\hat{\theta}_T^d > \hat{\theta}_T^{d^{(1)}}\right) = \Gamma^d.$$

- Then

$$\lim_{T\to\infty} \left(-\frac{1}{NT} \log \mathrm{R(T)}\right) = \min_{d \neq d^{(1)}} \Gamma^d.$$

The rate-optimal assignment: Lemma 2

From Glynn and Juneja (2004):

- Characterize Γ^d as a function of the treatment allocation share for each $\boldsymbol{d}$, $\bar{q}^d$.
- The posterior probability p^d_T of $\boldsymbol{d}$ being optimal converges at the same rate Γ^d.

Lemma

Suppose that $\bar{q}^d_T = m^d_T/(NT)$ converges to $\bar{q}^d$ for all $\boldsymbol{d}$, with $\bar{q}^{d^{(1)}} = 1/2$. Then

1. $\lim_{T\to\infty} -\frac{1}{NT}\log P\left(\hat{\theta}^d_T > \hat{\theta}^{d^{(1)}}_T\right) = \Gamma^d$, and
2. $\operatorname{plim}_{T\to\infty} -\frac{1}{NT}\log p^d_T = \Gamma^d$,

where

$$\Gamma^d = G^d(\bar{q}^d)$$

for a function $G^d : [0,1] \to \mathbb{R}$
that is finitely valued, continuous, strictly increasing in $\bar{q}^d$, and satisfies $G^d(0) = 0$.

The rate-optimal assignment: Lemma 3

- Characterize the allocation of observations across the treatments $\boldsymbol{d}$ which maximizes the rate of $\mathrm{R(T)}$.
- Our main result shows that exploration sampling converges to this allocation.

Lemma

The rate-optimal allocation $\bar{\boldsymbol{q}}$, subject to the constraint $\bar{q}^{d^{(1)}} = 1/2$, is given by the unique solution to the system of equations

$$\sum_{d \neq d^{(1)}} \bar{q}^d = 1/2 \quad \text{and} \quad G^d(\bar{q}^d) = \Gamma^* > 0 \text{ for all } d \neq d^{(1)} \tag{1}$$

for some Γ^*. No other allocation, subject to the constraint $\bar{q}^{d^{(1)}} = 1/2$, can achieve a faster rate of convergence of $\mathrm{R(T)}$ than Γ^*.

Bandits don't achieve good rates for exploration

- Thompson sampling 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 convergence** for policy regret!

- By contrast (easy to show): Any algorithm that assigns shares
  converging to non-zero shares for each treatment
  achieves **exponential convergence** for our objective.

- Our result (next slide): Exploration sampling achieves the
  (constrained) best exponential rate.

Exploration sampling is rate optimal

Theorem

Consider exploration sampling in a setting with fixed wave size $N_t = N \geq 1$. Assume that $\theta^{d^{(1)}} < 1$ and that the optimal policy $d^{(1)}$ is unique. As $T \to \infty$, the following holds:

1. The share of observations $\bar{q}_T^{d^{(1)}}$ assigned to the best treatment converges in probability to $1/2$.

2. The share of observations $\bar{q}_T^d$ assigned to treatment d converges in probability to a non-random share $\bar{q}^d$ for all $d \neq d^{(1)}$.
   $\bar{q}^d$ is such that $-\frac{1}{NT} \log p_t^d \to^p \Gamma^*$
   for some $\Gamma^* > 0$ that is constant across $d \neq d^{(1)}$.

3. Expected policy regret converges to 0 at the same rate Γ^*, that is,
   $-\frac{1}{NT} \log \mathrm{R(T)} \to^p \Gamma^*$.
   No other assignment shares $\bar{q}^d$ exist for which $\bar{q}^{d^{(1)}} = 1/2$ and $\mathrm{R(T)}$ goes to 0 at a faster rate than Γ^*.

Sketch of proof

Our proof draws on several Lemmas of Glynn and Juneja (2004) and Russo (2016).
Proof steps:

1. Each treatment is assigned infinitely often.
   $\Rightarrow p_T^d$ goes to $\mathbf{1}$ for the optimal treatment and to $\mathbf{0}$ for all other treatments.

2. Claim 1 then follows from the definition of exploration sampling.

3. Claim 2: Suppose p_t^d goes to $\mathbf{0}$ 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 the Lemmas discussed before.

Calibrated simulations

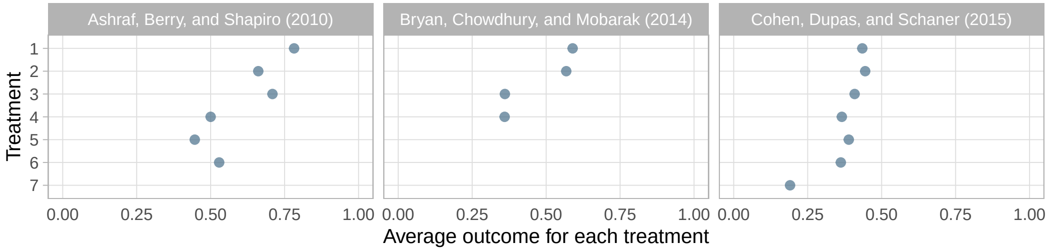
- Simulate data calibrated to estimates of 3 published experiments.
- Set θ equal to observed average outcomes for each stratum and treatment.
- Total sample size same as original.

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



Treatment arms labeled 1 up to 7:

- Ashraf et al. (2010): Kw 300 - 800 price for water disinfectant.
- Bryan et al. (2014): Migration incentives - cash, credit, information, and control.
- Cohen et al. (2015): Price of Ksh 40, 60, and 100 for malaria tablets, each with and without free malaria test, and control of Ksh 500.

Summary of simulation findings

- With two waves, relative to non-adaptive assignment:
  - Thompson reduces average policy 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).
- For in-sample regret, Thompson is best,
  followed closely by exploration sampling.

Ashraf, Berry, and Shapiro (2010)

Statistic	2 waves	4 waves	10 waves
Average policy regret			
exploration sampling	0.0017	0.0010	0.0008
expected Thompson	0.0022	0.0014	0.0013
non-adaptive	0.0051	0.0050	0.0051
Share optimal			
exploration sampling	0.978	0.987	0.989
expected Thompson	0.971	0.981	0.982
non-adaptive	0.933	0.935	0.933
Average in-sample regret			
exploration sampling	0.1126	0.0828	0.0701
expected Thompson	0.1007	0.0617	0.0416
non-adaptive	0.1776	0.1776	0.1776
Units per wave	502	251	100

Bryan, Chowdhury, and Mobarak (2014)

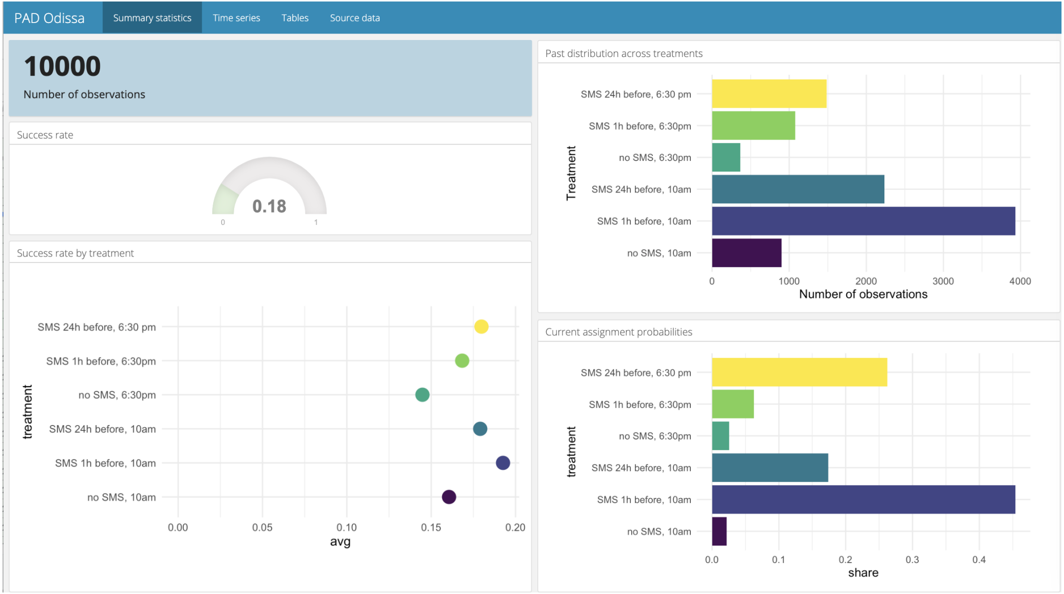
Statistic	2 waves	4 waves	10 waves
Average policy regret			
exploration sampling	0.0045	0.0041	0.0039
expected Thompson	0.0048	0.0044	0.0043
non-adaptive	0.0055	0.0054	0.0054
Share optimal			
exploration sampling	0.792	0.812	0.820
expected Thompson	0.777	0.795	0.801
non-adaptive	0.747	0.748	0.749
Average in-sample regret			
exploration sampling	0.0655	0.0386	0.0254
expected Thompson	0.0641	0.0359	0.0205
non-adaptive	0.1201	0.1201	0.1201
Units per wave	935	467	187

Cohen, Dupas, and Schaner (2015)

Statistic	2 waves	4 waves	10 waves
Average policy regret			
exploration sampling	0.0070	0.0063	0.0060
expected Thompson	0.0074	0.0065	0.0061
non-adaptive	0.0086	0.0087	0.0085
Share optimal			
exploration sampling	0.567	0.586	0.592
expected Thompson	0.560	0.582	0.589
non-adaptive	0.526	0.524	0.529
Average in-sample regret			
exploration sampling	0.0489	0.0374	0.0314
expected Thompson	0.0467	0.0345	0.0278
non-adaptive	0.0737	0.0737	0.0737
Units per wave	1080	540	216

Implementation in the field

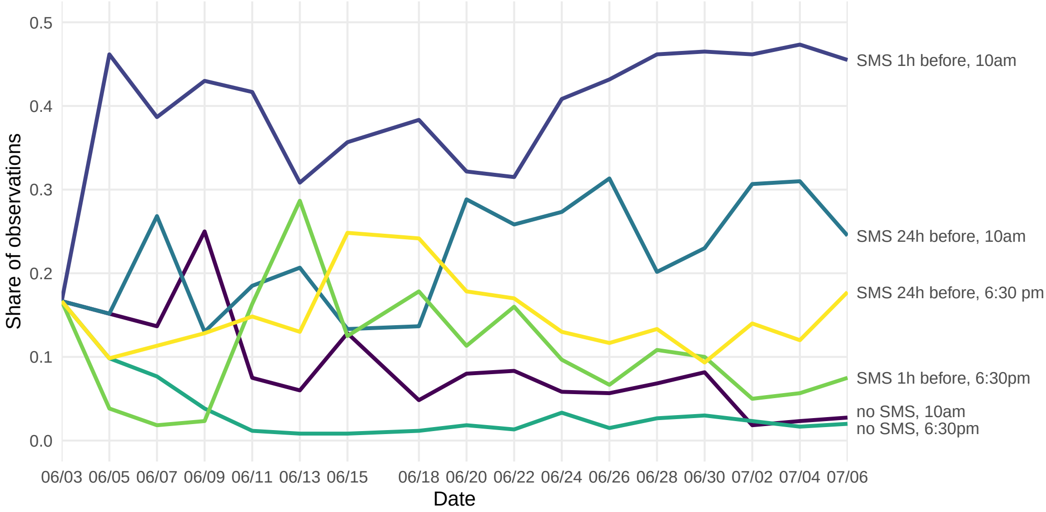
- NGO Precision Agriculture for Development (PAD),
  and Government of Odisha, India.
- Enrolling rice farmers into customized advice service by mobile phone.
- Waves of 600 farmers called through automated service; total of 10K calls.
- Outcome: did the respondent answer the enrollment questions?


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

Outcomes and posterior parameters

Treatment		Outcomes			Posterior		
Call time	SMS alert	m_T^d	r_T^d	r_T^d/m_T^d	mean	SD	p_T^d
10am	-	903	145	0.161	0.161	0.012	0.009
10am	1h ahead	3931	757	0.193	0.193	0.006	0.754
10am	24h ahead	2234	400	0.179	0.179	0.008	0.073
6:30pm	-	366	53	0.145	0.147	0.018	0.011
6:30pm	1h ahead	1081	182	0.168	0.169	0.011	0.027
6:30 pm	24h ahead	1485	267	0.180	0.180	0.010	0.126

Assignment shares over time



References

- *Glynn, P. and Juneja, S. (2004). A large deviations perspective on ordinal optimization. In* Proceedings of the 36th Winter simulation conference, *pages 577–585. Winter Simulation Conference.*
- *Russo, D. (2016). Simple bayesian algorithms for best arm identification. In* Conference on Learning Theory, *pages 1417–1418.*
- *Kasy, M. and Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice.* Econometrica, *89(1):113–132.*

- Interactive dashboard for treatment assignment:
  `https://maxkasy.shinyapps.io/exploration_sampling_dashboard/`