ADAPTIVE TREATMENT ASSIGNMENT IN EXPERIMENTS FOR POLICY CHOICE

Maximilian Kasy and Anja Sautmann

The goal of many experiments is to inform the choice between different policies. However, standard experimental designs are geared toward point estimation and hypothesis testing. We consider the problem of treatment assignment in an experiment with several non-overlapping waves where the goal is to choose among a set of possible policies (treatments) for large-scale implementation. We discuss the optimal experimental design as well as a computationally tractable approximation, based on a modification of Thompson sampling. Both learn from earlier waves by assigning more experimental units to the better-performing treatments in later waves, but to a lesser extent than algorithms for multi-armed bandit settings. Theoretical results and calibrated simulations demonstrate improvements in welfare, relative to both non-adaptive designs as well as standard Thompson sampling.

Keywords: Experimental design, field experiments, optimal policy.

1. introduction

The main objective of an academic researcher conducting a randomized controlled trial (RCT) is typically to generate a point estimate of the treatment effect and a corresponding standard error, in order to test the null hypothesis that the average effect equals 0. The research design is chosen to maximize power for tests of this null, for example by assigning an equal number of units to different treatments, and by stratifying the sample by pre-determined covariates (see for instance Athey and Imbens 2017). Such RCTs are designed to answer the question "Does this program have a significant effect?"

However, the objective of an NGO or government who considers conducting an experiment to evaluate its programs is often slightly different: instead of estimating effect size, they are interested in identifying and implementing the best out of several possible policies or policy variants. In other words, they would like to answer questions such as "Which program will have the largest effect?" We show that the objective of informing policy choice leads to design recommendations that are qualitatively different from standard RCT recommendations.

We consider an experimental setting with multiple waves of experimental units, and multiple treatments (policies). We assume that the outcome of interest is binary. At the beginning of each wave, the number of units assigned to each treatment arm is decided. After conclusion of the wave, prior beliefs about treatment effects are updated based on the observed success rates (outcomes) in the different treatment arms. Then treatments are assigned for the next wave, based on these updated beliefs. Once the experiment is concluded, one of the treatments is picked for full-scale implementation. The objective is to maximize the average outcomes for this full-scale implementation, net of the costs of treatment.

Department of Economics, Harvard University, maximiliankasy@fas.harvard.edu
J-PAL, Massachusetts Institute of Technology, sautmann@mit.edu.

Our setting is closely related to the well-known "multi-armed bandit" problem (cf. Weber et al., 1992; Bubeck and Cesa-Bianchi, 2012; Russo et al., 2018), but with the key difference that there is no "exploitation" motive, and thus no exploitation-exploration tradeoff. This is because in our setting the goal is to maximize outcomes after the experiment is concluded, but not during the experiment. We believe that this case is practically relevant, since in many settings indefinite experimentation is not feasible because of costs or political constraints. We focus on the boundary case where the experimental sample is negligible relative to the population of interest for conceptual clarity.

The policy choice problem described above defines a finite horizon dynamic stochastic optimization problem. The actions in each wave (period) are the different possible treatment assignments; the states are current beliefs over treatment effects; and transitions between states from period to period are determined by experimental outcomes. In the final period, the action consists in the choice of policy for implementation, after which welfare is realized as the average per-unit outcome net of costs.

This optimization problem can in principle be solved analytically using backward induction. In realistic settings, however, finding exact solutions quickly becomes infeasible, due to exploding state and action spaces. We propose the following assignment algorithm, which is a modified version of so-called Thompson sampling. In Thompson sampling, each unit is assigned to a given treatment d with probability equal to the posterior probability p_t^d (given past outcomes) that it is in fact the optimal treatment. We modify this prescription in two ways. First, taking into account that our setting has waves rather than sequential arrival of units, we do not independently assign each unit to a treatment based on the probabilities p_t^d, but instead we assign a corresponding share of each wave to the different treatments. Second, and more importantly, we replace the assignment shares p_t^d by shares equal to $q_t^d = S_t \cdot p_t^d \cdot (1 - p_t^d)$, where S_t is a normalizing constant. The shares q_t^d would arise if we ran conventional Thompson sampling sequentially, within a given wave, but forced the algorithm to never assign the same treatment twice in a row.

We show, using both simulations and theoretical results, that this modification improves expected welfare. It avoids assigning more than 50% of the sample to the highest-performing treatment, and in large samples it equalizes power for rejecting each of the sub-optimal treatments. This behavior is optimal for the convergence rate of welfare, while standard Thompson sampling is not, as discussed in Section 4.

In Section 5 we provide simulation evidence on the performance of modified Thompson sampling compared with alternative assignment algorithms, in particular a non-adaptive RCT (with equal treatment arm size) and standard Thompson sampling. We evaluate these algorithms according to the loss that is incurred from picking another than the highest-performing treatment option with some probability. Our simulations use parameters and sample sizes calibrated to data from three published experiments in development economics (Ashraf et al., 2010;

Bryan et al., 2014; Cohen et al., 2015). Confirming theoretical predictions, modified Thompson sampling consistently performs better than standard Thompson sampling, which in turn outperforms conventional non-adaptive designs. Furthermore, the gains from adaptive treatment assignment are larger when the experiment is divided into more waves, for the same total sample size.

In Section 6, we demonstrate the practical feasibility of our proposal in an experiment that uses the roll-out of a phone information campaign to rice farmers in Odisha to test different enrollment protocols. Despite small differences between the effects of alternative treatments, we are able to pin down the optimal treatment with high probability. As predicted by our theory, p_t^d converges to 0 at at comparable speed for all the sub-optimal treatments.

In a companion paper intended for practitioners, we will discuss additional empirical applications, non-binary outcomes, covariates and targeted treatment assignment, the choice of sample size, and statistical inference.

The idea of adaptive treatment assignment is almost as old as the idea of randomized experiments (Thompson, 1933). Adaptive experimental designs have been used in clinical trials (Berry, 2006), and in the targeting of online advertisements (Russo et al., 2018), but they have not yet entered the standard toolkit for RCTs in economics, see e.g. Duflo and Banerjee (2017). Under some conditions, the optimal solution to the Bandit problem can be expressed in terms of choosing the arm corresponding to highest "Gittins index," cf. Weber et al. (1992). In practice, most applications use heuristic algorithms rather than solving for the optimal assignment, such as the Upper Confidence Bound algorithm (UCB), and Thompson sampling (Russo et al., 2018). A fairly recent literature characterizes the expected regret of these algorithms, see for example Bubeck and Cesa-Bianchi (2012). Generalizations of the Bandit problem are discussed under the name of reinforcement learning in the machine learning literature (Ghavamzadeh et al. (2015) and Sutton and Barto (2018)). Lastly, Russo (2016) considers a problem closely related to ours, namely the problem of maximizing the probability of picking the best treatment (rather than maximizing expected welfare). Our theoretical analysis in Section 4 below draw son insights from this paper, and on the impossibility result of Bubeck et al. (2011).

2. SETUP

Consider a policymaker who wants to maximize the expected value of a binary outcome variable, i.e., a success rate. She has to choose between three or more different policies (treatments) and she can use an experiment that proceeds in multiple waves (repeated cross-sections). At the end of each experimental wave, outcomes are observed, and treatment assignment in subsequent waves can be based on these observed outcomes. After the experiment concludes, a treatment is chosen for large-scale implementation. Our goal is to derive optimal and approximately optimal experimental designs for this setting.

Treatments and potential outcomes. The experiment takes place in waves $t = 1, \ldots, T$. Each wave t is a new random draw of N_t experimental units $i = 1, \ldots, N_t$ from the population of interest (so that the waves are repeated cross-sections).

Each person or unit i in period t can receive one of k different treatments $D_{it} \in \{1, \ldots, k\}$, resulting in a binary outcome $Y_{it} \in \{0, 1\}$. Outcome Y_{it} is determined by the potential outcome equation $Y_{it} = \sum_{d=1}^{k} \mathbf{1}(D_i = d) \cdot Y_{it}^d$. This assumption implies in particular that there is no interference, i.e., outcomes are not affected by the treatments others receive. Random sampling means that the potential outcome vector $(Y_{it}^1, \ldots, Y_{it}^k)$ for unit i in period t is an i.i.d. draw from the population of interest. Each treatment d has a stationary unobserved average potential outcome (also known as average structural function) $\theta^d = E[Y_{it}^d]$.

Treatment assignment and state space during the experiment. Denote by $n_t^d = \sum_i \mathbf{1}(D_{it} = d)$ the number of units assigned to treatment d in wave t. The treatment assignment in wave t is summarized by the vector $\boldsymbol{n}_t = (n_t^1, \ldots, n_t^k)$ with $\sum_d n_t^d = N_t$. The experimenter's problem is to choose $\boldsymbol{n}_t$ at the beginning of wave t.

Denote $s_t^d = \sum_i \mathbf{1}(D_{it} = d, Y_{it} = 1)$ the number of successes (outcome $Y_{it} = 1$) among those in treatment group d in wave t. The outcome of wave t can be summarized by the vector $\boldsymbol{s}_t = (s_t^1, \ldots, s_t^k)$, where $s_t^d \leq n_t^d$, collecting the number of successes in each of the treatment groups in wave t. These outcomes are observed at the end of wave t. Treatment assignment in wave $t + 1$ can depend on the outcomes of waves 1 to t, and on a randomization device.

Denote the cumulative versions of these terms by $m_t^d = \sum_{t' \leq t} n_{t'}^d$, $r_t^d = \sum_{t' \leq t} s_{t'}^d$, and $\boldsymbol{m}_t = (m_t^1, \ldots, m_t^k)$, $\boldsymbol{r}_t = (r_t^1, \ldots, r_t^k)$. Thus, m_t^d is the total number of units assigned to treatment d in waves 1 through t, and r_t^d is the total number of successes among these units. With i.i.d. potential outcomes, all relevant information for the experimenter at the beginning of period $t + 1$ is summarized by $\boldsymbol{m}_t$ and $\boldsymbol{r}_t$.

Policy choice and welfare. After wave T, a policy $d^* \in 1, \ldots, k$ will be chosen and implemented, with the objective of maximizing the expected average of the outcome Y for the whole (remaining) population of interest, net of the unit cost c^d of treatment. This objective, per-capita expected social welfare of policy d at the end of the experiment, is given by

$$SW(d) = E[\theta^d | \boldsymbol{m}_T, \boldsymbol{r}_T] - c^d. \tag{2.1}$$

The optimal policy choice after the experiment is given by $d^* = \text{argmax}_d \, SW(d)$. In this formulation, social welfare does not include the outcomes of participants in the experiment. This implies that treatment assignment in each experimental wave is chosen to maximize learning and optimize the policy choice after wave T.

Excluding the welfare of participants from the optimization problem is justified if the number of experimental units is small relative to the population of interest. A concern for the welfare of participants could easily be added to our objective function, resulting in a hybrid setting between the bandit problem and the setting considered here.

Bayesian prior and posterior. Under our assumptions, Y^d has a Bernoulli distribution with unknown parameter θ^d: $Y^d \sim Ber(\theta^d)$. We assume that the policymaker holds prior belief $\theta^d \sim Beta(\alpha_0^d, \beta_0^d)$. The θ^d are mutually independent across d. A special case, and the default for applications later in this paper, is the uniform prior $\boldsymbol{\theta} \sim Uniform\left([0,1]^k\right)$, corresponding to $\alpha_0^d = \beta_0^d = 1$ for all d.

After the outcomes for periods $1, \ldots, t$ are realized, the posterior distribution is given by $\theta^d | \boldsymbol{m}_t, \boldsymbol{r}_t \sim Beta(\alpha_t^d, \beta_t^d)$, where $\alpha_t^d = \alpha_{t-1}^d + s_t^d = \alpha_0^d + r_t^d$ and $\beta_t^d = \beta_{t-1}^d + n_t^d - s_t^d = \beta_0^d + m_t^d - r_t^d$. Moreover, expected social welfare after period T, based on the expected success rate for d, is

$$SW(d) = \tfrac{\alpha_0^d + r_T^d}{\alpha_0^d + \beta_0^d + m_T^d} - c^d. \tag{2.2}$$

From the perspective of the experimenter, the outcomes of wave t, after making the treatment assignment decision $\boldsymbol{n}_t$, are subject to two sources of uncertainty: the uncertainty about $\boldsymbol{\theta}$, given $\boldsymbol{m}_{t-1}$ and $\boldsymbol{r}_{t-1}$, and the sampling uncertainty over the distribution of $\boldsymbol{s}_t$, given $\boldsymbol{\theta}$ and $\boldsymbol{n}_t$. The former is given by the $Beta(\alpha_{t-1}^d, \beta_{t-1}^d)$ distribution, the latter by Binomial distributions with parameters n_t^d and θ^d. Integrating out the unknown parameter $\boldsymbol{\theta}$, we get that the number of successes for each treatment in wave t follows a Beta-Binomial distribution,

$$\begin{aligned} P(s_t^d = s | \boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) &= E[P(s_t^d = s | \theta^d, n_t^d) | \boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t] \\ &= \binom{n_t^d}{s} \frac{B(\alpha_{t-1}^d + s, \beta_{t-1}^d + n_t^d - s)}{B(\alpha_{t-1}^d, \beta_{t-1}^d)}, \end{aligned} \tag{2.3}$$

where $B(\cdot, \cdot)$ is the beta function.

3. TREATMENT ASSIGNMENT

3.1. *Optimal assignment*

The choice of treatment assignment $\boldsymbol{n}_t$ for each $t = 1, \ldots, T$ is a dynamic stochastic optimization problem that can be solved using backward induction. The state at the end of wave $t-1$ is given by $(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1})$, and the action in t is given by $\boldsymbol{n}_t$. The transition between states is described by $\boldsymbol{m}_t = \boldsymbol{m}_{t-1} + \boldsymbol{n}_t$, $\boldsymbol{r}_t = \boldsymbol{r}_{t-1} + \boldsymbol{s}_t$, where the success probabilities are given by Equation (2.3).

Denote by V_t the value function after completion of wave t, that is, expected welfare assuming that all future treatment assignment decisions will be optimal, and that the optimal policy is implemented after the experiment. V_t is a function

of the state $(\boldsymbol{m}_t, \boldsymbol{r}_t)$. After the experiment is concluded, the value function is given by expected welfare for the optimal choice of policy, based on current beliefs:

$$V_T(\boldsymbol{m}_T, \boldsymbol{r}_T) = \max_d \left(E[\theta^d | \boldsymbol{m}_T, \boldsymbol{s}_T] - c^d\right) = \max_d \left(\tfrac{\alpha_0^d + r_T^d}{\alpha_0^d + \beta_0^d + m_T^d} - c^d\right). \tag{3.1}$$

Denote by U_t the action value function, given by expected welfare at the beginning of wave t when treatment assignment is $\boldsymbol{n}_t$, assuming all future assignment decisions will be optimal:

$$U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) = \sum_{\boldsymbol{s}: \boldsymbol{s} \leq \boldsymbol{n}_t} P(\boldsymbol{s}_t = \boldsymbol{s} \,|\, \boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) V_t\left(\boldsymbol{m}_{t-1} + \boldsymbol{n}_t, \boldsymbol{r}_{t-1} + \boldsymbol{s}\right), \tag{3.2}$$

where the probabilities for each vector of successes $P(\boldsymbol{s}_t = s | \boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t)$ are given by the Beta-Binomial distribution of Equation (2.3). The period t value function and the optimal treatment assignment satisfy

$$\begin{aligned} V_{t-1}(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}) &= \max_{\boldsymbol{n}_t:\, \sum_d n_t^d \leq N_t} U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) \\ \boldsymbol{n}_t^*(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}) &= \operatorname*{argmax}_{\boldsymbol{n}_t:\, \sum_d n_t^d \leq N_t} U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t). \end{aligned} \tag{3.3}$$

Together, these equations define a solution for the experimental design problem.

In Appendix A.2, we discuss optimal designs for a simple numerical example. In this example (i) adaptivity with two equal-sized waves dominates alternative splits, and in particular non-adaptive assignments, and (ii) the optimal treatment assignment in wave 2 assigns more units to those treatments that performed better in wave 1.

Computational cost of optimal solutions We can solve for the optimal treatment assignment using dynamic programming. This involves brute-force enumeration of all possible outcomes and actions. With larger sample sizes and a greater number of treatments, however, solving for the optimal assignment quickly becomes infeasible.

The most time-efficient approach for dynamic programming uses full memoization, where the the value function is calculated and stored for every possible state. At the end of wave t there are $\binom{M_t+k-1}{k-1} = O(M_t^{k-1})$ possible values m_t, and for each m_t there are $\prod_d m_t^d = O(M_t^k)$ possible values of r_t, so that the number of possible states at the end of wave t is of order $O(M_t^{2k-1})$.

Suppose $t < T$. Then for each of these states we need to calculate the value function, maximizing over the expected action value for each possible action n_{t+1}, where the expectation is over each possible realization of s_{t+1}. There are $\binom{N_{t+1}+k-1}{k-1} = O(N_{t+1}^{k-1})$ possible actions n_t, and $\prod_d n_{t+1}^d = O(N_{t+1}^k)$ possible realizations of $\boldsymbol{s}_{t+1}$ for each n_{t+1}, so that the required computation time for V_t at a given state is of order $O(N_{t+1}^{2k-1})$. For $t = T$, we only need to maximize over k possible actions (policy choices).

Collecting terms, we get that the computational time complexity for dynamic programming with full memoization in this setting is of order

$$\sum_{t=1}^{T-1} O\left((M_t N_{t+1})^{2k-1}\right) + O(M_T^{2k-1}k), \tag{3.4}$$

and the memory complexity is of order $\sum_{t=1}^{T} O\left(M_t^{2k-1}\right)$.

3.2. *Thompson sampling*

An alternative to full optimization is the use of simpler heuristic algorithms. Such algorithms are widely used for bandit problems, for instance in the placement of online ads. One of the most popular (and oldest) such algorithms is so-called Thompson sampling, originally proposed by Thompson (1933) in the context of clinical trials.

Consider the special case of our setting where each wave is of size 1, so that units arrive sequentially (and we can drop the subscript i). In each period t, assign treatment d with probability equal to the posterior probability, given past outcomes, that it is in fact the optimal treatment,

$$p_t^d = P(D_t = d|\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}) = P\left(d = \underset{d'}{\text{argmax}}\ (\theta^{d'} - c^{d'})|\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}\right). \tag{3.5}$$

This prescription is easy to implement, by sampling just one draw $\hat{\theta}_t$ from the posterior given $\boldsymbol{m}_{t-1}$ and $\boldsymbol{r}_{t-1}$, and setting $D_t = \text{argmax}_d(\hat{\theta}_t^d - c^d)$. In the context of the Beta-Binomial model outlined above, $\hat{\theta}_t$ is sampled from its Beta posterior. Thompson sampling can also be applied in much more general settings, with more complicated policy spaces, prior distributions, and likelihoods. An excellent overview can be found in Russo et al. (2018). This approach is easily adapted to batched settings such as ours, where posteriors are based on the outcomes of all preceding waves.

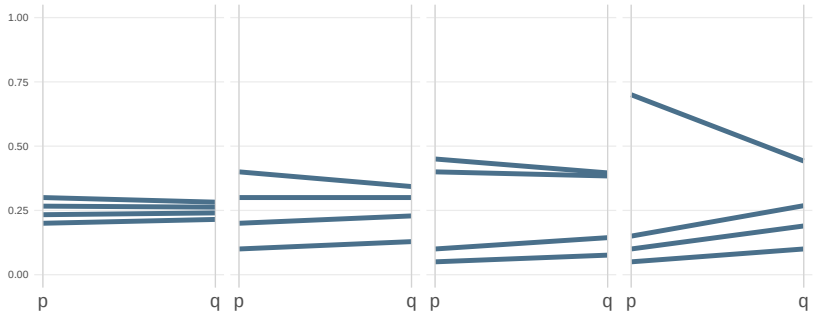
3.3. *Modified Thompson sampling*

We can improve on standard Thompson sampling in our context. We propose the following two modifications. The **first modification** is designed to reduce randomness in the treatment assignment. Rather than drawing each D_{it} independently from the distribution $(p_t^1, \ldots, p_t^k)$, we assign a non-random share p_t^d (up to required rounding) of observations in wave t to treatment d. We will refer to treatment assignment based on this modification alone as *expected Thompson sampling*.

The **second modification** replaces the assignment probabilities $(p_t^1, \ldots, p_t^k)$ with the following transformed probabilities:

$$q_t^d = S_t \cdot p_t^d \cdot (1 - p_t^d), \qquad S_t = \frac{1}{\sum_d p_t^d \cdot (1 - p_t^d)}. \tag{3.6}$$

FIGURE 3.1.— Illustration of modified Thompson probabilities



Notes: This figure shows examples of the mapping from the vector of Thompson probabilities $\boldsymbol{p}_t$ to the vector of modified Thompson probabilities $\boldsymbol{q}_t$.

We analyze this modification and its justification in Section 4 below. We will refer to treatment assignment based on both of these modifications as *modified Thompson sampling*. To get some initial intuition for the mapping from the vector of probabilities $\boldsymbol{p}_t$ to the transformed vector $\boldsymbol{q}_t$, Figure 3.1 plots a few examples.

4. ANALYSIS OF MODIFIED THOMPSON SAMPLING

In this section, we characterize the behavior of modified Thompson sampling. We first review results from the literature on the behavior of Thompson sampling and other bandit algorithms. Such algorithms assign a lot of observations to the best-performing treatment. This is good for the welfare of the experimental participants, but is not optimal for for the ability of a policymaker to distinguish the best treatment after the experiment.

We then provide a heuristic motivation of modified Thompson sampling, which arises if we force Thompson sampling to not assign the same treatment twice in a row, thereby improving power for comparisons of relevant alternatives. We finally present our key result, providing a theoretical characterization of modified Thompson sampling, and showing that it is (constrained) rate optimal for our objective.

4.1. *The large sample behavior of Thompson sampling*

In many bandit problems, the goal is to minimize average in-sample regret $\frac{1}{T}\sum_{t=1}^{T}\Delta^{D_t}$, where $\Delta^d = \max_{d'}\theta^{d'} - \theta^d$. Agrawal and Goyal (2012) (Theorem 2) have shown that in-sample regret for Thompson sampling (in the binary outcome

setting, with sequential arrival) satisfies the bound

$$\lim_{T\to\infty} E\left[\frac{\sum_{t=1}^{T}\Delta^{D_t}}{\log T}\right] \leq \left(\sum_{d\neq d^*}\frac{1}{(\Delta^d)^2}\right)^2. \tag{4.1}$$

As first shown by Lai and Robbins (1985), no adaptive experimental design can do better than this $\log T$ rate; the proof of this lower bound is reviewed in Section 2.3 of Bubeck and Cesa-Bianchi (2012).

This implies that Thompson sampling only assigns a share of units of order $\log(T)/T$ to treatments other than the optimal treatment, so that we effectively stop learning about the performance of suboptimal treatments very quickly. This behavior of Thompson sampling, which makes it a good choice for maximizing in-sample welfare, limits its benefits for ex-post policy choice. This intuition is formalized by Bubeck et al. (2011). Their Theorem 1 implies that any algorithm that achieves a $\log(T)/T$ rate for in-sample regret, such as Thompson sampling, can at most achieve a polynomial rate of convergence to 0 for the probability of choosing a sub-optimal treatment after the experiment, and thus for our objective.

This stands in contrast to algorithms which assign a fixed, non-zero share of observations to each treatment, such as conventional (non-adaptive) designs. Such algorithms, and more generally algorithms with shares converging to non-zero shares, achieve an exponential rate of convergence. We prove below that modified Thompson sampling achieves the best possible exponential rate of convergence, subject to some constraint.

We should emphasize that these are large-sample results. For realistic sample sizes, our simulations show that Thompson sampling considerably outperforms conventional non-adaptive designs, despite its slower convergence, while being outperformed by modified Thompson sampling.

4.2. *Alternating Thompson*

The intuition that better performance with respect to our objective can be achieved by assigning both the best treatment and its close competitors with comparable frequency suggests the following *alternating Thompson* algorithm, which leads to a heuristic motivation for our modified Thompson procedure.

Suppose that within a given wave we sequentially assign treatments based on the Thompson probabilities $\boldsymbol{p}$, except that we don't assign the same treatment twice in a row. This algorithm defines a Markov chain for the sequence of assigned treatments. The probability of transitioning from treatment d' to treatment $d \neq d'$ is given by $\frac{p^d}{1-p^{d'}}$. This Markov chain has a stationary distribution $\boldsymbol{q}$, where the stationary distribution satisfies the equations

$$q^d = \sum_{d'\neq d} q^{d'}\frac{p^d}{1-p^{d'}} \tag{4.2}$$

for $d \in \{1, \ldots, k\}$. Moreover, by the mean ergodic theorem, the assignment shares of the "alternating" algorithm converge to the stationary distribution characterized by Equation (4.2). We can solve explicitly for $\boldsymbol{q}$. Denote $S = \sum_d \frac{q^d}{1-p^d}$. Then Equation 4.2 can be rewritten as $\frac{q^d}{p^d} = S - \frac{q^d}{1-p^d}$, and some algebra yields $q^d = S \cdot p^d \cdot (1-p^d)$ and $S = \frac{1}{\sum_d p^d \cdot (1-p^d)}$. This implies that for large wave sizes the alternating Thompson algorithm assigns the same share of observations to each treatment as our modified Thompson algorithm.

4.3. *The large sample behavior of modified Thompson sampling*

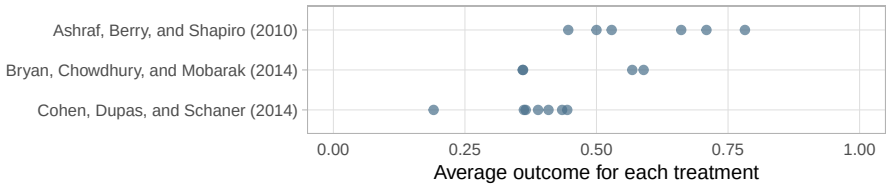
We now turn to our main characterization of modified Thompson sampling. The proof of the following Theorem builds on Russo (2016), and in particular on Proposition 7, as well as Lemma 12 through 14 in Appendix G.1 therein. Our theorem characterizes the behavior of modified Thompson sampling in settings with many waves and fixed wave sizes.

Theorem 1 *Consider the setting of Section 2 with fixed wave size $N_t = N$, and the modified Thompson algorithm as defined in Section 3.3. Assume that the optimal policy* $\operatorname{argmax}_d \theta^d$ *is unique. As $T \to \infty$, modified Thompson assignment satisfies the following:*

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 $\bar{q}^d$.*
3. *$\bar{q}^d$ is such that the posterior probability of d being optimal goes to* 0 *at the same exponential rate for all sub-optimal treatments.*
4. *No other assignment algorithm for which statement 1 holds has average regret going to* 0 *at a faster rate than modified Thompson sampling.*

The proof of this Theorem can be found in Appendix A.1; we proceed in several steps. First, we show that each treatment is assigned infinitely often. By a basic consistency result, this implies that p_T^d goes to 1 for the optimal treatment and to 0 for all other treatments. Claim 1 then follows from the definition of modified Thompson sampling. Second, we show claims 2 and 3 by contradiction. Suppose p_t^d goes to 0 at a faster rate for any one of the sub-optimal treatments d. Then modified Thompson sampling would effectively stop assigning this treatment d. This in turn allows the other sub-optimal treatments to "catch up." Lastly, efficiency (claim 4) holds because the algorithm balances the rate of convergence of posterior probabilities (or equivalently, of power) across treatments. That this is optimal follows from decreasing marginal returns of additional observations, for each treatment arm, in large samples.

FIGURE 5.1.— Average outcomes across treatment arms in published experiments



5. CALIBRATED SIMULATIONS

We next present simulation evidence on the performance of alternative treatment assignment algorithms, using parameter vectors and sample sizes calibrated to data from published experiments in development economics. The purpose of calibration is to "tie our hands" in choosing designs for our simulations. We opted for simplicity, rather than realism, in the assumptions driving our calibrations.

Experiments from the literature We consider the experiments discussed in Ashraf et al. (2010), Bryan et al. (2014), and Cohen et al. (2015). Our simulations use sample sizes equal to the original experiments. Appendix A.3 summarizes the context of these experiments. We make two simplifying assumptions. First, we ignore clustering in the sampling and treatment assignment of the original experiments. Second, we assume that the policymaker's goal is to maximize the average of the measured outcome.

Figure 5.1 shows the average outcomes across treatment arms for each of the three experiments. We set the vectors $\boldsymbol{\theta}$ equal to these average outcomes, for the purpose of our simulations. These vectors show interesting differences across the three experiments, which will be relevant for understanding the results of our simulations. For Ashraf et al. (2010), there are roughly evenly spaced average outcomes. This is a setting where it is comparatively easy to statistically detect which treatments are performing better, so that we would expect benefits of adaptation even for moderate sample sizes. For Bryan et al. (2014), there are two worse treatments (these two treatments are indistinguishable in Figure 5.1), and two better treatments that are very close. In this setting, it is easy to detect which two treatments perform better. Among these two, however, it takes a large amount of information to figure out which is the best. The returns of finding the best treatment among the top two, on the other hand, are not very large. For Cohen et al. (2015), the top 6 treatments are again roughly evenly spaced (the second and third treatment from the bottom are again indistinguishable in Figure 5.1). This setting is similar to Ashraf et al. (2010), except that the best treatments are closer and thus harder to distinguish.

Algorithms We compare four different algorithms. The first algorithm, which serves as a benchmark, is *non-adaptive* and assigns an equal share of units to each of the treatment arms. The second algorithm is standard *Thompson* sampling. The third algorithm, *expected Thompson*, assigns a non-random share of units in each wave based on the Thompson probabilities. The fourth algorithm, our preferred approach, is *modified Thompson* sampling as described in Section 3.3.

Performance criteria We evaluate the performance of these algorithms in terms of the distribution of regret across 50,000 simulation draws. Regret is given by the difference between the welfare generated by the optimal treatment, and welfare for the policy d^* with the highest posterior mean after conclusion of the experiment. That is,

$$d^* = \operatorname*{argmax}_d \; E[\theta^d | \boldsymbol{m}_T, \boldsymbol{r}_T], \qquad \text{regret} = \Delta^{d^*} = \max_d \theta^d - \theta^{d^*}.$$

For each of our simulations, the vector $\boldsymbol{\theta}$ is fixed across simulation draws, and thus the same holds for $\max_d \theta^d$, so that average regret is just a convenient renormalization of the average of welfare θ^{d^*}. We also report the share among our simulation draws for which the optimal treatment was chosen after conclusion of the experiment, that is for which regret $= 0$.

Simulation results Tables I through III show our simulation results. Appendix A.3 has figures providing further detail. The tables show the average of regret, and the probability that the optimal policy is chosen (corresponding to regret $= 0$) for each of the four algorithms considered, and for varying numbers of waves, holding total sample size constant.

There are several noticeable patterns across the simulations. First, modified Thompson sampling, as proposed in this paper, consistently performs better than expected Thompson sampling, which performs very similarly to Thompson sampling, and all of these outperform non-adaptive assignment. Second, adaptive designs with more waves consistently outperform designs with fewer waves (for the same total sample size). Third, the gains from adaptive design in terms of average regret are largest in the application to Ashraf et al. (2010), followed by Cohen et al. (2015). The gains for Bryan et al. (2014) are somewhat smaller. The gains in the probability of choosing the optimal treatment are even more pronounced.

The figures in Appendix A.3 reveal the following additional properties of modified Thompson sampling. The probability of choosing the best treatment is strictly larger than under non-adaptive assignment, for every setting considered. More generally, the distribution of regret under modified Thompson sampling first-order stochastically dominates the corresponding distribution under non-adaptive assignment. And lastly, for Ashraf et al. (2010) and Bryan et al. (2014), both approaches pick one of the best two treatments with high probability. For Cohen et al. (2015), the distribution is more dispersed, owing to smaller treatment differences.

TABLE I

ASHRAF, BERRY, AND SHAPIRO (2010)

Statistic	2 waves	4 waves	10 waves
Regret			
modified Thompson	0.0016	0.0010	0.0008
expected Thompson	0.0023	0.0014	0.0014
Thompson	0.0023	0.0014	0.0014
non-adaptive	0.0048	0.0050	0.0051
Share optimal			
modified Thompson	0.978	0.987	0.989
expected Thompson	0.970	0.981	0.982
Thompson	0.969	0.981	0.982
non-adaptive	0.937	0.935	0.933
Units per wave	502	251	100

TABLE II

BRYAN, CHOWDHURY, AND MOBARAK (2014)

Statistic	2 waves	4 waves	10 waves
Regret			
modified Thompson	0.0045	0.0041	0.0038
expected Thompson	0.0048	0.0044	0.0043
Thompson	0.0048	0.0044	0.0043
non-adaptive	0.0054	0.0055	0.0054
Share optimal			
modified Thompson	0.792	0.809	0.822
expected Thompson	0.777	0.798	0.802
Thompson	0.778	0.795	0.801
non-adaptive	0.750	0.747	0.748
Units per wave	935	467	187

TABLE III

COHEN, DUPAS, AND SCHANER (2014)

Statistic	2 waves	4 waves	10 waves
Regret			
modified Thompson	0.0069	0.0063	0.0061
expected Thompson	0.0072	0.0065	0.0061
Thompson	0.0073	0.0065	0.0062
non-adaptive	0.0087	0.0087	0.0086
Share optimal			
modified Thompson	0.565	0.587	0.591
expected Thompson	0.567	0.582	0.593
Thompson	0.563	0.579	0.591
non-adaptive	0.521	0.525	0.528
Units per wave	1080	540	216

6. IMPLEMENTATION IN THE FIELD

Precision Agriculture for Development (PAD) is testing a variety of different interactive voice response treatments in Odisha, India.[1] The purpose of these treatments is to enroll as many rice farmers as possible into Odisha's farmer information service. We designed an experiment in order to help PAD to choose the treatment that works best for their setting.

We tested six distinct treatment arms with different options for prior text messages and call times, cf. Table IV. The outcome is a binary variable describing successful call completion: it equals one if the call recipient answered five questions asked during the call (which enables processing by PAD).

Sample Selection and Treatment Assignment The only information PAD has about farmers are phone numbers provided by the government. PAD processed these numbers and set aside a batch of 10,000 numbers that were verified to be valid and are not on the Indian "do-not-disturb" list. PAD randomly selected 600 phone numbers for each wave. Starting on June 3 2019, experimental waves were carried out consecutively, where each wave took two days to administer (text messages are sent up to 24 hours ahead of time, there are two call times in the morning and the evening on the call day).

We used modified Thompson sampling, as described in Section 3.3, starting with a uniform prior over $\boldsymbol{\theta}$, in order to determine the assignment frequencies for each wave.

Findings [COMING SOON]

TABLE IV

TABULATED DATA

Treatment	Obervations	Successes	Share of successes
no SMS, 10am	743	126	0.170
SMS 1h before, 10am	2373	467	0.197
SMS 24h before, 10am	1319	247	0.187
no SMS, 6:30pm	291	41	0.141
SMS 1h before, 6:30pm	822	141	0.172
SMS 24h before, 6:30 pm	1052	187	0.178

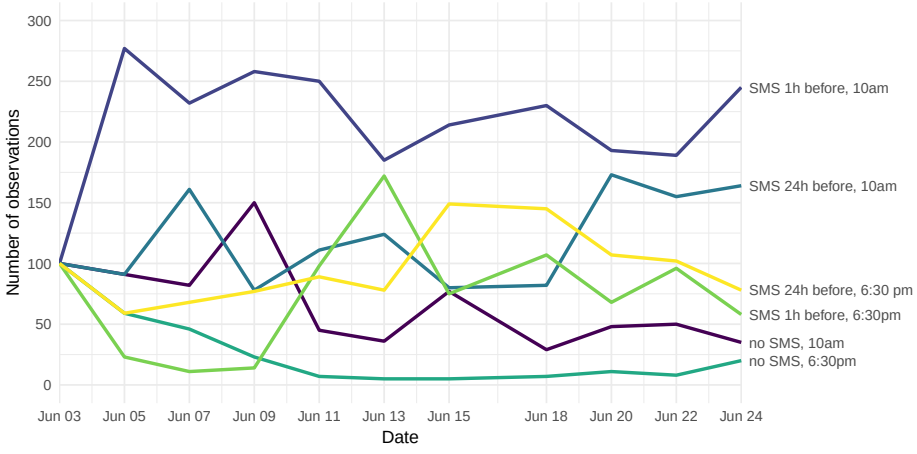
[1]We thank Grady Killeen for helping us implement this experiment.
The pre-analysis plan for this experiment was registered at
https://www.socialscienceregistry.org/trials/4263.
The R-code used for implementation can be found at
https://github.com/maxkasy/Precision-Agriculture-for-Development.

TABLE V

Posterior parameters

Treatment	Mean	Standard dev	Probability optimal
no SMS, 10am	0.170	0.014	0.032
SMS 1h before, 10am	0.197	0.008	0.649
SMS 24h before, 10am	0.188	0.011	0.209
no SMS, 6:30pm	0.143	0.020	0.007
SMS 1h before, 6:30pm	0.172	0.013	0.037
SMS 24h before, 6:30 pm	0.178	0.012	0.066

Figure 6.1.— Assignment frequencies over time.



REFERENCES

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1.

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.

Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, volume 1, pages 73–140. Elsevier.

Berry, D. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery*, 5(1):27–36.

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.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Bubeck, S., Munos, R., and Stoltz, G. (2011). Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852.

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.

Duflo, E. and Banerjee, A., editors (2017). *Handbook of Field Experiments*, volume 1. Elsevier.

Ghavamzadeh, M., Mannor, S., Pineau, J., and Tamar, A. (2015). Bayesian reinforcement learning: A survey. *Foundations and TrendsÂ® in Machine Learning*, 8(5-6):359–483.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Russo, D. (2016). Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418.

Russo, D. J., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. (2018). A Tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Weber, R. et al. (1992). On the Gittins index for multiarmed bandits. *The Annals of Applied Probability*, 2(4):1024–1033.

APPENDIX A: SUPPLEMENTARY APPENDIX

A.1. *Proof of Theorem 1*

Recall that under modified Thompson sampling, a share

$$q_t^d = \frac{p_t^d \cdot (1 - p_t^d)}{\sum_{d'} p_t^{d'} \cdot (1 - p_t^{d'})}$$

of wave t is assigned to treatment d, where p_t^d is the posterior probability that d is optimal.

1. **Each treatment is assigned infinitely often.**
   Suppose otherwise. Then there is some treatment which is only assigned a finite number of times, and is not assigned anymore after some wave t', so that $q_t^d = 0$ for $t > t'$. The posterior probability p_t^d of this treatment being optimal is bounded away from 0 for $t > t'$ by Lemma 14 in Russo (2016).

   Note now that under modified Thompson sampling, the denominator in the expression defining q_t^d is bounded above by 1, and thus the probability q_t^d of being assigned to treatment d is bounded below by $p_t^d \cdot (1 - p_t^d)$. It follows that q_t^d is bounded away from 0 when the same holds for p_t^d. The claim follows by contradiction.

2. **The share of observations assigned to the best treatment converges to** $1/2$ **as** $T \to \infty$**.**
   Since each treatment is assigned infinitely often, we have that $p_t^d \to 1$ for the optimal treatment d, again by Lemma 14 in Russo (2016).

   We can derive upper and lower bounds on q_t^d, for a given value of p_t^d, by considering the maximum and minimum of the expression defining q_t^d with respect to the vector $\boldsymbol{p}_t$, given p_t^d. The denominator of the expression defining q_t^d, $\sum_d p_t^{d'} \cdot (1 - p_t^{d'})$, is concave as a function of the vector $\boldsymbol{p}_t$. The maximum of q_t^d is therefore achieved at a corner of the simplex of possible values for $\boldsymbol{p}_t$ given $\boldsymbol{p}_t$. These corners are such that $p_t^{d'}$ is equal to 0 for all but two values of d'. For such $\boldsymbol{p}_t$ we get $q_t^d = 1/2$, and thus

   $$q_t^d \leq \frac{1}{2}$$

   for all values of $\boldsymbol{p}_t$ and all d.

   Reversely, again by concavity as well as symmetry of the denominator, the minimum of q_t^d with respect to the vector $\boldsymbol{p}_t$, given p_t^d, is achieved when $p_t^{d'}$ is equal to $\frac{1-p_t^d}{k-1}$ for all $d' \neq d$. We therefore get

   $$\begin{aligned} q_t^d &\geq \frac{p_t^d \cdot (1 - p_t^d)}{p_t^d \cdot (1 - p_t^d) + \sum_{d' \neq d} \frac{1-p_t^d}{k-1} \frac{k-2+p_t^d}{k-1}} \\ &= \frac{p_t^d}{p_t^d + \frac{k-2+p_t^d}{k-1}} = \frac{p_t^d}{p_t^d + 1 - \frac{1-p_t^d}{k-1}}, \end{aligned}$$

   where

   $$\frac{p_t^d}{p_t^d + 1 - \frac{1-p_t^d}{k-1}} \geq \frac{p_t^d}{p_t^d + 1} \to \frac{1}{2}$$

   as $p_t^d \to 1$. The claim follows.

3. **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.**
   Consider two treatments d, d'. By definition of q_t^d,

$$q_t^d \leq \frac{p_t^d \cdot (1-p_t^d)}{p_t^{d'} \cdot (1-p_t^{d'})} \leq 4 \frac{p_t^d}{p_t^{d'}},$$

   where the second inequality holds as long as $p_t^{d'} \leq 1/2$. Lemma 13 in Russo (2016) then implies that q_t^d converges to 0 at an exponential rate, along any subsequence of t for which the share of observations assigned to d exceeds the share $\bar{q}^d$.

   By Lemma 12 in Russo (2016), this in turn implies that the share of observations assigned to d has to converge to $\bar{q}^d$.

4. **No other assignment algorithm for which statement 1 holds has average regret going to 0 at a faster rate than modified Thompson sampling.**
   This is an immediate corollary of Proposition 7 Russo (2016), once we note that the rate of convergence of average regret to 0 is the same as the rate of convergence of the probability of choosing a sub-optimal treatment.

□

A.2. *Optimal design in a simple example*

In this section, we discuss optimal experimental designs in a simple example. Suppose we have ten experimental units that we can enroll in two waves. There are three treatments. The cost of all treatments is the same, so we set $\boldsymbol{c} = 0$ for simplicity. We impose a uniform prior for $\boldsymbol{\theta}$.

Dividing the sample between first and second wave. A first question to consider is how to divide the total sample of 10 units between the two waves. For each division $(N_1, 10 - N_1)$ between the two waves, we can calculate expected welfare V_0 at the outset of wave 1, using the value function derived in Section 3.1.

Figure A1 plots expected welfare as a function of the sample size N_1 in wave 1. The boundary cases $N_1 = 0$ and $N_1 = 10$ correspond to an experiment with only one wave. The optimal split assigns either 5 or 6 units to the first wave. Splitting the sample in this manner allows us to learn from the first-wave assignment (e.g. of two units per treatment if $N_1 = 6$) and then focus attention on the treatments with higher values in the second wave.[2]

Assigning treatments. Based on Figure A1, we set $N_1 = 6$, so that we have $N_2 = 4$ in the second wave. Driven by the symmetric prior, it is optimal to assign 2 units to each of the 3 treatments in wave 1. Optimal assignment in wave 2 depends on the outcomes of the first wave. We explore several scenarios in Figure A2. This figure plots expected welfare for any second-wave treatment assignment in the simplex $n_2^1 + n_2^2 + n_2^3 = 4$, conditional on first-wave outcomes. For each scenario, the number of successes in each treatment in the first wave determines the prior for treatment assignments in the second wave. Our uniform prior for $\boldsymbol{\theta}$ implies a Beta posterior, where for $s_1^d \in \{0, 1, 2\}$ we get $\alpha_1^d = 1 + s_1^d$ and $\beta_1^d = 1 + 2 - s_1^d$. This Beta posterior has a mean of $(1 + s_1^d)/4$.

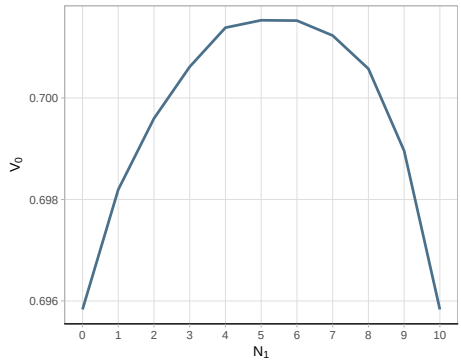
[2]The welfare differences across alternative designs are relatively small in this setting, owing to the small number of units involved. In our simulations calibrated to more realistic settings we fond quantitatively important differences.

Four scenarios The four scenarios we consider are $\boldsymbol{s}_1 = (1, 1, 1)$, $\boldsymbol{s}_1 = (1, 1, 2)$, $\boldsymbol{s}_1 = (1, 1, 0)$, and $\boldsymbol{s}_1 = (2, 2, 0)$. In the first scenario, each treatment had one success and one failure, leading to a posterior that is again symmetric across treatments. In this scenario, shown in the top left of Figure A2, it is optimal to assign 2 units to either of the three treatments, and 1 unit to the other two arms. In the second scenario, treatment 3 performed better than treatments 1 and 2. In this scenario, shown in the top right of Figure A2, it is optimal to assign 3 units to treatment 3, and 1 unit to either of the other two arms. In the third and fourth scenario, treatment 3 performed worse than treatments 1 and 2. In these scenarios, shown in the bottom part of Figure A2, it is optimal to assign no units to treatment 3, 3 units to either of treatment 1 or 2, and 1 to the other. Interestingly, this dominates (though not by much) the assignment of 2 units to each of treatment 1 and 2.

Discussion These examples show that dividing the sample equally between treatment arms is generally not optimal. Moreover, in each example, the largest number of units is assigned to the treatment arms with the highest expected return. This reflects that more precise effect estimates for treatment arms with low expected return are unlikely to affect the ultimate policy decision. This is true even though our objective function does not assign any weight to the welfare of experimental units; there is no exploitation motive.

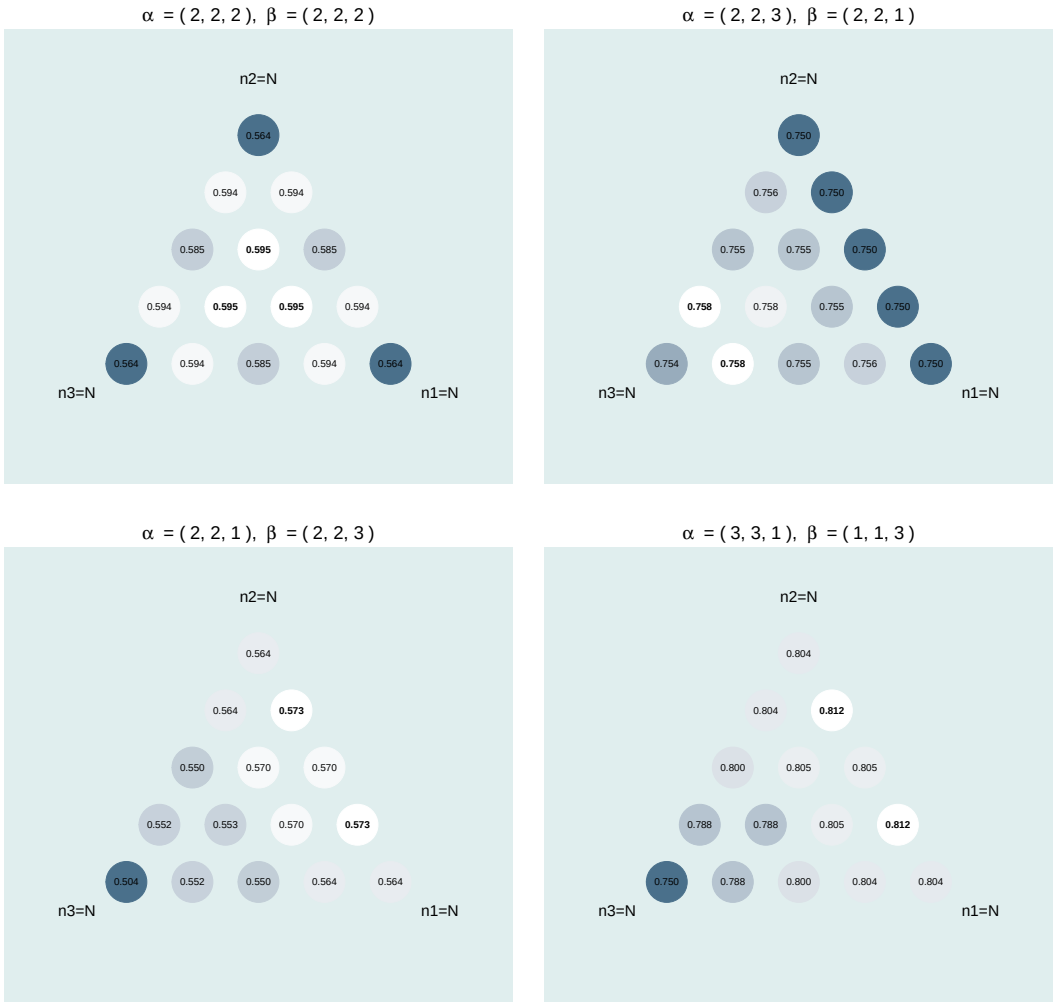
A last observation is that even with symmetric priors, a symmetric assignment is not necessarily optimal. Consider for instance the case $\alpha = (3, 3, 1)$, $\beta = (1, 1, 3)$. The prior distribution for treatments 1 and 2 is the same. The optimal design, however, assigns either more units to treatment 1 or to 2. This reflects a non-convexity in the value of information, due to the concave objective function $\max_d \left(E[\theta^d | \boldsymbol{m}_T, \boldsymbol{s}_T] - c^d\right)$. This situation is analogous to option pricing, where higher volatility can increase the value of a stock option which is only exercised for high profit realizations.

FIGURE A1.— Dividing the sample across waves.



Notes: The graph shows expected welfare V_0 as a function of the sample size N_1 in period 1, assuming a total sample size of 10 and three treatments, for a uniform prior.

FIGURE A2.— Expected welfare as a function of treatment assignment



Notes: This figure shows the expected welfare (action value function) U_2 for each possible treatment assignment $\boldsymbol{n}_2 = (n_2^1 + n_2^2 + n_2^3)$ in wave 2 (which is of size 4), taking as given the Beta-prior parameters $\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1$ which were determined by the outcomes of wave 1 (which is of size 6). For example, the upper right panel is for the case where treatment 1 and 2 each had one success, but treatment 3 had 2 successes. Note that the color scaling differs across the plots for better readability.

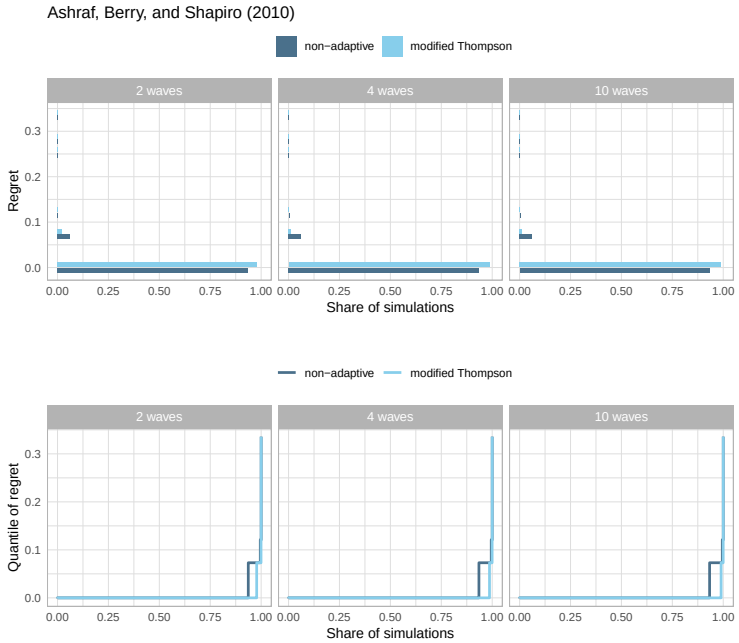
A.3. *Simulations*

Background on the experiments used for calibration Ashraf et al. (2010) conducted a field experiment in Zambia involving Clorin, a disinfectant. During a door-to-door sale of Clorin to about 1,000 households in Lusaka, each participating household was offered a bottle of Clorin for a randomly chosen offer price, at or below the retail price. The treatment in this experiment is the price offered, ranging from 300 to 800 Zambian Kwacha. The outcome is whether the household bought the bottle of Clorin.

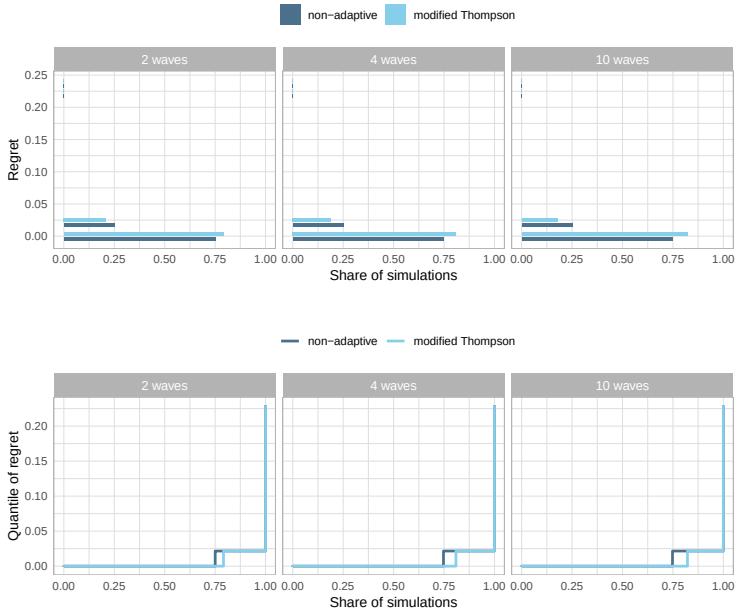
Bryan et al. (2014) conducted a field experiment in rural Bangladesh. Households were randomly assigned a cash or credit incentive of $8.50, conditional on a household member migrating during the 2008 monga (lean) season. This amount covers the round-trip travel cost. The treatments in this experiment are cash, credit, information, and a control group. The outcome is whether at least one household member migrated.

Cohen et al. (2015) conducted a field experiment in three districts of Western Kenya. Households were randomly assigned one of three subsidy levels for the purchase of artemisinin combination therapies (ACT), an antimalarial drug. They were also randomly offered a rapid detection test (RDT) for malaria. The treatments in this experiment are 3 subsidy levels with or without RDT, and a control group. The outcome is whether the household actually bought ACT.

Figures The figures compare the full distribution of regret between *non-adaptive* assignment and *modified Thompson*. They show the probability mass functions (histograms) and the quantile functions of the distribution of regret. Note in particular that a uniformly lower quantile function for modified Thompson sampling, relative to non-adaptive assignment, implies that its distribution of regret is first-order stochastically dominated. The integrated difference between the two quantile functions equals the decrease in average regret (increase in average welfare) that we gain from switching to modified Thompson sampling.



Bryan, Chowdhury, and Mobarak (2014)



Cohen, Dupas, and Schaner (2014)