Adaptive experiments for policy choice (preliminary draft)

Maximilian Kasy* Anja Sautmann†

December 5, 2018

Abstract

The goal of many of field experiments is to evaluate policies and inform future policy choices, but standard experimental designs are geared toward testing hypothesis. We consider the problem of assigning treatments in experiments with multiple waves, where after the experiment we choose among treatments to maximize expected outcomes net of costs. We discuss optimal treatment assignment, as well as a computationally tractable approximation. Optimal designs focus attention on better-performing policy options in later waves. Calibrated simulations demonstrate improvements in welfare, relative to conventional designs. Our setting is related to but different from multi-armed bandit settings.

Keywords: Experimental design, field experiments, optimal policy
JEL Codes: C93, C11, O22

1 Introduction

Field experiments have become part of the standard toolkit of economics (Banerjee et al., 2016). Field experiments can serve various ultimate goals. Two possible goals are the estimation of treatment effects, and testing for the presence of a non-zero treatment effect. Standard recommendations for the design of experiments are based on these goals. These recommendations (see for instance Athey and Imbens 2017) include (i) assigning an equal number of units to different treatments, (ii) if possible within strata defined by pre-determined covariates, and (iii) choosing the sample size based on power calculations for tests of the hypothesis of a zero average treatment effect.

Another important goal is to inform future policy choices. We will show that the objective of informing policy leads to design recommendations that are qualitatively different from standard recommendations. This is in particular the case when (i) there are multiple alternative policies (treatments) that we wish to compare and (ii) the experiment can be run in multiple waves.

Example To fix ideas, consider the example of an NGO that has the goal of improving newborn health by encouraging "kangaroo care" (skin-to-skin care) for prematurely born babies. The NGO knows that kangaroo care, implemented correctly and used consistently, works in principle (Conde-Agudelo et al., 2012). There is, however, a considerable number of implementation choices to be made, which determine the uptake and quality of kangaroo care that is de-facto received, such as setting the right incentives for health-care providers, or educating mothers. How should the NGO make these choices? We argue that the NGO should run an experiment in multiple waves. Initially they should try many different variants. As they learn which options appear to perform better, they

*Department of Economics, Harvard University, maximiliankasy@fas.harvard.edu
†Director of Research, Education, and Training, J-PAL, asautmann@povertyactionlab.org

should focus their experiment on the best performing options. This allows them to learn the optimal option with high certainty. Once the experiment is concluded, they can then recommend the best performing option for large scale implementation. As this example suggests, one way to think about our proposal is that it gives a principled approach for running pilot studies, or "tinkering" with policy options, in the spirit of "the economist as plumber" (Duflo, 2017). We will show that our proposal leads to consistently better policy recommendations, for the same sample size, relative to standard recommendations.

Setup More formally, we consider an experimental setting with multiple waves of experimental units, and multiple treatments. At the beginning of each wave, the number of units assigned to each treatment arm is decided. Then the assignments are implemented for this wave, and outcomes are observed, before treatments are assigned for the next wave. Once the experiment is concluded, one of the treatments is picked for large-scale implementation. The goal is to maximize the average outcomes for this large-scale implementation, net of the costs of treatment. We would like to design experiments that are optimal or close to optimal for this goal.

Note that, while having some commonalities, this setting is different from the well-known "multi-armed bandit" problem (cf. Bubeck and Cesa-Bianchi, 2012; Weber et al., 1992). In particular, there is no "exploitation" motive, and thus no exploitation-exploration tradeoff in our setting.

Optimal assignment The setup just sketched defines a dynamic stochastic optimization problem. The actions in each wave (time period) are treatment assignments, the state space are beliefs over treatment effects, and transitions are determined by experimental outcomes. The last action to take is the choice of policy for large-scale implementation, after which welfare is realized as average outcomes net of costs. This dynamic stochastic optimization problem can, in principle, be solved analytically using backward induction. For some small-scale examples, we will discuss analytic solutions and their qualitative features.In more realistic settings, however, finding exact solutions becomes quickly computationally infeasible, due to exploding state and action spaces.

Modified Thompson sampling These computational constraints motivate the use of approximate solutions that are computationally feasible. We will in particular propose the following assignment algorithm, which is a modified version of so-called Thompson sampling (Russo et al., 2018). For standard Thompson sampling, each unit is assigned to a given treatment with probability equal to the posterior probability (given past outcomes) that it is in fact the optimal treatment. This prescription is easy to implement, by sampling just one draw of the parameter vector from the posterior, and picking the optimal treatment corresponding to this parameter vector.

We modify this prescription in two ways. First, the same treatment may not be assigned twice in a row. Second, we apply Thompson sampling (with our first modification) repeatedly for each wave, and set the shares of units assigned to each treatment equal to the average shares across replicate draws. These modifications relative to standard Thompson sampling are motivated by the facts that (i) our setup does not have an "exploitation" motive, and (ii) we have waves, rather than units arriving sequentially.

Our baseline model does not take into account covariates. After discussing this baseline model, we consider covariates for stratified assignment algorithms and targeted treatment assignment policies. We propose, in particular, to use an "empirical Bayes" approach (Morris, 1983) in order to form posterior beliefs about the treatment effectiveness for each treatment arm and stratum. One of the advantages of this approach is that it avoids requiring the specification of prior distributions for the purpose of experimental design.

Simulation evidence We provide extensive simulation evidence on the performance of alternative assignment algorithms.We evaluate these algorithms in terms of the welfare they generate, via

informing a policy choice. Our simulations use parameters and sample sizes calibrated to data from published experiments in development economics (Ashraf et al., 2010; Bryan et al., 2014; Cohen et al., 2015).

Several patterns are confirmed by these simulations. Modified Thompson sampling consistently performs better than standard Thompson sampling, which in turn outperforms conventional non-adaptive designs. That is, the distribution of welfare (across simulations) under modified Thompson sampling stochastically dominates the alternatives. In particular, modified Thompson sampling generates the highest average welfare, and the highest probability of picking the optimal treatment. The gains from adaptive designs are larger when the experiment is divided into more waves, for the same total sample size. The same patterns hold when considering assignment stratified on covariates, and corresponding targeted treatment assignment policies.

Implementation in practice We next discuss a field experiment for which we implemented our recommended approach. [THIS IS WORK IN PROGRESS.]

Related literature The idea of adaptive experimentation is almost as old as the idea of randomized experiments; see for instance Thompson (1933). Adaptive experimental designs have been used in clinical trials (Berry, 2006), and in the targeting of online advertisements (Russo et al., 2018). By contrast, they have not entered the standard toolkit for field experiments in economics, see e.g. Duflo and Banerjee (2017).

The idea of adaptive experimental design has produced a large theoretical and practical literature, focused in particular on the so-called multi armed bandit problem. We will discuss the differences between our setting and the bandit problem in Section 2.2. Under some conditions (separability across treatment arms, infinite horizon), 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, rather than solving for the optimal assignment, heuristic algorithms are used far more commonly. Two popular algorithms are the Upper Confidence Bound algorithm (UCB), and Thompson sampling, cf. Russo et al. (2018). A large theoretical literature characterizes the expected regret of these algorithms, cf. Bubeck and Cesa-Bianchi (2012). Generalizations of the Bandit problem are discussed under the name of reinforcement learning in the machine learning literature, cf. Ghavamzadeh et al. (2015).

Roadmap The rest of this paper is structured as follows. In Section 2, we introduce our formal setup and solve for optimal treatment assignments. In Section 3 we discuss Thompson sampling and modified Thompson sampling. In Section 4 we consider optimal experimental designs in the context of simple examples. In Section 5, we provide simulation evidence on the relative performance of adaptive algorithms, using parameters calibrated to data from published field experiments. In Section 6, we discuss inference for adaptive designs. In Section 7 we present the results of a field experiment where we implemented our approach, demonstrating its practical feasibility. [TBD] Section 8 concludes. The supplementary appendix also provides additional simulation results and examples of optimal assignments.

2 Setup and optimal treatment assignment

Consider the following experimental setting. The experiment proceeds in multiple waves, and after conclusion of the experiment a policymaker observes the experimental outcomes and picks the policy maximizing expected welfare. The experimenter wishes to help the policymaker by designing an experiment which is maximally useful for policy choice. She thus aims to assign treatments so as to also maximize the policymaker's expected welfare. At the end of each experimental wave outcomes are observed, and treatment assignment in subsequent waves can be based on these observations.

The experimenter's problem thus takes the form of a dynamic stochastic optimization problem, which can in principle be solved using backward induction.

Treatments and potential outcomes The experiment takes place in waves $t = 1, \ldots, T$. In wave t there are N_t experimental units $i = 1, \ldots, N_t$. There is no overlap between successive waves, so that we have repeated cross-sections rather than panel data.

There is a treatment $D_{it} \in \{1, \ldots, k\}$ and an outcome $Y_{it} \in \{0, 1\}$. Outcomes Y_{it} are determined by the potential outcome equation

$$Y_{it} = \sum_{d=1}^{k} \mathbf{1}(D_{it} = d) Y_{it}^d.$$

This assumption implies in particular that there is no interference, i.e., outcomes are not affected by other units' treatment. We focus on binary outcomes in this paper for ease of exposition, and since it is the leading case for the applications considered below.

The vectors $(Y_{it}^0, \ldots, Y_{it}^k)$ are i.i.d. draws from the population of interest, across both i and t. Denote

$$\theta^d = E[Y_{it}^d], \qquad n_t^d = \sum_i \mathbf{1}(D_{it} = d), \qquad s_t^d = \sum_i \mathbf{1}(D_{it} = d, Y_{it} = Y_{it}^d = 1).$$

Thus θ^d is the average potential outcome for treatment value d (also known as average structural function), n_t^d is the number of units assigned to treatment d in wave t, and s_t^d is the number of "successes" (outcome $Y_{it} = 1$) among those in treatment group d in wave t.

Treatment assignment and state space Treatment assignment in wave t can depend on the outcomes of waves $1, \ldots, t-1$ and on a randomization device. Treatment assignment can be summarized by the vector

$$\boldsymbol{n}_t = (n_t^1, \ldots, n_t^k),$$

collecting the number of units assigned to each of the treatments $d = 1, \ldots, k$ in wave t, where $\sum_d n_t^d = N_t$. The experimenter's problem is to choose $\boldsymbol{n}_t$ at the beginning of wave t. The outcomes of wave t can be summarized by

$$\boldsymbol{s}_t = (s_t^1, \ldots, s_t^k),$$

collecting the number of "successes" among the units assigned to each of the treatments in wave t, where $s_t^d \leq n_t^d$. These outcomes are observed at the end of wave t, before treatment assignment takes place in wave $t + 1$.

Denote the cumulative versions of these terms by

$$m_t^d = \sum_{t' \leq t} n_{t'}^d \qquad r_t^d = \sum_{t' \leq t} s_{t'}^d$$
$$\boldsymbol{m}_t = (m_t^1, \ldots, m_t^k) \qquad \boldsymbol{r}_t = (r_t^1, \ldots, r_t^k),$$

and $M_t = \sum_{t' \leq t} N_{t'}$. 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, while M_t is the total number of units in waves 1 through t. Since we assumed i.i.d. potential outcomes across waves, all relevant information for the experimenter at the beginning of period $t + 1$ is summarized by $\boldsymbol{m}_t$ and $\boldsymbol{r}_t$. Put differently, $\boldsymbol{r}_t$ is a sufficient statistic for the parameter vector $\boldsymbol{\theta}$.

Policy choice and welfare Once the experiment is completed, after wave T, a policy $d^* \in 1, \ldots, k$ will be chosen. This policy will be implemented for everyone in the population of interest. The policy objective is to maximize the expected average of the outcome Y, net of the cost of treatment. The posterior expected social welfare of policy d, after completion of the experiment, is given by

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

where c^d is the unit cost of implementing policy d. $SW(d)$ is expressed in per-capita terms, so that the size of the population of interest does not matter. The optimal policy choice after the experiment is given by

$$d^\star = \underset{d}{\text{argmax}}\ SW(d).$$

Note also that, by assumption, social welfare does not include the outcomes of participants in the experiment. This is justified if the the number of participants is small relative to the size of the population of interest for which the policy will be implemented.

Bayesian prior and posterior Under our assumptions, $Y^d \sim Ber(\theta^d)$. Assume that the prior distribution of θ^d is given by

$$\theta^d \sim Beta(\alpha_0^d, \beta_0^d),$$

and that the θ^d are mutually independent across d. A special case, and the default for our simulations and 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.

The posterior distribution, after outcomes for periods $1, \ldots, t$ are realized, is then given by

$$\begin{aligned} \theta^d | \boldsymbol{m}_t, \boldsymbol{r}_t \sim Beta(\alpha_t^d, \beta_t^d) \qquad \alpha_t^d &= \alpha_{t-1}^d + s_t^d \qquad \beta_t^d = \beta_{t-1}^d + n_t^d - s_t^d \\ &= \alpha_0^d + r_t^d \qquad = \beta_0^d + m_t^d - r_t^d, \end{aligned}$$

and in particular

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

The belief of the experimenter at the beginning of period t about the outcomes of wave t, given her treatment assignment decision $\boldsymbol{n}_t$, reflects two sources of uncertainty: Her 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 is given by Binomial distributions with parameters n_t^d and θ^d. Integrating out the unknown parameter $\boldsymbol{\theta}$, we get that the experimenter's ex-ante belief about the outcomes of wave t follow a Beta-Binomial distribution,

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

2.1 Optimal assignment

This setting yields a dynamic stochastic optimization problem, which we consider next. We can, in principle, solve for optimal treatment assignments using backward induction. In practice, calculating optimal assignments is computationally very demanding; this is even more so in the context of the extensions considered below. These computational challenges motivate consideration of non-optimal but tractable alternatives below.

Value functions and backward induction The state at the end of period t is given by $(\boldsymbol{m}_t, \boldsymbol{r}_t)$. The action in period t is given by $\boldsymbol{n}_t$. The transitions between states are described by $\boldsymbol{m}_t = \boldsymbol{m}_{t-1} + \boldsymbol{n}_t$, $\boldsymbol{r}_t = \boldsymbol{r}_{t-1} + \boldsymbol{s}_t$ and $P(s_t^d = s|\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, n_t^d)$ given by Equation (1). The optimal experimental design can be derived using backward induction.

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 end of the experiment. V_t is a function of the state $(\boldsymbol{m}_t, \boldsymbol{r}_t)$. Starting at the end, we have

$$V_T(\boldsymbol{m}_T, \boldsymbol{r}_T) = \max_d \left(E[\theta^d|\boldsymbol{m}_T, \boldsymbol{s}_T] - c^d\right) = \max_d \left(\frac{\alpha_0^d + r_T^d}{\alpha_0^d + \beta_0^d + m_T^d} - c^d\right). \tag{2}$$

Denote by U_t expected welfare at the beginning of wave t, assuming that treatment assignment in wave t is based on $\boldsymbol{n}_t$, while all future treatment assignment decisions will be optimal,

$$U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) = E\left[V_t\left(\boldsymbol{m}_{t-1} + \boldsymbol{n}_t, \boldsymbol{r}_{t-1} + \boldsymbol{s}_t\right)|\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t\right],$$

where the expectation is taken over the Beta-Binomial distribution of Equation (1). Then we get that the period t value function and the optimal experimental design satisfy

$$\begin{aligned} V_{t-1}(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}) &= \max_{\boldsymbol{n}_t:\, \sum_d n_t^d \leq N_t} U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) \\ \boldsymbol{n}_t^*(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}) &= \operatorname*{argmax}_{\boldsymbol{n}_t:\, \sum_d n_t^d \leq N_t} U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t). \end{aligned} \tag{3}$$

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

Choosing sample size and number of waves The same value functions can be used not only to choose the treatment assignment vectors $\boldsymbol{n}_t$, they can also be used to choose other features of the experimental design, including the number and size of waves. For any choice of T and $(N_1, \ldots, N_T)$ we can, in principle, calculate expected welfare V_0. V_0 is increasing in each N_t, and increases when subdividing the same total sample size into more waves. Balancing these increases in expected welfare against the cost of experimentation yields optimal choices of sample size and the optimal number of waves. This is analogous to the power calculations commonly performed by researchers running field experiments.

Computational challenges Based on Equation (3), we can in principle solve for the optimal treatment assignment by brute-force enumeration of all possible outcomes and actions, since both have finite support. We will do so below in the context of illustrative examples with a small number of units, two waves, and three treatments. These examples will help build intuition for the form that the optimal assignment takes.

For more realistic settings, however, with larger sample sizes and a greater number of treatments, solving for the optimal assignment quickly becomes infeasible. This is due to the fast growth of the required number of calculations. To see this, consider the problem of calculating $\boldsymbol{n}_t^*$, assuming knowledge of V_t. For each choice of $\boldsymbol{n}_t$, evaluating U_t requires enumerating all possible realizations of $\boldsymbol{s}_t$, of which there are $\prod_d n_t^d = O(N_t^k)$ for a typical choice of $\boldsymbol{n}_t$. For each realization of $\boldsymbol{s}_t$ we need to evaluate the Beta-Binomial likelihood, and the value function V_t. In addition, in order to maximize over all possible $\boldsymbol{n}_t$, we need to calculate $U - t$ for each $\boldsymbol{n}_t$. There are $\binom{N_t+k-1}{k-1} = O(N_t^k)$ such possible assignment vectors.[1] We thus get that as N_t gets large the required computation time is of order $O(N_t^{2k})$ times the computation time for V_{t+1} and for the Beta-Binomal likelihood. This contrasts with the methods discussed in Section 3 below, for which computation time is only of order $O(N_t)$.

[1] This is known as the "number of weak compositions of N_t into k parts."

Table 1: Two decision problems

	Estimation	**Policy choice**
Action	Estimate $\hat{\boldsymbol{\theta}}$	Policy $d^* = \text{argmax}_{\,d}\, E[\theta^d \vert \boldsymbol{m}_T, r_T]$
Loss function	$\sum_d (\hat{\theta}^d - \theta^d)^2$	$-\theta^{d^*}$
Risk function	$\sum_d \text{Var}_\theta(\hat{\theta}^d) + \text{Bias}^2_\theta(\hat{\theta}^d)$	$-E_\theta[\theta^{d^*}]$
Bayes risk	$\sum_d E[(\hat{\theta}^d - \theta^d)^2]$	$-E[\max_d E[\theta^d \vert \boldsymbol{m}_T, r_T]]$

A large literature in structural econometrics considers methods for numerically finding approximate solutions to dynamic stochastic optimization problems, see for instance Judd (1998). Popular iteration-based methods depend on stationary settings to approximate value functions; such methods are not applicable in our setting since we are interested in a finite-horizon problem. More relevant in our case are approaches based on (i) interpolation, (ii) simulation, and (iii) restricted action spaces. Simulation-based methods, in particular, can control computation time by replacing full enumeration over all possible $\boldsymbol{n}_t$ and $\boldsymbol{s}_t$ by a random sample of fixed size for either vector. Appendix A.1 provides some discussion of these methods in the context of our setting.

2.2 Comparison to alternative experimental design problems

Estimation versus policy choice We have set up the experimental design problem as a decision problem, where the ultimate goal is to choose a policy in order to maximize expected outcomes. It is useful to contrast this setting with a more commonly considered one, where the goal of experimental design is to obtain precise estimates (of average potential outcomes or treatment effects). Table 1 provides such a comparison, where we assume for simplicity that $\boldsymbol{c} = 0$. In both cases, we wish to take an action after conclusion of the experiment, based on the outcome of the experiment. In the case of estimation, the action is given by the vector of estimates $\hat{\boldsymbol{\theta}}$, which might for instance be obtained by simple sample averages, or as a Bayesian posterior mean. In the case of policy choice, the action is given by d^*.

A common way to evaluate estimators is based on quadratic error loss. If we take the expectation of loss given the true parameter $\boldsymbol{\theta}$ we obtain the risk function. For unbiased estimators this risk function is given by the estimator variance. Averaging over a prior distribution for $\boldsymbol{\theta}$ gives Bayes risk. Analogously, the risk function for policy choice is given by $-E_\theta[\theta^{d^*}]$, the expected average outcome achieved by the chosen policy d^*. If we subtract the welfare of the oracle-optimal choice, we obtain $\text{Regret} = \max_d \theta^d - \theta^{d^*}$; taking expectations we obtain expected regret, which is a convenient re-centering of the risk function.

It is useful to keep the analogy between estimator variance and expected regret in mind for our subsequent discussions. The difference between these two objectives (minimizing variance versus minimizing expected regret) drives the difference in the resulting design recommendations. The analogy between them is useful when considering practical questions such as the choice of sample size.

Bandit problems versus policy choice Another experimental design problem that has received much attention is the multi-armed bandit problem. In its basic form, the multi-armed bandit problem can be described as follows. There are several treatments. A sequence of units to be treated arrives, one by one. Outcomes are observed for each unit before the next unit is treated. The experimenter cares about the outcomes (welfare) of the treated units themselves. This results in a tradeoff between exploitation (assigning the best possible treatment, given current knowledge, to the current unit) and exploration (learning which treatment is best, so as to make better choices for units arriving in the future).

The setup we consider in this paper differs in a number of important ways from the basic bandit problem. First, we consider units arriving in waves ("batched" assignment, in machine learning terminology), rather than sequentially. This is a more common setting in economic field experiments; sequential arrival is more common in clinical or online settings. Second, we consider a finite horizon setting, where the experiment stops after a pre-determined number of waves. This precludes characterizations using the Gittins index or similar devices, which rely on a stationary infinite horizon setting.

Third, and most importantly, we consider a different objective function. We consider a policy to be chosen after conclusion of the experiment, and want to maximize welfare for this policy choice. In our setting there is therefore no exploration-exploitation trade-off (though we could easily incorporate one). A common intuition for Bandit problems suggests that exploitation favors treatments with the best past outcomes, while exploration suggests to spread assignment across different treatment values. As our analysis shows, however, even with a pure exploration motive, there is a strong rationale to focus on the best performing treatments. This conclusion hinges on the ultimate objective of our analysis, maximizing welfare for a policy choice. This conclusion would not hold true for more conventional objectives such as the mean squared error of treatment effect estimators, or the power of statistical tests for no treatment effects.

3 Modified Thompson sampling

An alternative to full optimization is the use of simpler algorithms that approximate the optimal solution. Such approximate algorithms are widely used in the context of online dynamic experiments, for instance in the placement of ads. One of the most popular (and oldest) such algorithms is so-called Thompson sampling. Originally conceived by Thompson (1933) in the context of clinical trials, Thompson sampling can be defined as follows.

Thompson sampling Consider a special case of the setting described in Section 2 above where each wave is of size 1, so that units arrive sequentially (and we can drop the subscript i). In each period t, assign any treatment d with probability equal to the posterior probability, given past outcomes, that it is in fact the optimal treatment,

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

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

$$D_t = \underset{d}{\text{argmax}}\ (\hat{\theta}_t^d - c^d).$$

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 data likelihoods. An excellent overview can be found in Russo et al. (2018).

Modified Thompson sampling We can improve on standard Thompson sampling in our context. We propose the following two modifications. These modifications improve performance in our simulations. The **first modification** we propose stipulates that the same treatment is not to be assigned twice in a row,[2] In the setting with multiple observations i for wave t, this can be

[2]Similar modifications were discussed by Russo (2016), who considers the problem of best arm identification in the bandit setting.

implemented by sampling a draw $\hat{\theta}_{it}$ from the posterior given $\boldsymbol{m}_{t-1}$ and $\boldsymbol{r}_{t-1}$, and setting

$$D_{it} = \underset{d \neq D_{(i-1)t}}{\text{argmax}} \, (\hat{\theta}_{it}^{d} - c^{d}).$$

The **second modification** we propose applies Thompson sampling (with our first modification) repeatedly for each wave, and sets $\boldsymbol{n}_t$ to the averages of the corresponding frequencies across replicate Thompson draws, up to rounding. In our simulations and applications we use 10 replicate assignments for each wave. When we refer to "modified Thompson sampling" in the rest of this paper, it is understood that we mean these two modifications.

3.1 Justifications

Justifications of Thompson sampling in the bandit setting Thompson sampling is usually motivated by reference to the multi-armed bandit model. This model is similar to the one introduced in Section 2, except for the following differences: (i) It assumes that units arrive sequentially, corresponding to waves of size 1. (ii) It usually assumes an infinite horizon setting. (iii) Most importantly, the objective function in the multi-armed bandit problem is to maximizes the outcomes of the experimental units themselves, rather than to optimize the post-experimental policy choice, as in our setting. Treatment assignment in the bandit setting therefore involves a trade-off between exploitation (achieving good outcomes for the current unit) and exploration (learning about the right policy for future units).

In the context of the multi-armed bandit model, strong theoretical justifications exist for Thompson-sampling, see for instance Bubeck and Cesa-Bianchi (2012).[3] These justifications are based on characterizations of regret, where regret in this context is defined as the difference between the average outcome for the optimal treatment, $\max_d \theta^d$, and the average outcome for the actually assigned treatments, $\frac{1}{T}\sum_{t=1}^{T} \theta^{D_t}$. Thompson sampling achieves a regret of order $\log(T)/T$, as shown by Agrawal and Goyal (2012), and this rate can not be improved upon.

What are the features of Thompson sampling (and other related algorithms) which allow to achieve this rate? Intuitively, any algorithm which might hope to perform well has to avoid two opposite pitfalls. The first pitfall is to concentrate too soon on the treatment which appears to perform best. If this is done too soon, there is some chance that the wrong treatment is chosen and repeated many times. The second pitfall is to concentrate too late. In this case, too many units are assigned to sub-optimal treatments, even though we already know that they are sub-optimal. As it turns out, Thompson sampling achieves the right midpoint (in terms of rates) between these two pitfalls.

Justification of Thompson sampling in our setting Perhaps surprisingly, even though these justifications are based on the bandit setting, Thompson sampling also performs very well for our objective of post-experimental policy choice, as shown by the simulations presented in Section 5. The reason is that, even without an exploitation motive, optimal treatment assignment policies in our model focus on the best-performing options, to tease out which among these is in fact the best, rather than wasting experimental units on treatments that are clearly sub-optimal. This is confirmed by our examples of optimal assignments in Section 4 below.

Put differently, we would like to have sufficient power at the end of the experiment to reject all sub-optimal alternatives. Assume $\boldsymbol{c} = 0$ for simplicity, and denote by $\Delta^d = \max_{d'} \theta^{d'} - \theta^d$ the difference in expected outcomes between treatment d and the optimum. The choice of treatment d as a policy after the experiment yields a regret of Δ^d, which is in particular decreasing linearly in Δ^d. The power to reject treatment d as sub-optimal, on the other hand, is also decreasing in Δ^d.

[3] Similar justifications exist for other algorithms which we won't discuss here, including the Upper Confidence Bound algorithm, for appropriately chosen tuning parameters.

To equate power across suboptimal d, we would have to assign a number of units proportional to $(1/\Delta^d)^2$ to each of them. Given that regret is smaller for small Δ^d, we don't want to be quite as extreme, but still assign more units to treatments with smaller Δ^d. This is exactly what Thompson sampling achieves, in expectation.

Justification of modified Thompson sampling We can improve on standard Thompson sampling in our context, using our proposed modifications. The first modification we proposed above is motivated by the fact that we have no exploitation motive. This modification requires that the same treatment is not assigned twice in a row. To justify this modification, consider Thompson sampling in large samples. In large samples we would assign units to the optimal treatment with high probability, relative to its close competitors. But that means we stop learning about the performance of its competitors. To achieve an optimal rate of learning, we would instead assign the same number of observations to the closest competitor (while neglecting all worse treatments). Avoiding repeated assignment to the same treatment achieves this goal.

The second modification we proposed is motivated by the fact that in our setting units arrive in waves, rather than sequentially. The proposed modification applies Thompson sampling (with our first modification) repeatedly for each wave, and sets $\boldsymbol{n}_t$ to the averages of the corresponding frequencies across replicate Thompson draws. To justify this modification, note that it corresponds to a "Rao-Blackwellization" of Thompson sampling, reducing randomness and thereby increasing expected welfare.

3.2 Covariates, targeted assignment, and hierarchical priors

The model introduced in Section 2 did not involve any covariates. Suppose now that we additionally observe a covariate X_i with finite support for each unit i, before assigning a treatment D_i. Assume additionally that the optimal treatment assignment policy chosen by the policymaker at the end of the experiment might also condition treatment on X, assigning $d^*(x)$ to units characterized by $X = x$. We discuss three approaches to treatment assignment in this context, (i) considering each stratum as a separate experiment, (ii) (modified) Thompson sampling based on a hierarchical prior, and (iii) (modified) Thompson sampling based on an empirical Bayes approach. We ultimately recommend the latter, which does not require specification of prior parameters, but makes good use of cross-stratum information.

Considering each stratum as separate experiment How does this affect our analysis? One possibility would be to simply treat each stratum, defined by a value of X, as a separate experiment. Within strata, the previous analysis then applies almost verbatim. Note, however, we might not know in advance the covariate values for future waves. This implies that the strata-specific sample sizes are random ex-ante, thus adding an additional dimension of uncertainty to our optimization problem. Additionally, and more interestingly, we might want to leverage information from some strata in order to learn something about average potential outcomes for other strata. For example, if a medical treatment works well for patients aged 50-60 years, that might suggest that it also works well for patients aged 60-70 years.

Hierarchical priors Formally, this consideraton translates into a prior with statistical dependence between the vectors of average potential outcomes $\boldsymbol{\theta}^x = (\theta^{1x}, \ldots, \theta^{kx})$ across different strata x. One natural way to model such dependence is by constructing a hierarchical model, cf. Chapter 5 in Gelman et al. (2014).

Consider for instance the following generalization of the setting considered thus far. Recall that x indexes strata, and d denotes treatment values. Let θ^{dx} be the corresponding average potential

outcome. Assume

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

where π is some prior distribution for (α^d, β^d), and parameters are independent across treatment arms. It follows immediately that s_t^{dx}, the number of successes for treatment d and stratum x, follows a Beta-Binomial distribution given (α^d, β^d) and n_t^{dx}.

Intuitively, updating based on this prior works as follows. For each treatment d, consider the success rates s_t^{dx} across different strata x. Based on these success rates, learn the mean and dispersion of θ^{dx} across strata, as reflected in (α^d, β^d). Then use these as a prior, in conjunction with s_t^{dx} for a given stratum x, to learn about θ^{dx} for that stratum.

More formally, and in a slight change of notation relative to our baseline model, denote by $\boldsymbol{\theta}, \boldsymbol{m}_t, \boldsymbol{r}_t$ the vectors of parameters, cumulative trials and successes indexed by both d and x. We can sample from the posterior of $\boldsymbol{\theta}$ given $\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}$ as follows:

1. First, for each d draw a sample of (α^d, β^d) from the posterior based on the Beta-Binomial likelihood for the observed $\boldsymbol{r}_{t-1}^d$.

2. For each draw of (α^d, β^d), sample a draw of $\boldsymbol{\theta}^d$ from its Beta posterior given (α^d, β^d) and given $\boldsymbol{m}_{t-1}^d, \boldsymbol{r}_{t-1}^d$.

Based on the resulting draw of $\boldsymbol{\theta}$, Thompson sampling can be implemented by setting

$$D_{it} = \underset{d}{\text{argmax}} \, (\hat{\theta}_t^{dX_{it}} - c^d).$$

Similarly, modified Thompson sampling can be implemented by imposing that (i) the same treatment is not assigned twice in a row within strata, and (ii) setting assignment frequencies n_t^{dx} equal to their average over multiple replicate draws of Thompson sampling.

Empirical Bayes A third possibility, which is closely related to the hierarchical Bayes approach, is the use of an empirical Bayes approach. As before, assume that

$$\begin{aligned} Y_{it}^d | X_{it} = x, \theta^{dx}, (\alpha_0^d, \beta_0^d) &\sim Ber(\theta^{dx}) \\ \theta^{dx} | (\alpha_0^d, \beta_0^d) &\sim Beta(\alpha_0^d, \beta_0^d). \end{aligned}$$

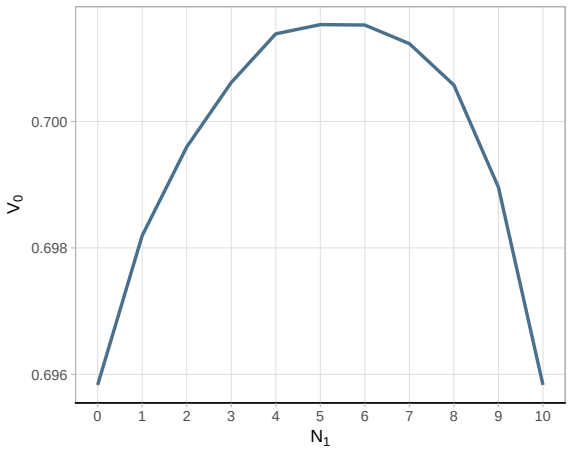
For the empirical Bayes approach, we do not require a prior distribution over (α_0^d, β_0^d). Instead, we generate "posterior" draws as follows:

1. First, for each d estimate (α^d, β^d) using **maximum likelihood**, based on the Beta-Binomial likelihood for the observed $\boldsymbol{r}_{t-1}^d$.

2. For each draw of (α^d, β^d), sample a draw of $\boldsymbol{\theta}^d$ from its Beta posterior given (α^d, β^d) and given $\boldsymbol{m}_{t-1}^d, \boldsymbol{r}_{t-1}^d$.

Then proceed as before. This is the approach we actually use in our simulations and applications using covariates below.

Figure 1: Dividing the sample across waves



Notes: This figure shows the 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.

4 Optimal design in a simple example

In this section, we consider optimal experimental designs in a simple example, in order to build intuition and to provide some additional motivation for our proposed modified Thompson sampling procedure. Suppose that we are in the setting of Section 2, with three treatments and two waves, and with a uniform prior for $\boldsymbol{\theta}$. Suppose that there are 10 units total, across the two waves, and that the cost of all treatments is the same, so that we can set $\boldsymbol{c} = 0$ for simplicity.

Dividing the sample between first and second wave The first question to consider in this experiment is how to divide the total sample of 10 units between the two waves. The value function derived in Section 2.1 provides an answer. For each division $(N_1, 10-N_1)$ between the two waves, we can calculate expected welfare, assuming that (i) the optimal policy is chosen after the experiment, and (ii) treatment assignment is optimal in both waves.

Figure 1 plots expected welfare as a function of the sample size N_1 in period 1. The boundary cases $N_1 = 0$ and $N_1 = 10$ correspond to having an experiment with only one wave. As can be seen, the optimal split assigns either 6 or 5 units to the first wave, and the remainder to the second wave. This results in higher expected welfare relative to the boundary cases, demonstrating the benefit of adaptive experimentation. Splitting the sample in two waves allows to re-optimize the design in the second wave, focusing attention on the potentially optimal treatment values.

Assigning treatments We next consider the optimal assignment of treatments in either wave. Based on Figure 1, set $N_1 = 6$, and correspondingly assume that we have $N_2 = 4$ units in the second wave. Treatment assignment in the first wave is straightforward. Driven by the symmetry of our setting, it is optimal to assign 2 units to each of the 3 treatment arms.

Treatment assignment in the second wave is more interesting. Optimal assignment depends on the outcomes of the first wave. We explore several scenarios in Figure 2.[4] 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

[4]The figures in Appendix A.3 explore similar scenarios for different sizes of wave N_2.

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 $(1+s_1^d)/4$.

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 2, 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 2, it is optimal to assign 3 units to treatment 3, and 1 unit to either of the other two arms.

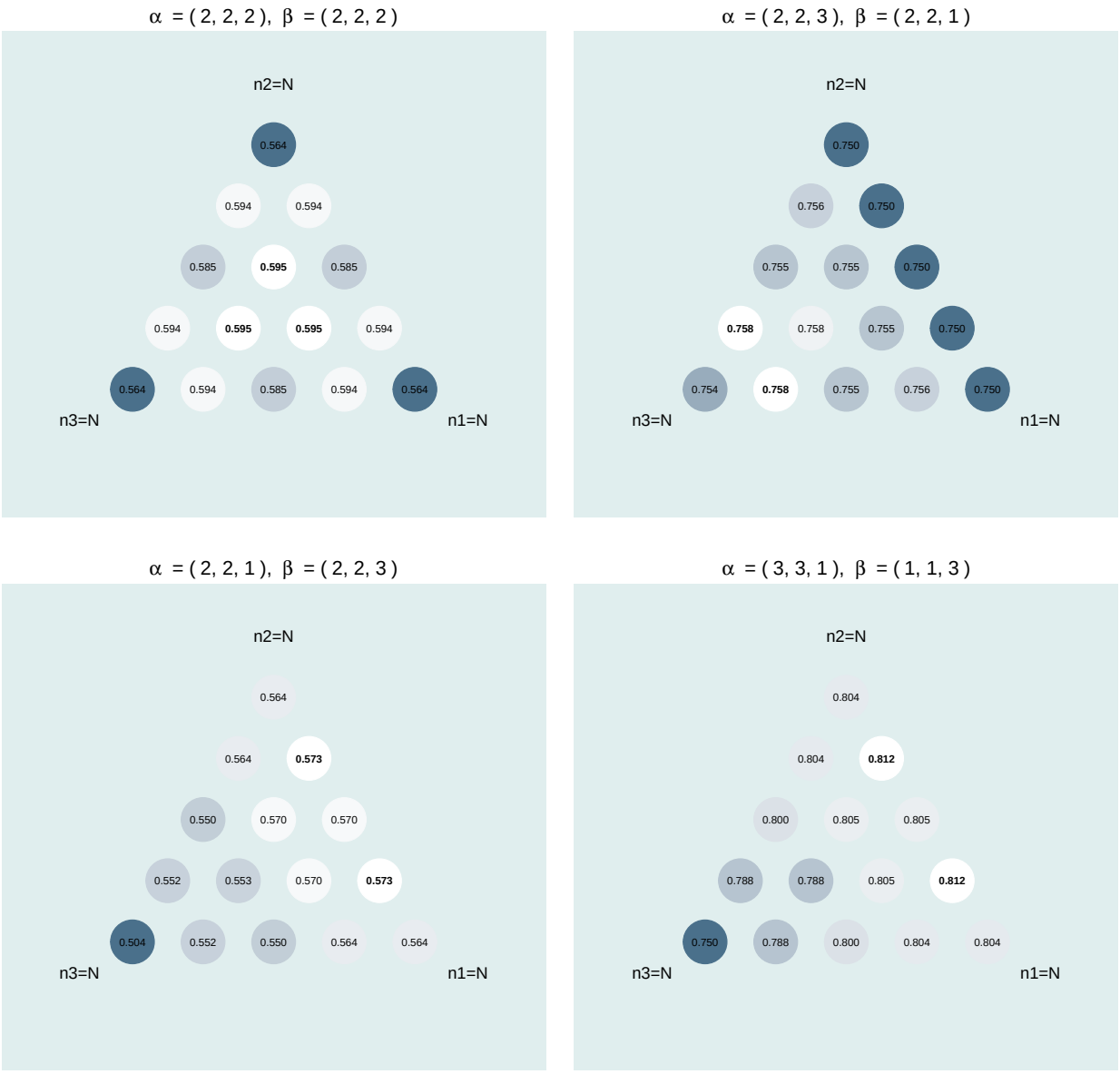
In the third and fourth scenario, treatment 3 performed worse than treatments 1 and 2. In these scenarios, shown in the bottom part of Figure 2, it is optimal to assign no units to treatment 3, 3 units to either of treatment 1 or 2, and 1 to the other. Interestingly, this dominates (though not by much) the assignment of 2 units to each of treatment 1 and 2.

Takeaways What are the messages that emerge from this discussion? We would like to emphasize three points. First, dividing the sample equally between treatment arms is in general not optimal, and can lead to important welfare losses relative to optimal treatment assignment. This is in particular the case when the prior means differ across arms.

Second, the largest number of treatments should be assigned to the treatment arms with the highest expected return. The reason is *not* that we care about the outcomes of experimental units (our objective function does not assign any weight to their welfare). Rather, this recommendation reflects the fact that for treatment arms which have little chance of being optimal, getting more precise estimates of their effect has little chance of affecting our ultimate policy decision. This feature of optimal treatment assignments is approximated by Thompson sampling and modified Thompson sampling.

Third, and for related reasons, but maybe somewhat counter-intuitively, even with symmetric priors a symmetric assignment is not necessarily optimal. Consider for instance the case $\alpha = (3,3,1)$, $\beta = (1,1,3)$. In this case, treatment 3 has the lowest prior mean, while the prior distribution for treatments 1 and 2 is the same. The optimal treatment assignment, however, assigns either more units to 1 or to 2. This reflects a non-convexity in the value of information, due to the concave objective function $\max_d \left(E[\theta^d|\boldsymbol{m}_T, \boldsymbol{s}_T] - c^d\right)$ (the concavity comes from considering the maximum). This situation is formally analogous to option pricing, where higher volatility can increase the value of a stock-option.

Figure 2: Expected welfare as a function of treatment assignment



Notes: This figure shows the expected welfare U_2 as a function of treatment assignment $\boldsymbol{n}_2 = (n_2^1 + n_2^{@} + n_2^3)$ in wave 2 (which is of size 4), taking as given the Beta-prior parameters $\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1$ determined by the outcomes of wave 1 (which is of size 6). Note that the color scaling differs across the plots for better readability.

5 Calibrated simulations

We next present simulation evidence on the performance of alternative treatment assignment algorithms, using parameter vectors $\boldsymbol{\theta}$ and sample sizes M_T 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 thus 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).

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.

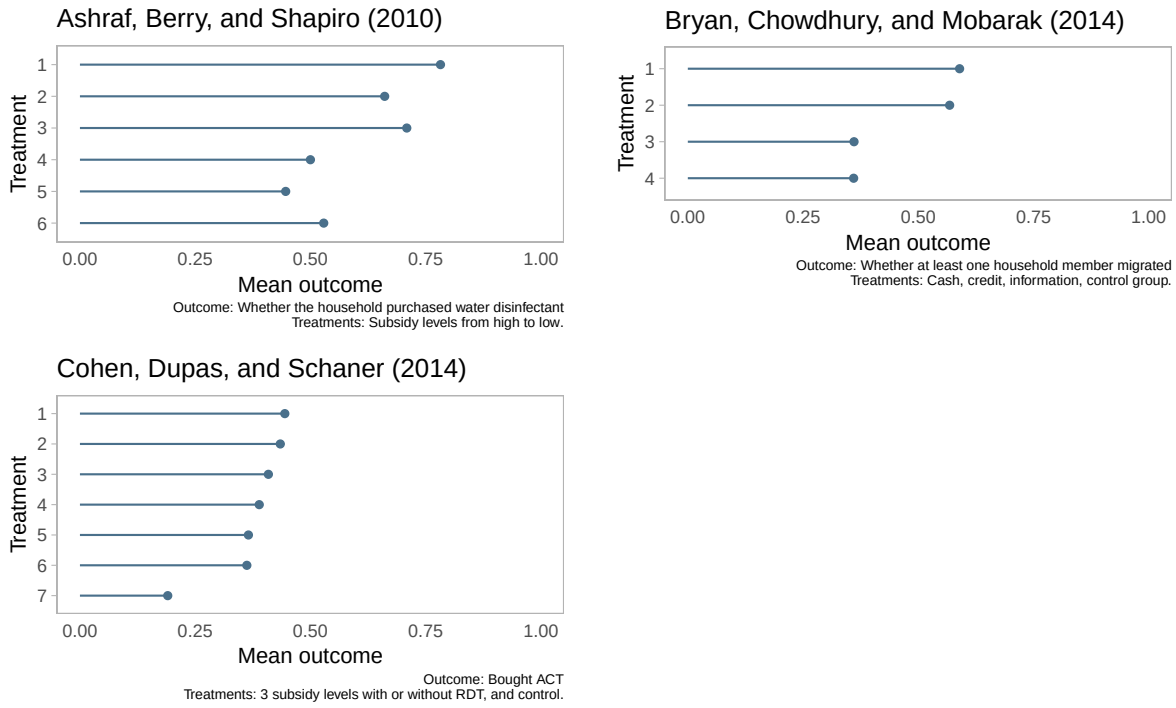
5.1 No covariates

We first consider simulations without covariates, corresponding to the setting discussed in Section 2. Throughout, we make two simplifying assumptions, to tie our hands and keep our analysis transparent. 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, and that the cost of each treatment is the same, w.l.o.g. $\boldsymbol{c} = 0$. This is of course unrealistic, but allows us to avoid ad-hoc assumptions regarding $\boldsymbol{c}$, and to focus on the benefit of adaptive assignment for a range of parameter vectors $\boldsymbol{\theta}$.

Calibrated parameter values Figure 3 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 ranging from .44 to .78 across 6 treatments. 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 with average outcomes of about .36, and two better treatments that are very close, with average outcomes of .57 and .59. 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, with average outcomes for these ranging from .36 to .44. This setting is similar to Ashraf et al. (2010), except that the best treatments are closer and thus harder to distinguish.

Figure 3: Average outcomes across treatment arms in published experiments



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. This is the conventional recommendation for experimental design. The second algorithm uses a simple heuristic to approximate features of the optimal assignment as discussed in Sections 2 and 4. This algorithm ranks treatments in terms of the posterior mean of their performance, ignores the worse half of treatments, and divide units equally between the better half of treatments. The third and fourth algorithm are Thompson sampling and modified Thompson sampling, as introduced in Section 3 above.

Performance criteria We evaluate the performance of these algorithms in terms of two statistics. The first statistic that we report is average regret, across 10,000 simulation draws. Since we set $\boldsymbol{c} = 0$ (or constant across treatment arms), 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,

$$d^* = \underset{d}{\text{argmax}} \; E[\theta^d | \boldsymbol{m}_T, \boldsymbol{r}_T],$$
$$\text{Regret} = \max_d \theta^d - \theta^{d^*}.$$

For each of our simulations the vector $\boldsymbol{\theta}$ is fixed, and thus the same holds for $\max_d \theta^d$, so that (in this context) average regret is just a convenient renormalization of the average of welfare θ^{d^*}. The second statistic that we report is the share among 10,000 simulation draws for which the optimal treatment was chosen after conclusion of the experiment, that is for which Regret $= 0$.

Simulation results Table 2 shows our simulation results for these settings. This table is based on total sample sizes equal to the original experiments. Tables A1 and A2 in the Appendix provide similar results for total sample sizes equal to half the original, and equal to 1.5 times the original sample size. There are several noticeable patterns across these simulations.

- The first pattern is that modified Thompson sampling, as proposed in this paper, consistently performs better than Thompson sampling, which in turn dominates the "best half" heuristic, which outperforms the non-adaptive design.
- The second pattern is that adaptive designs with more waves consistently outperform designs with fewer waves (for the same total sample size).
- The third pattern is that the gains from adaptive design 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.

Figure 4 provides more detail for some of these simulations; Figure A1 in the Appendix does the same for sample size half the original. Figure 4 compares the distribution of regret between non-adaptive assignment and our preferred method, modified Thompson sampling. Each treatment value d^*, chosen as the policy after the end of the experiment, corresponds to a value of regret of $\max_d \theta^d - \theta^{d^*}$. The oracle-optimal treatment in particular results in regret equal to 0. The dots in the figure show the probability of each treatment being chosen as policy under modified Thompson sampling; the other end of the corresponding lines show the probability of being chosen under non-adaptive assignment. This figure reveals 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.
- 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.

5.2 Targeting based on covariates

We next turn to simulations using covariates. Covariates used both for experimental treatment assignment, and for targeting of d^*, corresponding to the setting discussed in Section 3.2. We calibrate both the distribution across strata p_x and the conditional average potential outcomes θ^{dx} to the data.

Calibrated parameter values We choose the following covariates in order to define strata for targeting. Our choice of these covariates was guided by the requirement of observing units for each combination of covariate and treatment in the original experiments, so that we can calibrate θ^{dx} to observed averages. For Ashraf et al. (2010), we set X equal to a location indicator, corresponding to one of five low-income peri-urban areas in Lusaka. For Bryan et al. (2014), we set X equal to an indicator for literacy. For Cohen et al. (2015), we set X equal to an indicator for quartiles of distance to clinic.

Figure A2 in the Appendix shows the average outcomes for each treatment and covariate value across the applications considered. We again set the vectors $\boldsymbol{\theta}$ equal to these average outcomes, for the purpose of our simulations. This figure also shows the number of observations in each of the strata, which we use to calibrate the probabilities p_x of each covariate value in the population.

Algorithms We compare three algorithms. The first algorithm is non-adaptive assignment, stratified on covariates. In each wave and each stratum, as defined by a value of the covariates, we assign an equal share of units to each of the treatment arms. The second algorithm is Thompson sampling, using the empirical Bayes posterior discussed at the end of Section 3.2. We constrain the hyperparameters (α^d, β^d) to be less or equal than 20, thus limiting the influence that the hyperparameters have relative to the actual observations, for any given treatment and covariates combination. The third algorithm is modified Thompson sampling, using the empirical Bayes posterior, where alternation of treatments is imposed within each stratum in each wave.

Performance criteria The policies considered now target treatment assignment based on covariates, so that a unit with covariate value x is assigned to $d^*(x)$. Correspondingly, regret in this setting is defined via

$$d^*(x) = \underset{d}{\text{argmax}}\ E[\theta^{dx}|\boldsymbol{m}_T, \boldsymbol{r}_T] - c^d,$$
$$\text{Regret} = \sum_x p_x \left(\max_d \theta^{dx} - \theta^{d^*(x)} \right),$$

where p_x is the probability of covariate value x in the population, and $\boldsymbol{m}_T, \boldsymbol{r}_T$ are the cumulative trials and successes for all covariate and treatment combinations. We again report average regret across simulations, as well as the share of simulations for which the optimal policy function $d^*(\cdot)$ was chosen, that is for which Regret $= 0$.

Simulation results Table 3 shows our simulation results for these settings. This table is based on total sample sizes equal to the original experiments. Tables A3 and A4 in the Appendix again provide similar results for total sample sizes equal to half the original, and equal to 1.5 times the original sample size.

We would like to point out the following patterns.

- Again, modified Thompson sampling consistently performs better than Thompson sampling, which in turn dominates non-adaptive stratified assignment.
- Regret is larger than in the corresponding settings without covariates. This is owed to the fact that we are faced with the harder problem of figuring out the optimal treatment for each stratum, rather than just the unconditionally optimal treatment.
- Correspondingly, the share of simulations for which the optimal policy was chosen is smaller.
- There appears to be some return to more waves, but it is smaller than in the unconditional setting. We conjecture that this would be different for larger total sample size.

Table 2: Average regret and share optimal, calibrated parameter values, original sample size

2 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive	0.005	0.005	0.009
Regret, best half	0.003	0.004	0.007
Regret, Thompson	0.003	0.005	0.007
Regret, modified Thompson	0.001	0.004	0.007
Share optimal, non-adaptive	0.929	0.748	0.525
Share optimal, best half	0.965	0.802	0.560
Share optimal, Thompson	0.963	0.776	0.548
Share optimal, modified Thompson	0.981	0.800	0.571
Units per wave	502	935	1080
Number of treatments	6	4	7

4 waves

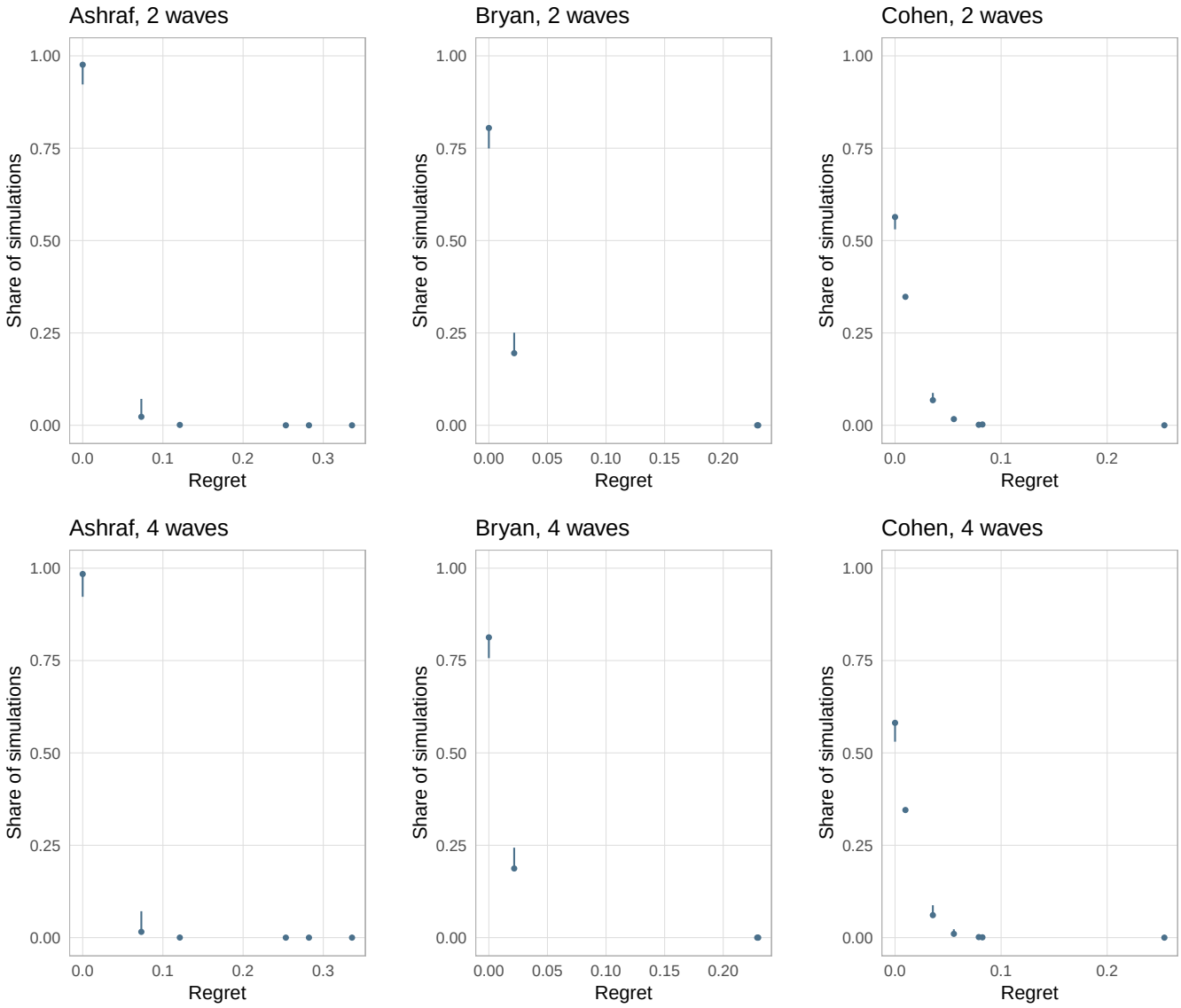
Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive	0.005	0.005	0.009
Regret, best half	0.002	0.004	0.007
Regret, Thompson	0.002	0.005	0.007
Regret, modified Thompson	0.001	0.004	0.007
Share optimal, non-adaptive	0.929	0.767	0.525
Share optimal, best half	0.977	0.794	0.555
Share optimal, Thompson	0.977	0.787	0.578
Share optimal, modified Thompson	0.985	0.810	0.563
Units per wave	251	467	540
Number of treatments	6	4	7

10 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive	0.004	0.005	0.009
Regret, best half	0.002	0.004	0.007
Regret, Thompson	0.002	0.004	0.006
Regret, modified Thompson	0.001	0.004	0.006
Share optimal, non-adaptive	0.942	0.748	0.525
Share optimal, best half	0.975	0.832	0.551
Share optimal, Thompson	0.979	0.810	0.593
Share optimal, modified Thompson	0.989	0.808	0.602
Units per wave	100	187	216
Number of treatments	6	4	7

Notes: This table shows average regret and the share of replications for which the optimal treatment was chosen across 10,000 simulation replications. Parameters are calibrated based on the data of published experimental studies, as shown in Figure 3.

Figure 4: Distribution of regret, original sample size



Notes: These figures compare the distribution of regret across 10,000 simulation replications for non-adaptive assignment and for modified Thompson sampling for samples equal to the original size, as in Table 2. The dot marks the share of simulations for which modified Thompson sampling yielded the corresponding regret, the other end of each line marks the share for non-adaptive assignment.

Table 3: Targeting on covariates: Average regret and share optimal, calibrated parameter values, original sample size

2 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive stratified	0.015	0.007	0.020
Regret, Thompson	0.011	0.006	0.018
Regret, modified Thompson	0.011	0.005	0.017
Share optimal, non-adaptive stratified	0.143	0.630	0.060
Share optimal, Thompson	0.207	0.698	0.060
Share optimal, modified Thompson	0.201	0.727	0.068
Units per wave	502	935	1080
Number of treatments	6	4	7
Number of strata	5	2	4

4 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive stratified	0.015	0.007	0.020
Regret, Thompson	0.010	0.005	0.017
Regret, modified Thompson	0.010	0.004	0.016
Share optimal, non-adaptive stratified	0.145	0.635	0.061
Share optimal, Thompson	0.206	0.721	0.071
Share optimal, modified Thompson	0.226	0.758	0.074
Units per wave	251	467	540
Number of treatments	6	4	7
Number of strata	5	2	4

10 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive stratified	0.015	0.007	0.020
Regret, Thompson	0.009	0.005	0.017
Regret, modified Thompson	0.009	0.004	0.016
Share optimal, non-adaptive stratified	0.146	0.647	0.058
Share optimal, Thompson	0.213	0.723	0.073
Share optimal, modified Thompson	0.243	0.781	0.073
Units per wave	100	187	216
Number of treatments	6	4	7
Number of strata	5	2	4

Notes: This table shows average regret and the share of replications for which the optimal targeted treatment policy was chosen across 10,000 simulation replications. Parameters are calibrated based on the data of published experimental studies, as shown in Figure A2 in the appendix.

Table 4: Inference for adaptive designs: A simple example

P	Y_1	Y_2	$\bar{Y}$	$\hat{\theta}^B$
$(1-\theta)^2$	0	0	0	1/3
$\theta(1-\theta)$	0	1	0	1/3
$\theta(1-\theta)$	1	0	1/2	1/2
θ^2	1	1	1	3/4
$E[\cdot\|\theta]$	θ	θ	$1/2 \cdot \theta(1+\theta)$	$1/3 + \theta/6 + \theta^2/4$
$E_{\theta\sim U[0,1]}[\cdot]$	1/2	1/2	5/12	1/2

6 Inference for adaptive designs

In this section we discuss inference in adaptive designs. Our discussion applies to the various adaptive algorithms considered above, including optimal assignment, best-half assignment, Thompson sampling, and modified Thompson sampling.

We will show, using a simple example, that adaptive treatment assignment biases conventional estimators, such as sample means, and corresponding tests. We then show that Bayesian inference, however, is un-affected by adaptivity. We can simply ignore adaptive assignment when calculating posterior distributions. We lastly discuss how to perform randomization tests for the null of no treatment effects. To do so, one needs to simply re-run the treatment assignment algorithm multiple times, leaving observed outcomes unchanged, to generate a randomization distribution of arbitrary test statistics.

Inference versus welfare maximization Our discussion of inference stands somewhat outside the framework introduced in Section 2. In this framework, the optimal policy d to choose at the end of the experiment is the one that maximizes the posterior expectation of θ^d, net of costs c^d. Quantifications of uncertainty have no further bearing on this choice. In practice, however, experiments might have secondary purposes in addition to informing policy choice. In particular, providing a plausible range of values (for instance, a confidence set) is central for academic publication of experimental results.

A minimal example Why does adaptivity matter for inference? Consider the following minimal example of an adaptive experiment. Suppose that we observe a binary random variable $Y_1 \sim Ber(\theta)$. If Y_1 is 0, we stop the experiment, if $Y_1 = 1$ we continue, and obtain another (independent) draw $Y_2 \sim Ber(\theta)$. Suppose an analyst ignores the fact that the experiment had a data-dependent stopping rule, based on which we decided whether to observe Y_2. Would this analyst draw correct conclusions?

Consider the sample mean $\bar{Y}$, which is equal to Y_1 if only one draw was observed (i.e., if $Y_1 = 0$), and equal to $(Y_1 + Y_2)/2$ if two draws were observed (i.e., if $Y_1 = 1$). What is the expectation of this sample mean? There are four possible combinations of values for (Y_1, Y_2), where Y_2 is only observed when $Y_1 = 1$. The combination $(0, 0)$, for instance, has probability $(1-\theta)^2$, and yields $\bar{Y} = Y_1 = 0$. Table 4 shows the remaining calculations which yield $E[\bar{Y}|\theta] = 1/2 \cdot \theta(1+\theta)$. In particular, whenever $\theta < 1$ we have that $\bar{Y}$ is downward-biased as an estimator of θ! This example illustrates that standard frequentist inference will not be valid for adaptive designs, without modification. Such a downward bias of the sample mean similarly holds for adaptive designs such as Thompson sampling: In these designs more observations for a particular treatment are obtained, on average, whenever that treatment performed well in earlier waves.

How about Bayesian inference? As it turns out, Bayesian inference ignoring the fact that sampling

was adaptive remains valid. To illustrate, assume that we start with a uniform (i.e., $Beta(1,1)$) prior for θ. Then the posterior mean $\hat{\theta}^B$ for θ is equal to $\frac{1+Y_1}{2+1}$ if one draw of Y_1 is observed, and equal to $\frac{1+Y_1+Y_2}{2+2}$ if two draws are observed. This posterior mean is in fact the same whether or not the we take into account that observability of Y_2 depended on the realization of Y_1! In particular, the prior mean of $\hat{\theta}^B$ is indeed equal to $1/2$, the prior mean of θ. This contrasts with the prior mean of $\bar{Y}$, which is equal to $5/12$.

General validity of standard Bayesian inference This validity of standard Bayesian inference is not specific to this example. Bayesian inference that ignores adaptivity remains in fact correct in the context of any adaptive experimental design setting. This follows from the following simple proposition.

Assumption 1 *Consider a set of units $i = 1, 2, \ldots, N$ with potential outcomes $(Y_i^1, \ldots, Y_i^k)$. Denote the likelihood of Y_i^d, indexed by the parameter vector $\boldsymbol{\theta}$, by $f_{\boldsymbol{\theta}}^d(Y_i^d)$.*
Let $H_i = (D_1, \ldots, D_{i-1}, Y_1, \ldots, Y_{i-1})$ be the history of treatments and outcomes before i. Suppose that the treatment assigned to unit i is a function of H_i, as well as possibly a randomization device U_i.[5]

Proposition 1 *Under Assumption 1, the likelihood of $(D_1, \ldots, D_N, Y_1, \ldots, Y_N)$ equals*

$$\prod_{i=1}^{N} f_{\boldsymbol{\theta}}^{D_i}(Y_i),$$

up to a constant that does not depend on $\boldsymbol{\theta}$.

Proof: This follows immediately from the factorization

$$P(D_1, \ldots, D_N, Y_1, \ldots, Y_N | \boldsymbol{\theta}) = \prod_{i=1}^{N} P(D_i | H_i, \boldsymbol{\theta}) \cdot P(Y_i | D_i, H_i, \boldsymbol{\theta}),$$

where $P(D_i|H_i, \boldsymbol{\theta})$ does not depend on θ, and $P(Y_i|D_i, H_i, \boldsymbol{\theta}) = f_{\boldsymbol{\theta}}^{D_i}(Y_i)$. □

Construction of valid randomization tests A popular method for inference in randomized experiments are randomization tests (also known as permutation tests). Consider again the setting of Assumption 1, and the following null hypothesis.

H0 1 *For each unit i, the potential outcomes are the same for all treatments,*

$$Y_i^d = Y_i \quad \forall\, i \,\forall\, d.$$

A randomization test in this setting can be constructed by simulating treatment vectors $\tilde{D}_i$ using a scheme that iterates from $i = 1 \ldots n$, drawing $\tilde{U}_i$ from the same distribution as U_i, defining

$$\tilde{H}_i = (\tilde{D}_1, \ldots, \tilde{D}_{i-1}, Y_1, \ldots, Y_{i-1}),$$

and setting

$$\tilde{D}_i = d_i(\tilde{H}_i, \tilde{U}_i).$$

The following proposition is immediate.

Proposition 2 *Consider the setting of Assumption 1.*

1. *Under the null hypothesis 1, $\tilde{H}_n$ has the same distribution as H_n, conditional on $(Y_1, \ldots, Y_n)$.*
2. *Let $T(H_n)$ be any test statistic, and let T^α be the $(1-\alpha)$ quantile of the distribution of $T(\tilde{H}_n)$ over draws of $\tilde{U}_1, \ldots, \tilde{U}_n$. Then a test which rejects iff $T(H_n) > T^\alpha$ controls size at level α.*

[5] This setting includes as a special case treatment assignment in waves, and binary outcomes, as in Section 2.

7 Implementation in the field

8 Conclusion

A Supplementary Appendix

A.1 Numerical methods for approximating the optimal assignment

A.2 Additional simulation results

Table A1: Average regret and share optimal, calibrated parameter values, half of original sample size

2 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive	0.014	0.007	0.014
Regret, best half	0.008	0.006	0.012
Regret, Thompson	0.008	0.006	0.012
Regret, modified Thompson	0.007	0.006	0.011
Share optimal, non-adaptive	0.829	0.685	0.428
Share optimal, best half	0.901	0.741	0.475
Share optimal, Thompson	0.904	0.708	0.475
Share optimal, modified Thompson	0.912	0.717	0.493
Units per wave	251	467	540
Number of treatments	6	4	7

4 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive	0.012	0.007	0.014
Regret, best half	0.007	0.005	0.011
Regret, Thompson	0.006	0.006	0.011
Regret, modified Thompson	0.007	0.006	0.011
Share optimal, non-adaptive	0.850	0.682	0.428
Share optimal, best half	0.912	0.751	0.461
Share optimal, Thompson	0.918	0.724	0.500
Share optimal, modified Thompson	0.912	0.726	0.494
Units per wave	125	233	270
Number of treatments	6	4	7

10 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive	0.012	0.007	0.014
Regret, best half	0.008	0.005	0.012
Regret, Thompson	0.006	0.006	0.011
Regret, modified Thompson	0.006	0.006	0.010
Share optimal, non-adaptive	0.850	0.681	0.428
Share optimal, best half	0.901	0.762	0.457
Share optimal, Thompson	0.927	0.733	0.492
Share optimal, modified Thompson	0.926	0.731	0.496
Units per wave	50	93	108
Number of treatments	6	4	7

Notes: This table shows average regret and the share of replications for which the optimal treatment was chosen across 10,000 simulation replications. Parameters are calibrated based on the data of published experimental studies, as shown in Figure 3.

Table A2: Average regret and share optimal, calibrated parameter values, 1.5 times original sample size

2 waves

Statistic	Ashraf	Bryan	Cohen
regret, non-adaptive	0.002	0.004	0.006
regret, best half	0.001	0.003	0.006
regret, Thompson	0.001	0.004	0.005
regret, modified Thompson	0.001	0.003	0.005
share optimal, non-adaptive	0.972	0.793	0.599
share optimal, best half	0.990	0.847	0.594
share optimal, Thompson	0.990	0.835	0.625
share optimal, modified Thompson	0.992	0.839	0.606
units per wave	753	1402	1620
number of treatments	6	4	7

4 waves

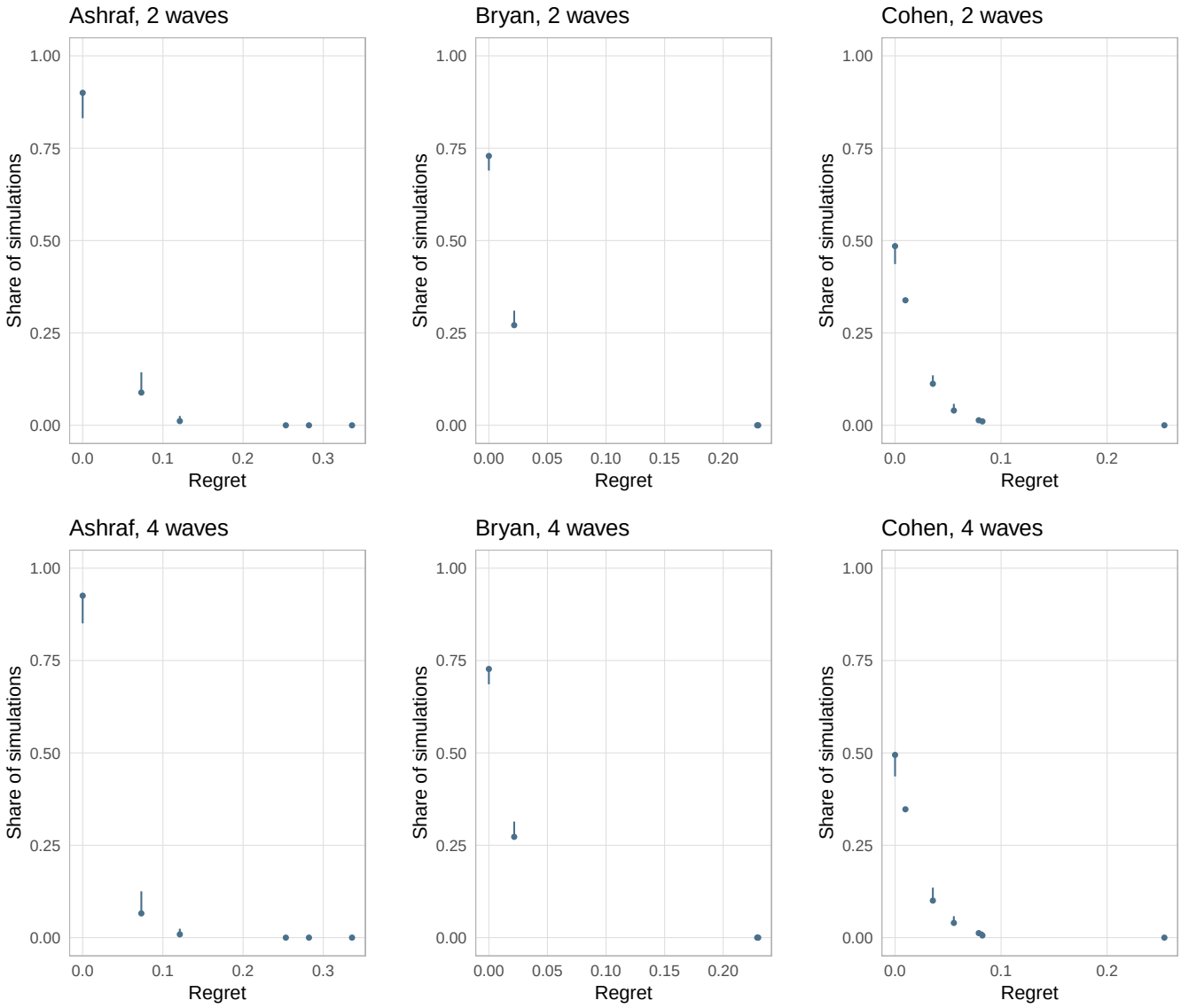
Statistic	Ashraf	Bryan	Cohen
regret, non-adaptive	0.002	0.004	0.006
regret, best half	0.001	0.003	0.005
regret, Thompson	0.000	0.003	0.005
regret, modified Thompson	0.000	0.003	0.004
share optimal, non-adaptive	0.972	0.793	0.599
share optimal, best half	0.993	0.867	0.633
share optimal, Thompson	0.995	0.845	0.626
share optimal, modified Thompson	0.997	0.858	0.675
units per wave	376	701	810
number of treatments	6	4	7

10 waves

Statistic	Ashraf	Bryan	Cohen
regret, non-adaptive	0.003	0.004	0.006
regret, best half	0.000	0.003	0.005
regret, Thompson	0.000	0.003	0.005
regret, modified Thompson	0.000	0.003	0.004
share optimal, non-adaptive	0.962	0.809	0.599
share optimal, best half	0.994	0.874	0.611
share optimal, Thompson	0.996	0.846	0.625
share optimal, modified Thompson	0.998	0.872	0.648
units per wave	150	280	324
number of treatments	6	4	7

Notes: This table shows average regret and the share of replications for which the optimal treatment was chosen across 5000 simulation replications. Parameters are calibrated based on the data of published experimental studies, as shown in Figure 3.

Figure A1: Distribution of regret, half of original sample size



Notes: These figures compare the distribution of regret across 10,000 simulation replications for non-adaptive assignment and for modified Thompson sampling for samples half the original size, as in Table A1. The dot marks the share of simulations for which modified Thompson sampling yielded the corresponding regret, the other end of each line marks the share for non-adaptive assignment.

Figure A2: Average outcomes across treatment arms and strata in published experiments

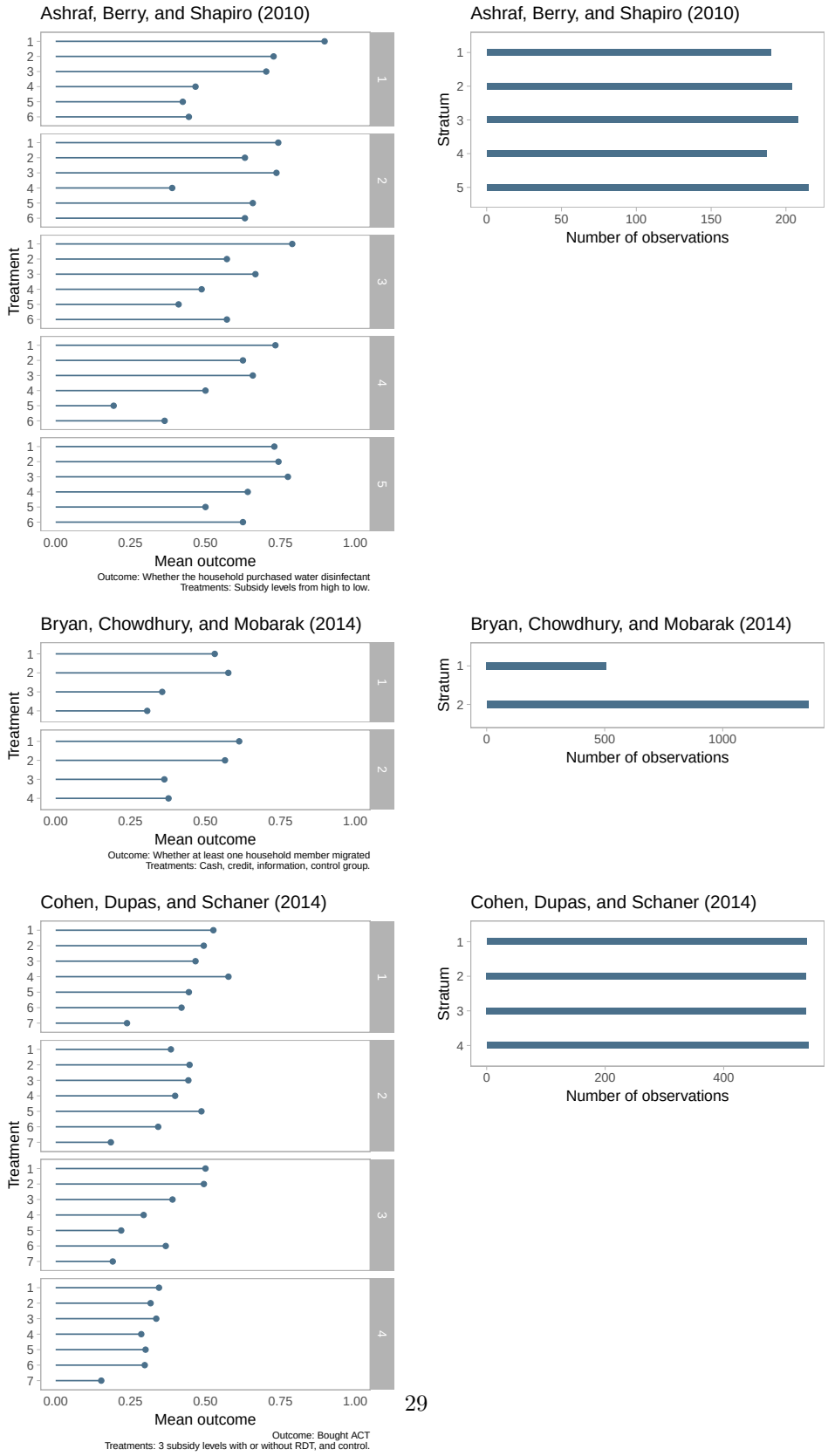


Table A3: Targeting on covariates: Average regret and share optimal, calibrated parameter values, half of original sample size

2 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive stratified	0.023	0.012	0.029
Regret, Thompson	0.018	0.010	0.026
Regret, modified Thompson	0.018	0.009	0.026
Share optimal, non-adaptive stratified	0.083	0.477	0.027
Share optimal, Thompson	0.114	0.540	0.033
Share optimal, modified Thompson	0.117	0.567	0.030
Units per wave	251	467	540
Number of treatments	6	4	7
Number of strata	5	2	4

4 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive stratified	0.024	0.011	0.029
Regret, Thompson	0.017	0.009	0.025
Regret, modified Thompson	0.017	0.008	0.025
Share optimal, non-adaptive stratified	0.086	0.488	0.031
Share optimal, Thompson	0.125	0.558	0.040
Share optimal, modified Thompson	0.122	0.587	0.033
Units per wave	125	233	270
Number of treatments	6	4	7
Number of strata	5	2	4

10 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive stratified	0.023	0.012	0.029
Regret, Thompson	0.017	0.010	0.025
Regret, modified Thompson	0.016	0.008	0.024
Share optimal, non-adaptive stratified	0.091	0.486	0.026
Share optimal, Thompson	0.130	0.530	0.035
Share optimal, modified Thompson	0.137	0.615	0.038
Units per wave	50	93	108
Number of treatments	6	4	7
Number of strata	5	2	4

Notes: This table shows average regret and the share of replications for which the optimal targeted treatment policy was chosen across 10,000 simulation replications. Parameters are calibrated based on the data of published experimental studies, as shown in Figure A2.

Table A4: Targeting on covariates: Average regret and share optimal, calibrated parameter values, 1.5 times original sample size

2 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive stratified	0.011	0.005	0.016
Regret, Thompson	0.008	0.004	0.014
Regret, modified Thompson	0.008	0.003	0.013
Share optimal, non-adaptive stratified	0.207	0.737	0.085
Share optimal, Thompson	0.255	0.779	0.099
Share optimal, modified Thompson	0.263	0.815	0.104
Units per wave	753	1402	1620
Number of treatments	6	4	7
Number of strata	5	2	4

4 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive stratified	0.011	0.005	0.016
Regret, Thompson	0.007	0.003	0.013
Regret, modified Thompson	0.007	0.003	0.012
Share optimal, non-adaptive stratified	0.209	0.747	0.080
Share optimal, Thompson	0.269	0.817	0.107
Share optimal, modified Thompson	0.287	0.832	0.116
Units per wave	376	701	810
Number of treatments	6	4	7
Number of strata	5	2	4

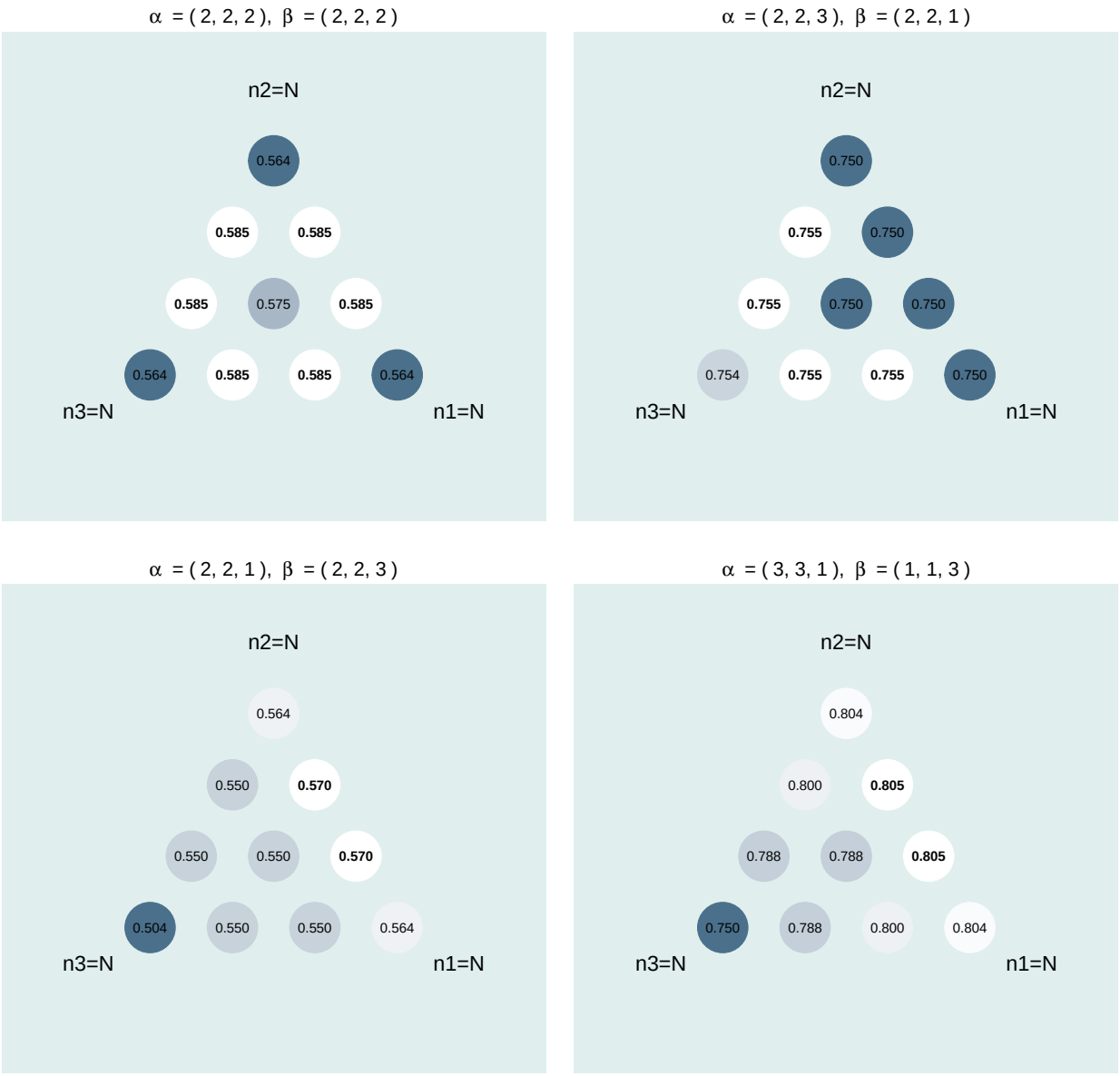
10 waves

Statistic	Ashraf	Bryan	Cohen
Regret, non-adaptive stratified	0.011	0.005	0.016
Regret, Thompson	0.007	0.003	0.013
Regret, modified Thompson	0.007	0.002	0.012
Share optimal, non-adaptive stratified	0.197	0.736	0.077
Share optimal, Thompson	0.274	0.809	0.110
Share optimal, modified Thompson	0.286	0.858	0.119
Units per wave	150	280	324
Number of treatments	6	4	7
Number of strata	5	2	4

Notes: This table shows average regret and the share of replications for which the optimal targeted treatment policy was chosen across 10,000 simulation replications. Parameters are calibrated based on the data of published experimental studies, as shown in Figure A2.

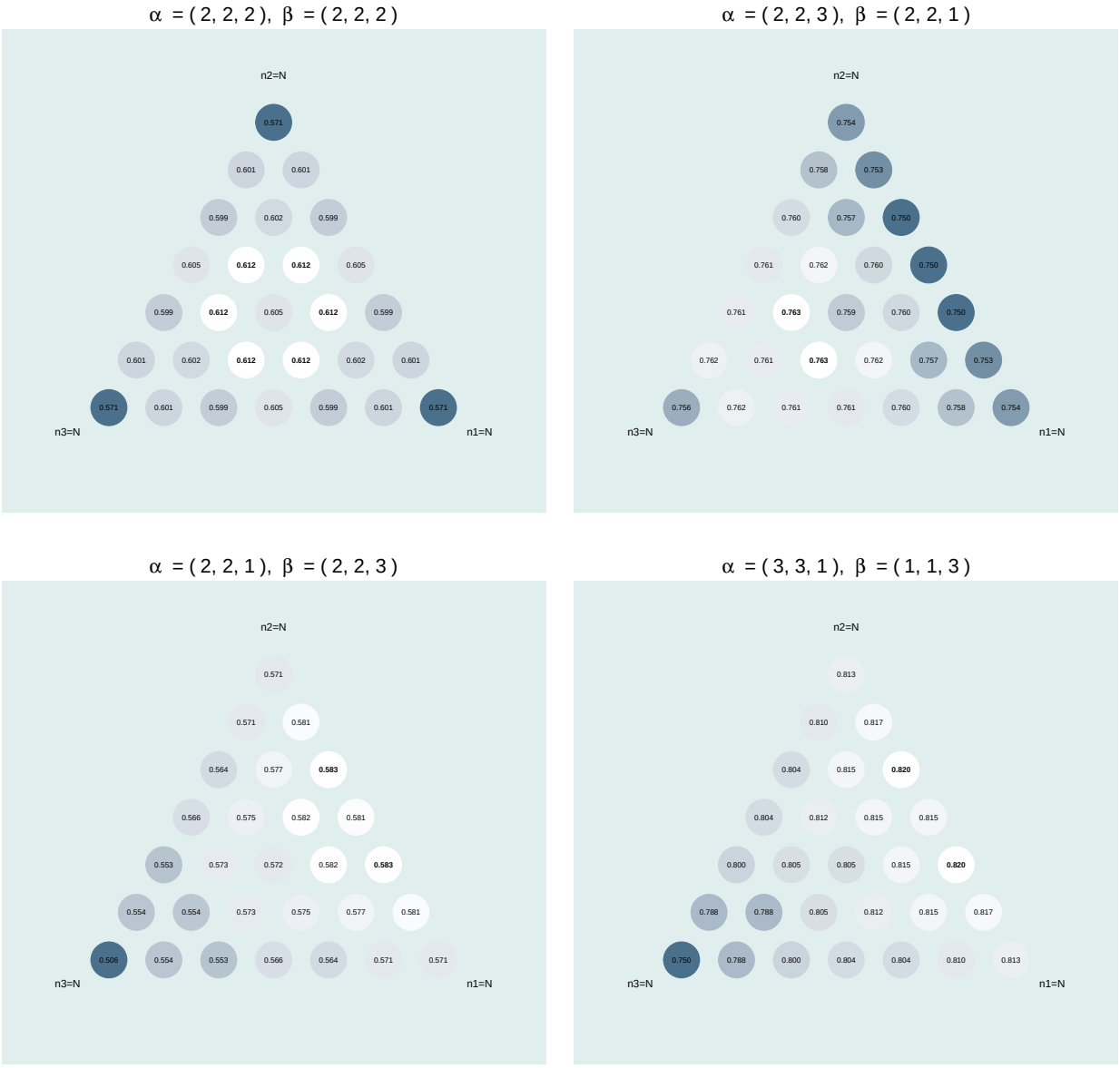
A.3 Additional optimal design example plots

Figure A3: Expected welfare as a function of treatment assignment, sample size 3



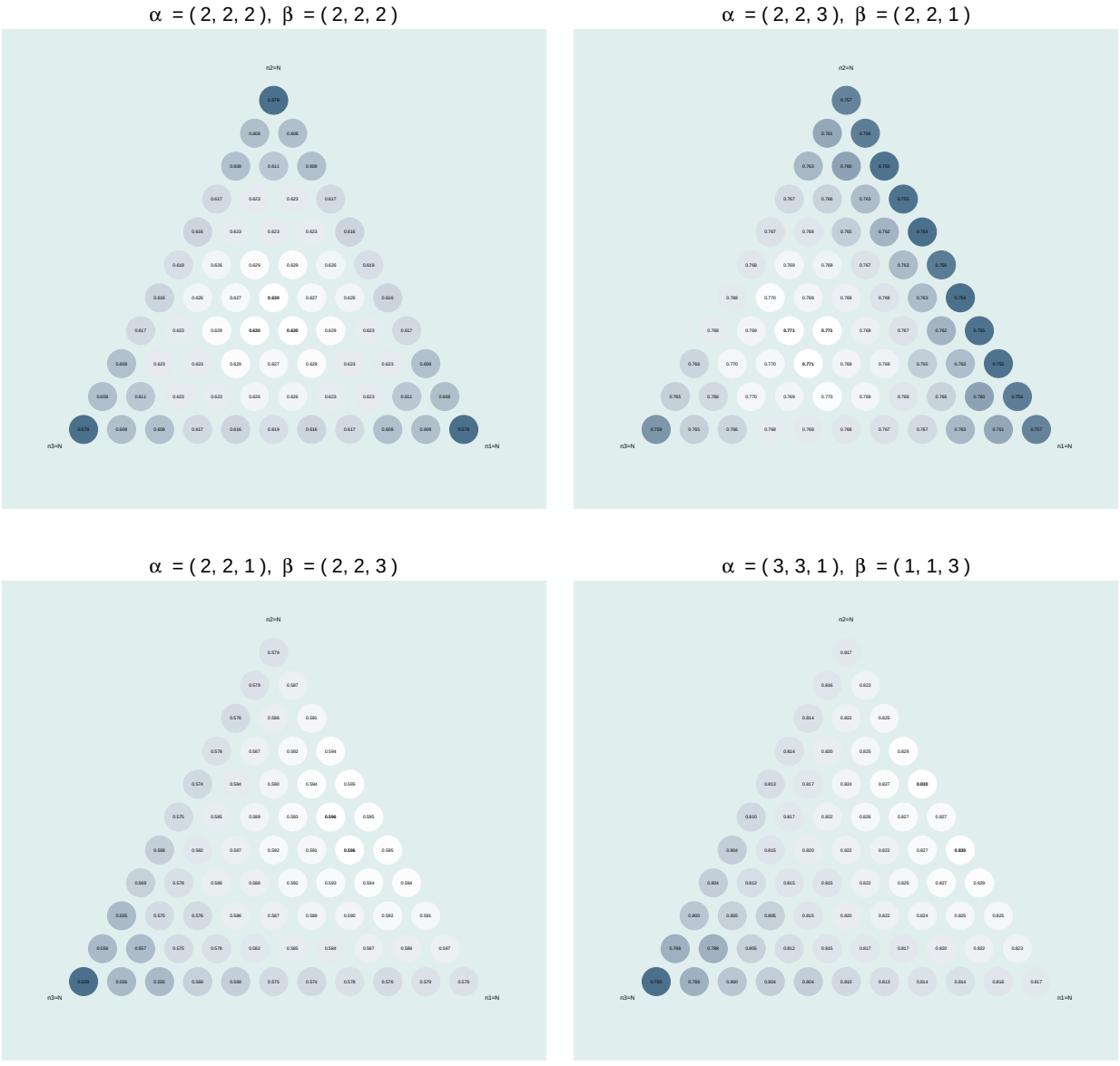
Notes: This figure shows the expected welfare U_2 as a function of treatment assignment $\boldsymbol{n}_2$ in wave 2 (size 3), taking as given the *Beta*-prior parameters $\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1$ determined by the outcomes of wave 1 (size 6). Note that the color scaling differs across figures for better readability.

Figure A4: Expected welfare as a function of treatment assignment, sample size 6



Notes: This figure shows the expected welfare U_2 as a function of treatment assignment $\boldsymbol{n}_2$ in wave 2 (size 6), taking as given the *Beta*-prior parameters $\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1$ determined by the outcomes of wave 1 (size 6). Note that the color scaling differs across figures for better readability.

Figure A5: Expected welfare as a function of treatment assignment, sample size 10



Notes: This figure shows the expected welfare U_2 as a function of treatment assignment $\boldsymbol{n}_2$ in wave 2 (size 10), taking as given the *Beta*-prior parameters $\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1$ determined by the outcomes of wave 1 (size 6). Note that the color scaling differs across figures for better readability.

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 experimentsa. In *Handbook of Economic Field Experiments*, volume 1, pages 73–140. Elsevier.

Banerjee, A., Duflo, E., and Kremer, M. (2016). The influence of randomized controlled trials on development economics research and on development policy. *Mimeo MIT*.

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.

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.

Conde-Agudelo, A., Belizán, J. M., and Diaz-Rossello, J. (2012). Cochrane review: Kangaroo mother care to reduce morbidity and mortality in low birthweight infants. *Evidence-Based Child Health: A Cochrane Review Journal*, 7(2):760–876.

Duflo, E. (2017). Richard T. Ely lecture: The economist as plumber. *American Economic Review*, 107(5):1–26.

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

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Taylor & Francis.

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.

Judd, K. L. (1998). *Numerical methods in economics*. MIT press.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):pp. 47–55.

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.

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.