Adaptive treatment assignment in experiments for policy choice

Maximilian Kasy[*] Anja Sautmann[†‡]

December 2, 2019

Abstract

Standard experimental designs are geared toward point estimation and hypothesis testing, and bandit algorithms are geared toward in-sample outcomes. Here, we consider treatment assignment in an experiment with several waves for choosing the best among a set of possible policies (treatments) at the end of the experiment. We propose a computationally tractable assignment algorithm that we call "exploration sampling," where assignment probabilities in each wave are an increasing concave function of the posterior probabilities that each treatment is optimal. We prove an asymptotic optimality result for this algorithm and demonstrate improvements in welfare in calibrated simulations over both non-adaptive designs and bandit algorithms. An application to selecting between six different recruitment strategies for an agricultural extension service in Odisha, India, demonstrates practical feasibility.

KEYWORDS: EXPERIMENTAL DESIGN, FIELD EXPERIMENTS, OPTIMAL POLICY, MULTI-ARMED BANDITS

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 conducting an experiment to evaluate its programs is often different: instead of estimating effect sizes, they are interested in identifying and implementing the best out of several possible

[*] Department of Economics, Harvard University, maximiliankasy@fas.harvard.edu.

[†] J-PAL, Massachusetts Institute of Technology, sautmann@mit.edu.

[‡] We thank Isaiah Andrews, Guido Imbens, Robbie Minton, Ashesh Rambachan, and numerous seminar participants for feedback on this manuscript. We thank Precision Agriculture for Development, Shawn Cole, and Grady Killeen for helping us implement our procedure in the field.

policies or policy variants. In other words, they would like to answer "Which program will have the largest effect?"

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 outcomes. Then treatments are assigned for the next wave. 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.

This setting defines a finite-horizon dynamic stochastic optimization problem. It can be solved analytically using backward induction, but finding exact solutions is computationally challenging. We therefore propose a new assignment algorithm, "exploration sampling," a modification of Thompson sampling. In Thompson sampling, the probability that a treatment d is assigned to a given experimental unit arriving at t is equal to the posterior probability p_t^d (given outcomes up to $t-1$) that this treatment is in fact optimal. For exploration sampling, we replace p_t^d with the assignment share $q_t^d = S_t \cdot p_t^d \cdot (1 - p_t^d)$, where S_t is a normalizing constant.

We show with theoretical results and simulations that this modification improves expected welfare. It avoids assigning more than 50% of the sample to the highest-performing treatment, and in large samples equalizes power for rejecting each of the sub-optimal treatments. This is optimal for the convergence rate of welfare (while standard Thompson sampling is not). As discussed in Section 7, the algorithm and its characterization extend to settings with heterogeneous treatment costs, non-binary outcomes, non-linear objectives, and targeted treatment assignment based on covariates.

The idea of adaptive treatment assignment is almost as old as that of randomized experiments (Thompson, 1933). Adaptive experimental designs have been used for example in clinical trials (Berry, 2006; FDA, 2018) and in the targeting of online advertisements (Russo et al., 2018), but they are not yet common in economics. Our setting is closely related to multi-armed bandit problems (Scott, 2010), but with the key difference that there is no "exploitation" motive, and thus no exploitation-exploration tradeoff. Under some conditions, the optimal solution to the bandit problem can be expressed in terms of choosing the arm corresponding to the highest "Gittins index," cf. Weber et al. (1992). In practice, most applications use heuristic algorithms such as the Upper Confidence Bound algorithm (UCB) and Thompson sampling (Russo et al., 2018). A recent literature characterizes the expected regret of these algorithms, see for example Bubeck and Cesa-Bianchi (2012). Approximations to our optimization problem are also considered in the literature on Bayesian optimization, Frazier (2018). Russo (2016) considers the closely related problem of maximizing the probability of picking the best treatment (rather than maximizing expected welfare). Our theoretical analysis in Section 4 below draws on insights from this paper, on the characterization of oracle-optimal allocations in Glynn and Juneja (2004), 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, that is, 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.

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

Each person or unit i in period t can receive one of k different treatments $D_{it} \in \{1, \dots, k\}$, resulting in a binary outcome $Y_{it} \in \{0, 1\}$ 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 $(Y_{it}^1, \dots, Y_{it}^k)$ for unit i in period t is an i.i.d. draw from the population of interest. Each treatment d has a stationary 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 $\boldsymbol{n}_t = (n_t^1, \dots, n_t^k)$ with $\sum_d n_t^d = N_t$. Denote $s_t^d = \sum_i \mathbf{1}(D_{it} = d, Y_{it} = 1)$ the number of successes ($Y_{it} = 1$) in treatment group d in wave t. The outcome of wave t can be summarized by the vector $\boldsymbol{s}_t = (s_t^1, \dots, s_t^k)$, where $s_t^d \leq n_t^d$. These outcomes are observed at the end of wave t.

Denote the cumulative versions of these terms from 1 to t by $m_t^d = \sum_{t' \leq t} n_{t'}^d$, $r_t^d = \sum_{t' \leq t} s_{t'}^d$, and $\boldsymbol{m}_t = (m_t^1, \dots, m_t^k)$, $\boldsymbol{r}_t = (r_t^1, \dots, r_t^k)$. With i.i.d. potential outcomes, total observations and successes $(\boldsymbol{m}_t, \boldsymbol{r}_t)$ are sufficient statistics for the likelihood of the data given $\boldsymbol{\theta}$ and summarize all relevant information for the experimenter at the beginning of period $t + 1$.

Bayesian updating. 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 independent across d. A special case, and the default for the applications later in this paper, is the uniform prior, corresponding to $\alpha_0^d = \beta_0^d = 1$ for all d.

After the outcomes for periods $1, \dots, t$ 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 $\beta_t^d = \beta_{t-1}^d + n_t^d - s_t^d = \beta_0^d + m_t^d - r_t^d$.

Policy choice and regret. After wave T, the experimenter implements a policy $d_T^* \in 1, \dots, k$, with the objective of maximizing the expected average of the outcome Y for the whole (remaining) population. At the conclusion of the experiment, per-capita expected social welfare of policy d is given by

$$SW_T(d) = E_T[\theta^d | \boldsymbol{m}_T, \boldsymbol{r}_T] = \tfrac{\alpha_0^d + r_T^d}{\alpha_0^d + \beta_0^d + m_T^d}$$

and the optimal policy choice is $d_T^* = \text{argmax}_{\,d}\, SW_T(d)$. Denote the true optimal treatment $d^{(1)} = \arg\max_{d'} \theta^{d'}$, and call $\Delta^d = \theta^{d^{(1)}} - \theta^d$ be the *policy regret* when choosing treatment d, relative to the optimal treatment. Then maximizing expected social welfare is equivalent to minimizing the expected regret at T, $E_T[\Delta^d|\boldsymbol{m}_T, \boldsymbol{r}_T] = \theta^{d^{(1)}} - SW_T(d)$, and d_T^* is the minimizing choice. Note that the objective considered in the bandit literature is *in-sample regret* $\frac{1}{M}\sum_{i,t} \Delta^{D_{it}}$ rather than policy regret $\Delta^{d_T^*}$. Disregarding the welfare of participants in the experiment is justified if their number is small relative to the population of interest.

Treatment assignment The experimenter chooses treatment assignment $\boldsymbol{n}_t$ at the beginning of wave t. Treatment assignment can depend on the outcomes of waves 1 to $t-1$, and on a randomization device. We will evaluate treatment assignment algorithms based on expected social welfare, or equivalently expected policy regret (conditional on the true $\boldsymbol{\theta}$), for the policy d_T^*:

$$\text{R(T)} = E\left[\Delta^{d_T^*}|\boldsymbol{\theta}\right] = \sum_d \Delta^d \cdot P\left(d_T^* = d|\boldsymbol{\theta}\right), \tag{2.1}$$

where T denotes the number of waves of the experiment, and the expectation is over all possible success realizations and treatment assignment choices.

Optimal treatment assignment. The choice of treatment assignment $\boldsymbol{n}_t$ for each $t = 1, \ldots, T$ is a dynamic stochastic optimization problem that can in principle be solved using backward induction, with full enumeration of all possible states and actions. 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 follow a Beta-Binomial distribution, $P(s_t^d = s|\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) = E[P(s_t^d = s|\theta^d, n_t^d)|\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t] = \binom{n_t^d}{s}\frac{B(\alpha_{t-1}^d + s, \beta_{t-1}^d + n_t^d - s)}{B(\alpha_{t-1}^d, \beta_{t-1}^d)}$. In the online supplement we discuss the derivation of value functions and the corresponding optimal assignment functions, which map the state $(\boldsymbol{m}_t, \boldsymbol{r}_t)$ into the assignment $\boldsymbol{n}_t$. Under optimal treatment assignment, R(T) is minimized. We show in numerical examples with two waves that the optimal treatment assignment in wave 2 assigns more units to those treatments that performed better in wave 1.

Computational complexity We show in the online supplement that the time complexity for dynamic programming with full memorization in this setting is of order $\sum_{t=1}^{T-1} O\left((M_t N_{t+1})^{2k-1}\right) + O(M_T^{2k-1}k)$, and the memory complexity is of order $\sum_{t=1}^{T} O\left(M_t^{2k-1}\right)$. An alternative to full optimization is the use of simpler heuristic algorithms, widely used for bandit problems. This reduces computational complexity but may increase expected policy regret R(T). Below, we first briefly discuss one of the most popular (and oldest) bandit algorithm, so-called Thompson sampling, originally proposed by Thompson (1933). We then propose a new, closely related adaptive algorithm we call exploration sampling, and show that for exploration sampling R(T) converges to zero at a constrained optimal rate.

3 Thompson sampling

We next define Thompson sampling and review its large-sample behavior, in order to compare it with our proposed algorithm below. Consider the special case of our setting where units arrive sequentially. 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,

$$p_t^d = P(D_t = d | \boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}) = P\left(d = \underset{d'}{\operatorname{argmax}}\ \theta^{d'} | \boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}\right). \tag{3.1}$$

This prescription can be implemented by sampling one draw $\hat{\theta}_t$ from the posterior given $\boldsymbol{m}_{t-1}$ and $\boldsymbol{r}_{t-1}$, and setting $D_t = \operatorname{argmax}_d \hat{\theta}_t^d$. In the context of the Beta-Binomial model above, $\hat{\theta}_t$ is sampled from its Beta posterior. Thompson sampling can be applied much more generally; an excellent overview is in Russo et al. (2018). When treatment assignment takes place in waves, it is natural to adapt Thompson sampling by assigning a non-random share p_t^d (up to rounding) of observations in wave t to treatment d, in order to reduce randomness.[1] We will refer to this method of assignment as *expected Thompson sampling.*

The large-sample behavior of Thompson sampling In many bandit problems, the goal is to minimize average in-sample regret $E(\frac{1}{T}\sum_{t=1}^{T}\Delta^{D_t})$. 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

$$\lim_{T\to\infty} E\left[\frac{\sum_{t=1}^{T}\Delta^{D_t}}{\log T}\right] \leq \left(\sum_{d\neq d^{(1)}} \frac{1}{(\Delta^d)^2}\right)^2. \tag{3.2}$$

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 result 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 benefits in-sample welfare, but is not optimal for ex-post policy choice.

Bubeck et al. (2011) formalize this intuition. 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 policy regret. This contrasts with algorithms which assign a fixed, non-zero share of observations to each treatment, such as conventional (non-adaptive) designs. In general, algorithms that converge to non-zero shares achieve an exponential rate of convergence.

[1] The remainder after rounding is still assigned randomly, so that expected shares remain equal to p_t^d.

4 Exploration Sampling

Based on Thompson sampling, we propose a modified treatment assignment algorithm we call *exploration sampling.* It replaces the Thompson assignment shares $(p_t^1, \ldots, p_t^k)$ with the following transformed shares:

$$q_t^d = S_t \cdot p_t^d \cdot (1 - p_t^d), \qquad S_t = \tfrac{1}{\sum_d p_t^d \cdot (1-p_t^d)}. \tag{4.1}$$

This modification shifts weight away from the best performing option to its close competitors. Since there is at most one d for which $p_t^d > 1/2$, we have that across d (given S_t) the mapping from p_t^d to q_t^d is monotonically increasing and concave.

Heuristic motivation. Exploration sampling would arise if we used Thompson sampling but never assigned the same treatment twice in a row, thus improving power for comparisons of relevant alternatives. Suppose that within a given wave, we sequentially draw treatment assignments based on the Thompson probabilities $\boldsymbol{p}$, but if necessary, the draw is repeated until the current unit is assigned a different treatment from the previous unit. This algorithm defines a Markov chain for the sequence of assigned treatments where 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 $\boldsymbol{q}$ that satisfies $q^d = \sum_{d' \neq d} q^{d'} \frac{p^d}{1-p^{d'}}$ for all d. By the mean ergodic theorem, the assignment shares converge to this stationary distribution. Solving for $\boldsymbol{q}$ yields Equation (4.1). Thus, for large wave sizes, this algorithm assigns the same share of observations to each treatment as exploration sampling.

4.1 The large-sample behavior of exploration sampling

Our key result shows that exploration sampling achieves the best possible exponential rate of convergence, subject to the constraint that half the observations end up assigned to the best treatment. It achieves in particular a better exponential rate than non-adaptive assignment, and converges much faster than Thompson sampling, which only converges polynomially. In a second characterization, we show that for a large first wave, exploration sampling splits the second wave equally between the two best treatments.

Many waves Theorem 1 characterizes the behavior of exploration sampling in settings with many waves and fixed wave sizes. Let $\bar{q}_t^d = \frac{1}{t}\sum_{s \leq t} q_t^d$, and write "$\to^p$" for convergence in probability.

Theorem 1 *Consider the exploration sampling algorithm of Section 2 with fixed wave size $N_t = N \geq 1$. Assume that $\theta^{d^{(1)}} < 1$ and that the optimal policy $d^{(1)}$ is unique. As $T \to \infty$, the following holds:*

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

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

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

The proof of Theorem 1 can be found in Appendix A, where we first state six preliminary Lemmas before proceeding to the main proof. Lemmas 1 through Lemma 3, drawing on Glynn and Juneja (2004), characterize the oracle optimal allocation $\bar{\boldsymbol{q}}$ of observations across the treatments d. This allocation maximizes the rate of convergence of policy regret to 0, as T goes to ∞. This allocation asymptotically equalizes the power of tests comparing the optimal treatment to each suboptimal treatment. Lemmas 4 through 6, drawing on Russo (2016), leverage results on posterior consistency and the rate of convergence of posterior probabilities to give sufficient conditions for $\bar{\boldsymbol{q}}_T$ to converge to $\bar{\boldsymbol{q}}$.

The main proof of Theorem 1 then proceeds in several steps. First, we show that each treatment is assigned infinitely often. This implies that p^d_T goes to 1 for the optimal treatment and to 0 for all other treatments. Claim 1 then follows from the definition of exploration sampling. Second, we show claims 2 and 3 by contradiction. Suppose p^d_t goes to 0 at a faster rate for any one of the sub-optimal treatments d. Then exploration 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.

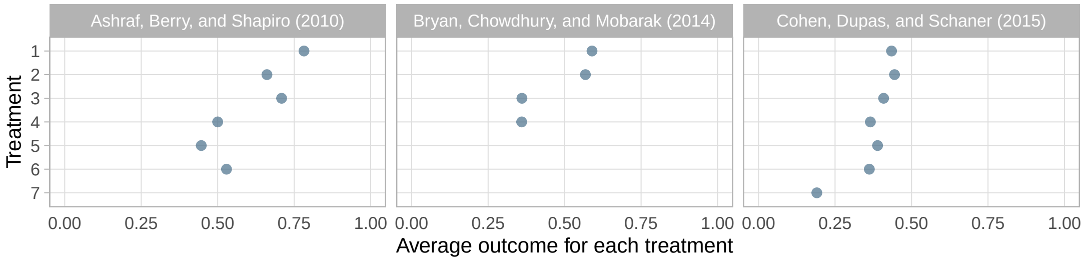
Large first wave Next we consider the case of large wave size N_1. With finite T, the potential for adaptivity is limited, so no optimality guarantees exist. We can nonetheless characterize the behavior of adaptive algorithms. For large N_1, with high probability Thompson sampling assigns all observations in the second wave to the best performing treatment, while exploration sampling splits the second wave equally between the best two treatments.

Proposition 1 *Consider the setting of Section 2, and assume that both the optimal policy $d^{(1)}$ and the second best policy $d^{(2)} = \operatorname{argmax}_{d \neq d^{(1)}} \theta^d$ are unique. Then the following holds.*

1. *Suppose that treatment is assigned using Thompson sampling. Then the second period assignment shares satisfy $p^{d^{(1)}}_2 \to^p 1$ as $N_1 \to \infty$, and $p^d_2 \to^p 0$ for $d \neq d^{(1)}$.*

2. *Suppose that treatment is assigned using exploration sampling. Then, as $N_1 \to \infty$, the second period assignment shares satisfy $q^d_2 \to^p \frac{1}{2}$ for $d \in \{d^{(1)}, d^{(2)}\}$, and $q^d_2 \to^p 0$ for $d \notin \{d^{(1)}, d^{(2)}\}$.*

Figure 5.1: Average treatment outcomes in experimental data.



Notes: Treatment arms labeled 1 up to 7: Ksh 300 - 800 price for water disinfectant, treatment 1 (Ksh 300) is optimal (Ashraf et al.); migration incentives - cash, credit, information, and control, treatment 1 (cash) is optimal (Bryan et al.); price of Ksh 40, 60, and 100 for malaria tablets, each with and without free malaria test, and control of Ksh 500, treatment 2 (Ksh 40, no test) is optimal (Cohen et al.).

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 the calibration is to "tie our hands" in choosing research designs and data characteristics for our simulations.

Experiments. We use data from the experiments in Ashraf et al. (2010), Bryan et al. (2014), and Cohen et al. (2015). They each have multiple treatments and a binary outcome (more information is in the online supplement). Our simulations use the same sample sizes as the original experiments, but for simplicity ignore any clustering. We assume that the policymaker's goal is to maximize the average measured outcome (which was not necessarily the goal of the original experiments).

Figure 5.1 shows the average outcomes across treatment arms for each of the three experiments. We set the "true" parameter vectors $\boldsymbol{\theta}$ equal to these average outcomes for the purpose of our simulations. For Ashraf et al. (2010), average outcomes are roughly evenly spaced. This makes it comparatively easy to statistically detect which treatments are performing better, so we would expect benefits of adaptation even for moderate sample sizes. For Bryan et al. (2014), two treatments are clearly better, but they are also very close. It is easy to distinguish these from the other two treatments, but it takes a large amount of information to figure out which is the best, and the returns to doing so are not very large. For Cohen et al. (2015), the top six treatments are again roughly evenly spaced, but the best treatments are close together and thus hard to distinguish.

Simulation results. We compare three different algorithms. The *non-adaptive* algorithm assigns an equal share of units to each of the treatment arms and serves as a benchmark. *Expected Thompson* assigns a non-random share of units in each wave based on the Thompson probabilities. *Exploration sampling*, our preferred approach, is as described in Section 4.

We evaluate the performance of these algorithms using the distribution of policy regret across 100,000 simulation draws. Recall that policy regret is given by $\Delta^{d^*_T} = \max_d \theta^d - \theta^{d^*_T}$. For each experiment, the vector $\boldsymbol{\theta}$, and in particular $\max_d \theta^d$, is fixed across simulation draws.

Table 1 shows performance metrics for each of the three algorithms considered, and for varying numbers of waves, holding total sample size constant. We report average policy regret, as well as the share of simulation draws for which the optimal treatment $d^{(1)}$ would be chosen after T. Finally, we report average in-sample regret, $\frac{1}{M}\sum_{i,t}\Delta^{D_{it}}$.

The results show that exploration sampling consistently outperforms expected Thompson sampling in terms of average policy regret, and both outperform non-adaptive assignment. Adaptive designs with more waves perform better than designs with fewer waves (for the same total sample size). The gains from adaptive designs are largest in the application to Ashraf et al. (2010), followed by Cohen et al. (2015).

The probability of choosing the best treatment is strictly larger than under non-adaptive assignment as well; figures in the online supplement furthermore show that the distribution of policy regret under exploration sampling first-order stochastically dominates the 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.

In-sample regret is the objective of bandit algorithms, for which Thompson sampling is rate-optimal, but exploration sampling is not. Here, Thompson performs best among the algorithms compared. However, in the same way that Thompson sampling outperforms non-adaptive assignment for average policy regret, the fact that exploration sampling shifts sampling toward the better performing options means that it dominates non-adaptive assignment in terms of in-sample regret.

6 Implementation in the field

Precision Agriculture for Development (PAD) is an NGO that works with the government of Odisha, India, to provide a phone-based personalized agricultural extension service to rice farmers. We designed an experiment using exploration sampling as in Section 4 to help PAD choose among a variety of different call methods to enroll farmers.[2] The outcome is a binary variable describing call completion: It equals one if the call recipient answered five questions asked during the call (which enables processing). PAD tested six treatments that combined automated voice calls in the morning or evening with possible text message alerts sent up to 24 hours ahead.

The sample was selected from a list of phone numbers provided by the government. PAD set aside a batch of 10,000 valid numbers that are not on the Indian "do-not-disturb" list, and

[2] The pre-analysis plan for this experiment was registered at `www.socialscienceregistry.org/trials/4263`. The R-code used for implementation can be found at `github.com/maxkasy/Precision-Agriculture-for-Development`.

Table 1: Simulation results

Ashraf, Berry, and Shapiro (2010)

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

Bryan, Chowdhury, and Mobarak (2014)

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

Cohen, Dupas, and Schaner (2015)

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

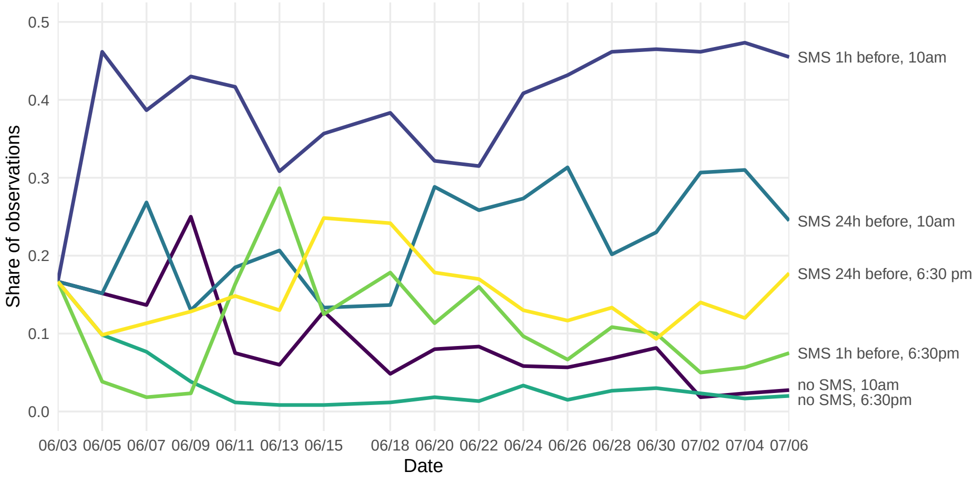
Notes: Average policy regret, share optimal, and average in-sample regret across 100,000 simulation draws. The vector $\boldsymbol{\theta}$ equals average outcomes in the original experiment in all draws. The total sample size is as in the original experiment.

Table 2: Outcomes of the adaptive experiment for PAD (Odisha).

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

Notes: For each treatment arm: total observations, total successes, share of successes, posterior mean and standard deviation of the success rate, and probability that the arm is optimal. 10,000 units and 17 waves.

Figure 6.1: Assignment frequencies over time.



randomly selected waves of 600 phone numbers for testing. Starting on June 3, 2019, a new experimental wave was started every other day and completed the next day (with a one-day delay on June 17).[3] The success rate of each treatment arm was estimated starting with a uniform prior over $\boldsymbol{\theta}$ in order to determine the assignment frequencies for each consecutive wave.

Findings Table 2 shows treatment assignments and successes, and the posterior mean and standard deviation of the success rate θ^d, as well as the posterior probability p_T^d that each treatment is optimal. Figure 6.1 plots the assignment shares over time. The figure shows that one treatment was assigned the most units from wave 2 onwards, but some closely competing treatments got a high share of observations, especially in early waves. The number of observations per wave assigned to each of the treatments did stabilize towards later waves, as predicted by our characterization of exploration sampling in Theorem 1.

[3] The schedule got delayed by one day starting June 17 due to internet connectivity issues in Odisha.

Calling farmers at 10am after a text message an hour ahead of time is with over 75 percent probability the treatment with the greatest success rate, estimated to be 19.3 percent. Across treatments, higher success rates are associated with a higher number of observations, and correspondingly smaller posterior standard deviation. This holds by design for exploration sampling. Cumulatively, nearly 40 percent of farmers received the most successful type of call, whereas under four percent received the least successful call (at 6:30pm without a text message alert). Based on the posterior estimated success rates, exploration sampling not only improved learning, but also increased overall success rates within the experiment (18.04%) compared to a standard design with equal assignment to treatment arms (where the estimated success rate based on posterior means would be 17.15%).

7 Extensions

More general outcome distributions and objective functions. Above, we assumed that outcomes are binary and social welfare is linear in the success rate θ^d. This setting immediate generalizes in a number of ways. For normal outcomes with appropriately modified posterior probabilities, the definition of exploration sampling and the characterizations of Section 4 apply verbatim. We can also allow for known costs c^d of treatment d.[4]

More generally, outcomes may follow some other distribution F, and the planner may be interested in other moments of F besides the mean, such as the variance (inequality) or certain percentiles (poverty levels). For example, suppose there is a family of outcome distributions F that can be parameterized with a vector θ^d, and assume that the planner learns about θ^d in order to maximize $E[U(\theta^d)]$, where U is a utility function that may depend on these parameters in complex ways.

The exploration sampling algorithm generalizes in a straightforward way to this setting, as long as we can simulate draws of $\boldsymbol{\theta} = (\theta^d)_{d=1}^k$ from the posterior in each t. We can simulate $\boldsymbol{\theta}$, determine the utility-maximizing treatment for each draw, calculate the share p_t^d of draws for which d is optimal, and obtain the exploration sampling shares q_t^d by formula (4.1). We conjecture that exploration sampling improves learning, and the asymptotic characterizations in section 4.1 hold, in a range of more general settings of this form. In the specific case where outcomes are binary, but regret Δ^d is replaced with $U(\theta^{d^{(1)}}) - U(\theta^d)$, our convergence results apply directly for any utility function U that is increasing in θ^d.

Covariate-specific treatment assignment. Instead of one treatment for the whole population, the planner may be interested in the optimal treatment assignment $d_T^*(\cdot)$ that maps covariates X into treatments D. Such assignment rules are studied in the literature on contextual bandits (Dudik et al., 2011; Dimakopoulou et al., 2017) and on targeted treatment assignment based on observational data (e.g. Kitagawa and Tetenov, 2018).

[4] In the binary outcome setting with treatment costs c^d, Theorem 1 needs the additional conditions $\theta^{d^{(1)}} - c^{d^{(1)}} < 1 - c^d$ and $-c^{d^{(1)}} < \theta^d - c^d \quad \forall d$. This condition ensures that the problem is “hard” enough, so that we can not exclude any treatment from being optimal without observations on that treatment.

A natural adaptation of our algorithm to targeted treatment assignment policies uses hierarchical Bayesian models. Consider the case of binary outcomes Y and discrete covariates X. Let $\theta^{dx} = E[Y^d|X = x]$. We might model $\theta^{dx} \sim Beta(\alpha_0^d, \beta_0^d)$, with $(\alpha_0^d, \beta_0^d) \sim \pi$ distributed independently across d for some prior π. We can sample from the posterior using Markov-chain Monte Carlo, and based on such samples estimate $p_t^{dx} = P\left(d = \text{argmax}\ _{d'}(\theta^{d'} - c^{d'})|X = x, \boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}\right)$. The exploration sampling algorithm for this setting then uses stratum-specific conditional assignment shares $q_t^{dx} = S_t^x \cdot p_t^{dx} \cdot (1 - p_t^{dx})$.

Combining bandit and policy objectives. When the experimental sample is not negligible relative to the policy population, the planner may consider a combination of policy regret and in-sample regret. This is easily accommodated in the fully optimal assignment discussed in the supplement. Adapting the exploration sampling algorithm is less immediate. A possible alternative approach could build on the knowledge gradient method discussed in the literature on Bayesian optimization (Frazier, 2018), which is based on a one-period truncated version of the dynamic optimization problem; its approximation to the value function could be modified to take into account in-sample welfare.

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.

Dimakopoulou, M., Athey, S., and Imbens, G. (2017). Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. (2011). Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*.

FDA (2018). Adaptive designs for clinical trials of drugs and biologics, guidance for industry. *U.S. Department of Health and Human Services, Food and Drug Administration*.

Frazier, P. I. (2018). A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press.

Glynn, P. and Juneja, S. (2004). A large deviations perspective on ordinal optimization. In *Proceedings of the 36th conference on Winter simulation*, pages 577–585. Winter Simulation Conference.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616.

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.

Scott, S. L. (2010). A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658.

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.

A Proofs

Consider an experiment of length T that uses exploration sampling. We are interested in the behavior of R(T) as $T \to \infty$. Throughout this appendix, all probability statements are frequentist (conditional on the true $\boldsymbol{\theta}$). We start with three lemmas that characterize the rate-optimal treatment allocation $\bar{\boldsymbol{q}}$, based on results from Glynn and Juneja (2004). We then restate some lemmas from Russo (2016) that help establish the convergence of exploration sampling to this rate-optimal allocation, before proceeding to the proof of Theorem 1 based on these Lemmas.

A.1 The rate-optimal allocation

Lemma 1 states that the rate of convergence of expected policy regret R(T) to zero is equal to the slowest rate of convergence Γ^d across $d \neq d^{(1)}$ for the probability of d being estimated to be better than $d^{(1)}$. Lemma 2 draws on the theory of large deviations to characterize Γ^d as a function of the treatment allocation share for each d, $\bar{q}^d$. It also states that the posterior probability p_T^d of d being optimal, and therefore the Thompson allocation shares, converge at the same rate Γ^d. Lemma 3 characterizes the allocation of observations across the treatments d which maximizes the rate of R(T).

Lemma 1 *Denote the estimated success rate of d at time T by* $\hat{\theta}_T^d = \frac{1+r_T^d}{2+m_T^d}$. *Assume that the optimal policy* $d^{(1)}$ *is unique. Suppose that* $\lim_{T\to\infty} -\frac{1}{NT} \log P\left(\hat{\theta}_T^d > \hat{\theta}_T^{d^{(1)}}\right) = \Gamma^d$ *for all d. Then*

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

Proof: The expected policy regret can be written as $\mathrm{R(T)} = \sum_d \Delta^d \cdot P\left(\operatorname{argmax}_{d'} \hat{\theta}_T^{d'} = d\right)$. We can bound this below and above:

$$\left(\min_{d\neq d^{(1)}} \Delta^d\right)\cdot\left(\max_{d\neq d^{(1)}} P\left(\hat{\theta}_T^d > \hat{\theta}_T^{d^{(1)}}\right)\right) \leq \mathrm{R(T)}$$
$$\leq \left((k-1)\max_{d\neq d^{(1)}} \Delta^d\right)\cdot\left(\max_{d\neq d^{(1)}} P\left(\hat{\theta}_T^d > \hat{\theta}_T^{d^{(1)}}\right)\right).$$

The claim follows. □

Lemma 2 *Suppose that* $\bar{q}_T^d = m_T^d/(NT)$ *converges to* $\bar{q}^d$ *for all d, with* $\bar{q}^{d^{(1)}} = 1/2$. *Then*

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

where $\Gamma^d = G^d(\bar{q}^d)$ *for a function* G^d *that is finitely valued, continuous, strictly increasing in* $\bar{q}^d$, *and satisfies* $G^d(0) = 0$.

Proof: The first claim follows from the arguments in Glynn and Juneja (2004), Section 2. The second claim follows from Proposition 5 and Lemma 2 in Russo (2016). The function G^d is given by $G^d(\bar{q}^d) = \inf_x\left(\bar{q}^{d^{(1)}}\cdot I^{d^{(1)}}(x) + \bar{q}^d\cdot I^d(x)\right)$, where I^d is the Legendre transform of the log moment generating function of Y^d, and $\bar{q}^{d^{(1)}} = 1/2$. □

Lemma 3 *The rate-optimal allocation* $\bar{\boldsymbol{q}}$, *subject to the constraint* $\bar{q}^{d^{(1)}} = 1/2$, *is given by the unique solution to the system of equations*

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

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

Proof: Based on Lemma 1 and part (2) of Lemma 2, the allocation that maximizes the convergence of R(T) is given by $\arg\max_{\bar{\boldsymbol{q}}} \min_{d \neq d^{(1)}} G^d(\bar{q}^d)$, subject to the constraints $\sum_d \bar{q}^d = 1$ and $\bar{q}^{d^{(1)}} = 1/2$. The claim follows from the properties of G^d in Lemma 1. □

Theorem 1 in Glynn and Juneja (2004) states an unconstrained version of this result, maximizing the probability of picking the optimal policy, rather than minimizing policy regret.

For the remainder of this section, we take Equation(A.1) as the definition of $\bar{\boldsymbol{q}}$. Theorem 1 states that $\bar{\boldsymbol{q}}_T$ converges to $\bar{\boldsymbol{q}}$ for exploration sampling.

A.2 Sufficient conditions for convergence to the optimal allocation

To prove Theorem 1, we draw on several Lemmas from Russo (2016),[5] which we restate here in our notation. Lemma 4 shows that $\bar{q}_t^d$ converges to the optimal share $\bar{q}^d$ as in Equation (A.1) if the assignment share is "self-correcting", in the sense that the current assignment q_t^d is small whenever total assignment $\bar{q}_t^d$ to d is too high. Lemma 5 shows that if $\bar{q}_t^d$ is too high, then p_t^d converges to zero exponentially faster than $p_t^{d'}$ for some other $d' \neq d^{(1)}$, which will allow us to show that q_t^d will be such that Lemma 4 applies. Lemma 6 is a posterior consistency result for adaptive assignments.

Lemma 4 (Lemma 12 of Russo (2016)) *Suppose that* $\bar{q}_t^{d^{(1)}} \to^p 1/2$ *and*

$$\sum_{t=1}^{\infty} q_t^d \cdot \mathbf{1}(\bar{q}_t^d > \bar{q}^d + \delta) < \infty$$

for all $d \neq d^{(1)}$ *and all* $\delta > 0$*, with probability* 1*. Then* $\bar{q}_t^d \to^p \bar{q}^d$*.*

Lemma 5 (Lemma 13 of Russo (2016)) *Suppose that* $\bar{q}_t^{d^{(1)}} \to^p 1/2$ *and consider any* $d \neq d^{(1)}$ *and any* $\delta > 0$*. Then there exists a* $\delta' > 0$ *and a sequence* $\epsilon_t \to 0$ *such that for all* t

$$\bar{q}_t^d > \bar{q}^d + \delta \quad \Rightarrow \quad \frac{p_t^d}{\max_{d' \neq d^{(1)}} p_t^{d'}} \leq \exp(-t(\delta' + \epsilon_t)).$$

Lemma 6 (Proposition 4 of Russo (2016)) *Denote the posterior probability after wave* t *that* θ *is in some set* A *by* $P_t^d(A)$*. Suppose that* $\sum_t q_t^d = \infty$*. Then* $P_t^d([\theta - \epsilon, \theta + \epsilon]) \to^p 1$ *for any* $\epsilon > 0$ *and the true* θ*. Suppose that* $\sum_t q_t^d < \infty$*. Then* $\inf_t P_t^d(A) > 0$ *for any open interval* A*.*

[5] Russo (2016) also proposes algorithms that take the form of assigning units with some fixed probability β to the arm estimated to be best, and the remainder to the second best. Our setting differs from that of Russo (2016), since (1) our objective is to maximize expected welfare rather than the probability of picking the best treatment, (2) Russo's algorithm requires the tuning parameter β rather than using an endogenous assignment rule for the first best treatment.

A.3 Proof of Theorem 1

Recall that under exploration 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.

Step 1: each treatment is assigned infinitely often.
Suppose that treatment d is only assigned finitely often. Then there is some wave t' after which d is not assigned anymore, and the posterior probability p_t^d of d being optimal is bounded away from 0 for all $t > t'$. To see this, note that the posterior (Beta) distribution for θ^d assigns positive mass to any interval $(1-\epsilon, 1]$ for any $\epsilon > 0$. Let $\epsilon = (1-\theta^{d^{(1)}})/2 > 0$. For any other d' and any t, there is positive posterior probability that $\theta^{d'} < 1-\epsilon$: If $\sum_t q_t^{d'} = \infty$ then the posterior probability that $\theta^{d'} < 1-\epsilon$ converges to 1 by Lemma 6. If $\sum_t q_t^{d'} < \infty$, then the posterior probability that $\theta^{d'} < 1-\epsilon$ remains bounded away from 0, again by Lemma 6. It follows that $\operatorname{plim\,inf}_{t\to\infty} p_t^d > 0$.

Under exploration sampling, the denominator in the expression defining q_t^d is bounded above by 1, and thus the assignment share q_t^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. This implies that treatment d will be assigned again with probability 1 after t'; contradiction.

Step 2: the share of observations $\bar{q}_T^{d^{(1)}}$ assigned to the best treatment converges in probability to $1/2$ as $T \to \infty$.
We can derive upper and lower bounds on q_t^d, by considering the maximum and minimum of the expression defining q_t^d with respect to the vector $\boldsymbol{p}_t$, for a given $p_t^d > 0$.

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 p_t^d. These corners are such that $p_t^{d'}$ is equal to 0 for all but two values of d' (one of them d). For any 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.

In the reverse, again by concavity and 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

$$q_t^d \geq \frac{p_t^d \cdot (1-p_t^d)}{p_t^d \cdot (1-p_t^d) + \sum_{d'\neq d} \frac{1-p_t^d}{k-1}\left(1 - \frac{1-p_t^d}{k-1}\right)} = \frac{p_t^d}{p_t^d + \left(1 - \frac{1-p_t^d}{k-1}\right)} \geq \frac{p_t^d}{p_t^d + 1}.$$

Since each treatment is assigned infinitely often, we have that $p_t^{d^{(1)}} \to^p 1$ in probability by Lemma 6 and uniqueness of the optimal policy $d^{(1)}$, and therefore $\frac{p_t^{d^{(1)}}}{p_t^{d^{(1)}}+1} \to^p \frac{1}{2}$. We get that $q_t^{d^{(1)}} \to^p 1/2$ in probability. The claim for $\bar{q}_T^{d^{(1)}}$ follows by the law of large numbers.

Step 3: the share of observations $\bar{q}_T^d$ assigned to treatment d converges in probability to a non-random share $\bar{q}^d$, for all d.
$\bar{q}^d$ is such that $-\frac{1}{NT}\log p_T^d \to^p \Gamma^*$ for all $d \neq d^{(1)}$ and some $\Gamma^* > 0$ that is constant across $d \neq d^{(1)}$.

Consider a sub-optimal treatment d and a subsequence of t where the share of observations allocated to d up to time t exceeds $\bar{q}^d$. Lemma 5 implies that along this subsequence, the posterior probability p_t^d that d is optimal has to go to zero exponentially faster than $p_t^{d'}$, for at least some suboptimal d' (we assume wlog $k > 2$, since the claims of Theorem 1 are immediate for $k = 2$). Note now that by definition of q_t^d

$$q_t^d \leq \frac{p_t^d \cdot (1 - p_t^d)}{p_t^{d'} \cdot (1 - p_t^{d'})} \leq 2 \frac{p_t^d}{p_t^{d'}},$$

where the second inequality holds as long as $p_t^{d'} \leq 1/2$. Since $\frac{p_t^d}{p_t^{d'}}$ converges to 0 at an exponential rate, the same holds for q_t^d. Thus, along any subsequence where the share of observations allocated to d up to time t exceeds $\bar{q}^d$ we have that q_t^d goes to 0 sufficiently fast (e.g., at an exponential rate). By Lemma 4, we have that the share of observations assigned to d has to converge to $\bar{q}^d$. The second part of the claim follows from the definition of G^d in Lemma 2 and the definition of $\bar{\boldsymbol{q}}$ in Equation A.1.

Step 4: expected policy regret converges to 0 at the same rate, that is, $-\frac{1}{NT} \log \mathrm{R(T)} \to^p \Gamma^*$. No other assignment shares $\bar{q}^d$ exist for which $\bar{q}^{d^{(1)}} = 1/2$ and $\mathrm{R(T)}$ goes to 0 at a faster rate than Γ^*.

By Lemma 1 and Lemma 3. □

A.4 Proof of Proposition 1

To show this result, let us consider the posterior probabilities p_2^d that d is optimal after wave 1, for each d. Note first that $p_2^{d^{(1)}} \to^p 1$ as $N_1 \to \infty$ by consistency of posteriors (Schwartz's theorem, e.g. Theorem 6.16 in Ghosal and Van der Vaart 2017). Further, from Proposition 5 in Russo (2016), it follows that $p_2^d / p_2^{d^{(2)}} \to^p 0$ for $d \notin \{d^{(1)}, d^{(2)}\}$: The posterior probability of the set Θ^d of values for $\boldsymbol{\theta}$ which are such that d is the optimal treatment converges in probability to 0 at an exponential rate equal to the KL-divergence of the closest element in Θ^d to the true data generating $\boldsymbol{\theta}$. Given uniqueness of $d^{(2)}$, this KL-divergence is strictly larger for any $d \notin \{d^{(1)}, d^{(2)}\}$ than for $d^{(2)}$, and $p_2^d / p_2^{d^{(2)}} \to^p 0$ follows.

1. From these conditions on $\boldsymbol{p}_2$, the claim for Thompson sampling follows immediately.
2. To show the claim for exploration sampling, recall that $q_2^d = \frac{p_2^d \cdot (1 - p_2^d)}{\sum_{d'} p_2^{d'} \cdot (1 - p_2^{d'})}$. Note that under these conditions on $\boldsymbol{p}_2$, for $d \notin \{d^{(1)}, d^{(2)}\}$

$$
\begin{aligned}
0 \le \frac{p_2^d}{1-p_2^{d^{(1)}}} \le \frac{p_2^d}{p_2^{d^{(2)}}} &\to^p 0 \\
\frac{p_2^{d^{(2)}}}{1-p_2^{d^{(1)}}} = 1 - \sum_{d \notin \{d^{(1)}, d^{(2)}\}} \frac{p_2^d}{1-p_2^{d^{(1)}}} &\to^p 1, \quad \text{and thus} \\
\frac{1}{2} \ge \frac{p_2^{d^{(1)}} \cdot (1-p_2^{d^{(1)}})}{\sum_{d'} p_2^{d'} \cdot (1-p_2^{d'})} \ge \frac{p_2^{d^{(1)}} \cdot (1-p_2^{d^{(1)}})}{\sum_{d' \in \{d^{(1)}, d^{(2)}\}} p_2^{d'} \cdot (1-p_2^{d'})} &\to^p \frac{1}{2}, \\
\frac{1}{2} \ge \frac{p_2^{d^{(2)}} \cdot (1-p_2^{d(2)})}{\sum_{d'} p_2^{d'} \cdot (1-p_2^{d'})} \ge \frac{p_2^{d^{(2)}} \cdot (1-p_2^{d(2)})}{\sum_{d' \in \{d^{(1)}, d^{(2)}\}} p_2^{d'} \cdot (1-p_2^{d'})} &\to^p \frac{1}{2}. \quad \square
\end{aligned}
$$