ADAPTIVE TREATMENT ASSIGNMENT IN EXPERIMENTS FOR POLICY CHOICE

Maximilian Kasy and Anja Sautmann

Standard experimental designs are geared toward point estimation and hypothesis testing, while bandit algorithms are geared toward in-sample outcomes. Here, we instead 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 first 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. These might in turn be used for testing the null hypothesis that the average effect equals 0. The research design is chosen to minimize the estimation error or to maximize the 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 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. 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.\(^1\)

We provide theoretical results and simulations showing 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

\(^1\)An interactive app implementing exploration sampling is available at [https://maxkasy.shinyapps.io/exploration_sampling_dashboard/](https://maxkasy.shinyapps.io/exploration_sampling_dashboard/)
the convergence rate of welfare (while standard Thompson sampling is not).
In Section 8 we discuss how 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 (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 in-sample regret of these algorithms, see for example Bubeck and Cesa-Bianchi (2012); Kock and Thyrsgaard (2017); Kock et al. (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).

Approximations to our optimization problem are also considered in the literature on Bayesian optimization, Frazier (2018).

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 $t = 1, \ldots, T$. Each wave $t$ is a new random sample of $N_t$ experimental units $i = 1, \ldots, N_t$ drawn from the population of interest (so that the waves are repeated cross-sections, and each unit is treated only once). The total sample size is $M = \sum_{t=1}^{T} N_t$. Each person or unit $i$ in period $t$ can receive one of $k$ different treatments $D_{it} \in \{1, \ldots, k\}$, resulting in a binary outcome $Y_{it} \in \{0, 1\}$ determined by the potential outcome equation $Y_{it} = \sum_{d=1}^{k} 1(D_i = d) \cdot Y_{id}$. This assumption implies in particular that there is no interference, i.e., outcomes are not affected by the treatments others receive. Random sampling means that the potential outcome vector $(Y_{it}^1, \ldots, Y_{it}^k)$ for unit $i$ in period $t$ is an i.i.d. draw from the population of interest. Each treatment $d$ has a stationary 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 $n^d_t = \sum_i 1(D_{it} = d)$ the number of units assigned to treatment $d$ in wave $t$. The treatment assignment in wave $t$ is summarized by the vector $\mathbf{n}_t = (n^1_t, \ldots, n^k_t)$ with $\sum_d n^d_t = N_t$. Denote $s^d_t = \sum_i 1(D_{it} = d, Y_{it} = 1)$ 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 $\mathbf{s}_t = (s^1_t, \ldots, s^k_t)$, where $s^d_t \leq n^d_t$. These outcomes are observed at the end of wave $t$. Denote the cumulative versions of these terms from 1 to $t$ by $m^d_t = \sum_{\tau=1}^{t} n^d_{\tau}$, $r^d_t = \sum_{\tau=1}^{t} s^d_{\tau}$, and $\mathbf{m}_t = (m^1_t, \ldots, m^k_t)$, $\mathbf{r}_t = (r^1_t, \ldots, r^k_t)$. With i.i.d. potential outcomes, the total observations and successes in each treatment arm $(\mathbf{m}_t, \mathbf{r}_t)$ are sufficient statistics for the likelihood of the data given $\theta$ (regardless of the wave in
which each unit was observed), and they summarize all relevant information for the experimenter at the beginning of period $t + 1$.

**Bayesian updating.** Under our assumptions, $Y_{dit}^d$ has a Bernoulli distribution with unknown parameter $\theta^d$: $Y_{dit}^d \sim Ber(\theta^d)$. We assume that the policymaker holds prior belief $\theta^d \sim Beta(\alpha^d_0, \beta^d_0)$. The $\theta^d$ are independent across $d$. A special case, and the default for the applications later in this paper, is the uniform prior, corresponding to $\alpha^d_0 = \beta^d_0 = 1$ for all $d$. After the outcomes for periods $1, \ldots, t$ are realized, the posterior distribution is given by $\theta^d|m_t, r_t \sim Beta(\alpha^d_t, \beta^d_t)$, where $\alpha^d_t = \alpha^d_{t-1} + s^d_t = \alpha^d_0 + r^d_t$ and $\beta^d_t = \beta^d_{t-1} + n^d_t - s^d_t = \beta^d_0 + m^d_t - r^d_t$.

**Policy choice and regret.** After wave $T$, the experimenter implements a policy $d^*_T \in 1, \ldots, k$, 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 $SW_T(d) = E[\theta^d|m_T, r_T] = \frac{\alpha^d_0 + r^d_T}{\alpha^d_0 + \beta^d_0 + m^d_T}$ and the optimal policy choice is $d^*_T \in \arg\max_d SW_T(d)$. Denote the true optimal treatment $d^{(1)} \in \arg\max_d \theta^d$, and let $\Delta^d = \theta^{d^{(1)}} - \theta^d$ be the policy regret when choosing treatment $d$, relative to the optimal treatment. 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 $\Delta^{d^*_T}$, as considered here. Disregarding the welfare of participants in the experiment is justified if their number is small relative to the population of interest.

**Expected regret.** The experimenter chooses the treatment assignment $n_t$ at the beginning of wave $t$. This 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 $\theta$), for the
policy \(d^*_T\):

\[
R_{\theta}(T) = E \left[ \Delta d^*_T | \theta \right] = \sum_d \Delta^d \cdot P \left( d^*_T = d | \theta \right),
\]

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 \(n_t\) 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 \((m_{t-1}, r_{t-1})\), and the action in \(t\) is given by \(n_t\). The transition between states is described by \(m_t = m_{t-1} + n_t, r_t = r_{t-1} + s_t\), where the success probabilities follow a Beta-Binomial distribution, \(P(s^d_t = s | m_{t-1}, r_{t-1}, n_t) = \binom{n_t^d}{s_t} \frac{B(\alpha_{t-1}^d + s, \beta_{t-1}^d + n_t^d - s)}{B(\alpha_{t-1}^d, \beta_{t-1}^d)}\). In the online supplement we discuss the derivation of value functions and the corresponding optimal assignment functions, which map the state \((m_t, r_t)\) into the assignment \(n_t\). Under optimal treatment assignment, \(R_{\theta}(T)\) is minimized. We use numerical examples with two waves to show that the optimal treatment assignment in wave 2 assigns more units to those treatments that performed better in wave 1.

**Computational complexity.** We also show in the online supplement that the time complexity for dynamic programming with full memoization (storing all intermediate results) in this setting is of order \(\sum_{t=1}^{T-1} O \left( (M_t N_{t+1})^{2k-1} \right) + O(M_{T_k-1}^{2k-1} k)\), 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 heuristic algorithms, which are widely used for bandit problems. This reduces computational complexity but may increase expected policy regret \(R_{\theta}(T)\). Below, we first briefly discuss one of the most popular (and oldest) bandit algorithms, so-called Thompson sampling, originally proposed by [Thompson (1933)]. We then propose a new, closely related algorithm that we call exploration sampling.
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 first the special case of our setting where units arrive one at a time. In each period $t$, assign treatment $d$ with probability $p_d^t$ equal to the posterior probability, given past outcomes, that it is in fact the optimal treatment,

$$p_d^t = P \left( d = \arg\max_{d'} \theta^d | m_{t-1}, r_{t-1} \right).$$

This prescription can be implemented by sampling one draw $\hat{\theta}_t$ from the posterior given $m_{t-1}$ and $r_{t-1}$, and setting $D_t = \arg\max_d \hat{\theta}_t^d$. 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 number $\lfloor p_d^t N_t \rfloor$ of observations in wave $t$ to treatment $d$, in order to reduce randomness. The remainder of observations are assigned randomly so that expected shares remain equal to $p_d^t$. 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

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

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 re-
viewed in Section 2.3 of Bubeck and Cesa-Bianchi (2012). This result implies that Thompson sampling only assigns a share of units of order 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.

4. EXPLORATION SAMPLING

Based on Thompson sampling, we propose a modified treatment assignment algorithm which we call \textit{exploration sampling}. It replaces the Thompson assignment shares \((p^1_t, \ldots, p^k_t)\) with the following transformed shares:

\[
q^d_t = S_t \cdot p^d_t \cdot (1 - p^d_t), \quad S_t = \frac{1}{\sum_d p^d_t (1 - p^d_t)}.
\]

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

\textbf{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 \( p \), but the draw is repeated, if necessary, 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 \( q \) 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 \( q \) 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, Theorem 1, shows that exploration sampling achieves the best possible exponential rate of convergence, subject to the constraint that in the limit half the observations are 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, Proposition 1, 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 size \( N_t = N \geq 1 \). Let \( \bar{q}_T^d = m_t^d/(N_t) \) denote the share of observations assigned to \( d \) over all waves until \( t \), and write \( \rightarrow_p^* \) for convergence in probability.

**Theorem 1** Consider exploration sampling in the setting of Section 2 with fixed wave size \( N_t = N \geq 1 \). Assume that the optimal policy \( d^{(1)} \) is unique and that \( \theta^{d^{(1)}} < 1 \). 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 \( \rho^{d^{(1)}} \equiv \plim_{T \to \infty} \bar{q}_T^{d^{(1)}} = 1/2 \).
2. The share of observations \( \bar{q}_T^d \) assigned to each treatment \( d \neq d^{(1)} \)
converges in probability to a non-random share \( \rho^d \). \( \rho^d \) is such that
\[
- \frac{1}{NT} \log p^d_T \to^p \Gamma^* \text{ for some } \Gamma^* > 0 \text{ that is constant across } d \neq d^{(1)}.
\]

3. Expected policy regret converges to 0 at the same rate \( \Gamma^* \), that is,
\[
- \frac{1}{NT} \log R_\theta(T) \to \Gamma^*.
\]
No algorithm with limit assignment shares \( \hat{\rho} \neq \rho \) and \( \hat{\rho}^{d^{(1)}} = 1/2 \) exists for which \( R_\theta(T) \) goes to 0 at a faster rate than \( \Gamma^* \).

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 draw on Glynn and Juneja (2004) and characterize the oracle optimal allocation of observations across the treatments \( d \neq d^{(1)} \). This allocation maximizes the rate of convergence of policy regret to 0 as \( T \) goes to infinity by asymptotically equalizing 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 \( \bar{q}_T \) to converge to this allocation.

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 claim 2 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 3) holds because the algorithm balances the rate of convergence of posterior probabilities (or equivalently, of power) across treatments. This is optimal because the rate of convergence of policy regret is dominated by the slowest rate of convergence across treatments.

**Large first wave.** Consider now the case of large wave size \( N_1 \). With small
The potential for adaptivity is limited, so no optimality guarantees exist.

We can nonetheless characterize the behavior of adaptive algorithms. For large $N$, 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 $d^{(2)} = \arg\max_{d \neq d^{(1)}} \theta^d$ are unique. Then the following holds.

1. Suppose that treatment is assigned using Thompson sampling. Then, as $N \to \infty$, the second period assignment shares satisfy $p_{d^{(1)}2} \to^p 1$, and $p_{d \neq d^{(1)}2} \to^p 0$ for $d \neq d^{(1)}$.

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

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)](https://example.com), [Bryan et al. (2014)](https://example.com), and [Cohen et al. (2015)](https://example.com). 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](https://example.com) shows the average outcomes across treatment arms for each
Figure 5.1.— Average treatment outcomes in experimental data.

Notes: Treatment arms labeled 1 up to 7: Kw 300 - 800 price for water disinfectant, treatment 1 (Kw 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.).

of the three experiments. We set the “true” parameter vectors $\theta$ 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 $\Delta^* = \max_{d} \theta_{d} - \theta^*_{d}$. For each experiment, the vector $\theta$, and in
particular $\max_d \theta^d$, is fixed across simulation draws.

Table I 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, because exploration sampling shifts sampling toward the better performing options, 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 government partners to provide a phone-based personalized agricultural ex-
TABLE I
SIMULATION RESULTS
ASHRAF, BERRY, AND SHAPIRO (2010)

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

BRYAN, CHOWDHURY, AND MOBARAK (2014)

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

COHEN, DUPAS, AND SCHANER (2015)

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

Notes: Average policy regret, share optimal, and average in-sample regret across 100,000 simulation draws. The vector \( \mathbf{\theta} \) equals average outcomes in the original experiment in all draws. The total sample size is as in the original experiment.
tension service to 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 rice farmers in one state in India. 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 partner. PAD set aside a batch of 10,000 valid numbers that are not on the Indian “do-not-disturb” list, and 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). The success rate of each treatment arm was estimated starting with a uniform prior over $\theta$ in order to determine the assignment frequencies for each consecutive wave.

**Findings.** Table II shows treatment assignments and success rates, and the posterior mean and standard deviation of the success rate $\theta^d$, as well as the posterior probability $p^d_T$ 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.

---

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

3 The schedule got delayed by one day starting June 17 due to internet connectivity issues in India.
TABLE II
OUTCOMES OF THE ADAPTIVE EXPERIMENT FOR PAD.

<table>
<thead>
<tr>
<th>Treatment Call Time</th>
<th>SMS Alert</th>
<th>$m_T^d$</th>
<th>$r_T^d$</th>
<th>$r_T^d/m_T^d$</th>
<th>mean</th>
<th>SD</th>
<th>$p_T^d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10am</td>
<td>-</td>
<td>903</td>
<td>145</td>
<td>0.161</td>
<td>0.161</td>
<td>0.012</td>
<td>0.009</td>
</tr>
<tr>
<td>10am</td>
<td>1h ahead</td>
<td>3931</td>
<td>757</td>
<td>0.193</td>
<td>0.193</td>
<td>0.006</td>
<td>0.754</td>
</tr>
<tr>
<td>10am</td>
<td>24h ahead</td>
<td>2234</td>
<td>400</td>
<td>0.179</td>
<td>0.179</td>
<td>0.008</td>
<td>0.073</td>
</tr>
<tr>
<td>6:30pm</td>
<td>-</td>
<td>366</td>
<td>53</td>
<td>0.145</td>
<td>0.147</td>
<td>0.018</td>
<td>0.011</td>
</tr>
<tr>
<td>6:30pm</td>
<td>1h ahead</td>
<td>1081</td>
<td>182</td>
<td>0.168</td>
<td>0.169</td>
<td>0.011</td>
<td>0.027</td>
</tr>
<tr>
<td>6:30 pm</td>
<td>24h ahead</td>
<td>1485</td>
<td>267</td>
<td>0.180</td>
<td>0.180</td>
<td>0.010</td>
<td>0.126</td>
</tr>
</tbody>
</table>

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.

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. ALTERNATIVES TO EXPLORATION SAMPLING

Theorem 1 shows that exploration sampling achieves the best possible exponential rate of convergence of expected policy regret, subject to the constraint that in the limit half the observations are assigned to the best treatment, or $\gamma \equiv \text{plim } \bar{q}_{d}^{(1)} = 1/2$. This result raises several questions.

Other algorithms. A first question is whether there are other algorithms that achieve the same form of constrained efficiency. The answer is yes; Proposition 3 in Appendix A.6 shows that there is a larger class of algorithms for which the assignment shares for $d \neq d^{(1)}$ converge to the oracle-optimal limit shares characterized by Lemmas 1 to 3. In these algorithms, the vector of assignment shares in wave $t$, $q_{t}$, is given by a function $\psi(p_{t})$ of the Thompson shares $p_{t}$ which satisfies the properties stated in Appendix A.6. An algorithm satisfying these properties might for instance be constructed by setting $q_{d}^{*} = \gamma$, and $q_{d}^{t} = [(1 - \gamma)/(1 - p_{d}^{t})] \cdot p_{d}^{t}$ for $d \neq d^{*}$.

We will call the limit allocation of such an algorithm $\rho_{\gamma}$. In this notation the assignment shares under exploration sampling converge to $\rho_{1/2}$.

Relative to algorithms such as the one just proposed and those proposed in Russo (2016), exploration sampling has additional advantages that are not reflected in the large-deviation asymptotics of Theorem 1. In these al-

---

Russo (2016) proposes another algorithm satisfying these conditions, which takes the form of assigning units with a fixed probability $\gamma$ to the arm with the highest $p_{d}^{t}$, and the remainder to the arm with the second highest $p_{d}^{t}$. 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, and (2) Russo’s algorithm is less well suited for the batched setting, as it assigns all observations to two treatments. Unlike exploration sampling, Russo’s algorithm and the alternative algorithms characterized here require a tuning parameter $\gamma$. 
algorithms, there tends to be initial instability in assignment shares, due to
the fixed share \( \gamma \) assigned to \( d_T \), and due to initial uncertainty about which
treatment is best. Exploration sampling leads to more stable assignment
shares in initial periods. This advantage would be reflected in worst-case
(local to 0) asymptotics, the analysis of which is left for future research.

Different share for the top treatment. A second question is whether
\( \gamma = 1/2 \) is a good choice. The following proposition, which is related to
Lemma 3 in [Russo (2016)], provides a partial answer. It implies that \( \gamma = 1/2 \)
provides a tight bound on the rate of convergence of expected policy re-
gret, relative to the efficient unconstrained allocation which uses the (un-
known) rate-maximizing choice of \( \gamma \). Consider again the setting of Section 2. Suppose that an algorithm is used under which \( \bar{q}_T \) converges in proba-
bility to the limit allocation \( \rho \), that is rate-optimal for policy regret, sub-
ject to \( \rho^{d(1)} = \gamma \). Denote by \( \Gamma^*_\gamma \) the rate of convergence of \( R_\theta(T) \), that is,
\( \Gamma^*_\gamma = \lim_{T \to \infty} \left( -\frac{1}{NT} \log R_\theta(T) \right) \), for such assignment shares.

**Proposition 2** Under these assumptions, \( \Gamma^*_\gamma \geq \min (\gamma, 1 - \gamma) \cdot \max_{\gamma'} \Gamma^*_{\gamma'} \),
and in particular, \( \Gamma^*_{1/2} \geq \frac{1}{2} \cdot \max_{\gamma'} \Gamma^*_{\gamma'} \).

Data-dependent share for the top treatment. A third question is
whether it would be possible to do better by choosing \( \gamma \) in a data-dependent
manner. Asymptotically, the answer is again yes. In Section 5 of [Glynn
and Juneja (2004)], an algorithm with data-dependent \( \gamma \) is proposed. This
algorithm is based on a plug-in procedure that works as follows. Given the
available data, form an estimate of \( \theta \). Based on this estimate, calculate the
optimal sample shares \( \rho \) for maximizing the asymptotic rate of convergence
\( \Gamma \) of policy regret, as derived from large deviations theory. Use these plug-
in shares for the allocation of treatments. [Glynn and Juneja (2004)] prove
consistency of this approach under some conditions.
8. EXTENSIONS

More general outcome distributions and objective functions. Above, we assumed that outcomes are binary and social welfare is linear in the success rate \( \theta^d \). This setting immediately 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 with minor modifications. We can also allow for known per-unit costs \( c^d \) of treatment \( d \).

More generally, population 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, we may assume that there is a family of outcome distributions \( F \) that can be parameterized with a vector \( \theta^d \), and the planner learns about \( \theta^d \) 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 \( \theta = (\theta^d)_{d=1}^k \) from the posterior in each \( t \). We can simulate \( \theta \), determine the utility-maximizing treatment for each draw, calculate the share \( p^d_t \) of draws for which \( d \) is optimal, and obtain the exploration sampling shares \( q^d_t \) 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 \( \Delta^d \) is replaced with \( U(\theta^{d(1)}) - U(\theta^d) \), our convergence results apply verbatim for any utility function \( U \) that is increasing in \( \theta^d \).

---

5 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 \( -c^{d(1)} < \theta^d - c^d \) \( \forall d \). These conditions ensure that the problem is "hard" enough, so that we cannot exclude any treatment from being optimal without observations on that treatment. They are immediately satisfied if the costs are the same for all \( d \).
**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 [Dudík et al., 2011; Dimakopoulou et al., 2017; Bastani et al., 2017; Dimakopoulou et al., 2018] and on targeted treatment assignment based on observational data (e.g. Kitagawa and Tetenov, 2018).

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 \(\theta^{dx} \sim Beta(\alpha^d_0, \beta^d_0)\) independently across \(x\) given \(d\), with \((\alpha^d_0, \beta^d_0) \sim \pi\) distributed independently across \(d\) for some prior \(\pi\). We can sample from the posterior using Markov-chain Monte Carlo, and based on such samples estimate \(p^{dx}_t = P(d = \text{argmax}_d(\theta^{d'} - c^{d'})|X = x, m_{t-1}, r_{t-1})\). The exploration sampling algorithm for this setting then uses stratum-specific conditional assignment shares \(q^{dx}_t = S^*_t \cdot p^{dx}_t \cdot (1 - p^{dx}_t)\).

**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


Consider an experiment of length $T$ that uses exploration sampling. We are interested in the behavior of $R_{\theta}(T)$ as $T \to \infty$. Throughout this appendix, all probability statements are frequentist (conditional on the true $\theta$). We start with three lemmas that characterize the rate-optimal treatment allocation $\bar{q}$, 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.

APPENDIX A: PROOFS
A.1. The rate-optimal allocation

Lemma 1 states that the rate of convergence of expected policy regret $R_\theta(T)$ to zero is equal to the slowest rate of convergence $\Gamma^d$ across $d \neq d^{(1)}$ for the probability of $d$ being estimated to be better than $d^{(1)}$. For a given limit assignment share $\rho^{d^{(1)}} = \gamma$ for $d^{(1)}$, Lemma 2 draws on the theory of large deviations to characterize $\Gamma^d$ as a function of the treatment allocation share for each $d \neq d^{(1)}$, $\rho^d$. 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 $\Gamma^d$. Lemma 3 characterizes the allocation of observations across the treatments $d \neq d^{(1)}$ which maximizes the rate of $R_\theta(T)$.

**Lemma 1** Denote the estimated success rate of $d$ at time $T$ by $\hat{\theta}_T^d = \frac{\alpha_0^d + r_T^d}{\alpha_0^d + \beta_0^d + m_T^d}$. Assume that the optimal policy $d^{(1)}$ is unique. Suppose that $\lim_{T \to \infty} \frac{1}{NT} \log P(\hat{\theta}_T^d > \hat{\theta}_T^{d^{(1)}}) = \Gamma^d$ for all $d$. Then

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

**Proof:** The expected policy regret can be written as $R_\theta(T) = \sum_d \Delta^d \cdot P(\arg\max_d \hat{\theta}_T^d = d)$. We can bound this from below and above:

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

The claim follows.

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

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

2. $\plim_{T \to \infty} \frac{1}{NT} \log p_T^d = \Gamma^d$,

where $\Gamma^d = G^d(\rho^d)$ for a function $G^d : [0, 1] \to \mathbb{R}$ that is finitely valued,
continuous, strictly increasing in \( \rho^d \), 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 \( G^d(\rho^d) = \inf_x (\gamma \cdot I^{d(1)}(x) + \rho^d \cdot I^d(x)) \), where \( I^d \) is the Legendre transform of the log moment generating function of \( Y^d \).

**Lemma 3** The rate-optimal allocation \( \rho_\gamma \), subject to \( \rho_\gamma^{d(1)} = \gamma \), is given by the unique solution to the system of equations

\[
(A.1) \quad \sum_{d \neq d^{(1)}} \rho_\gamma^d = 1 - \gamma \quad \text{and} \quad G^d(\rho_\gamma^d) = \Gamma^* > 0 \quad \text{for all} \ d \neq d^{(1)}
\]

for some \( \Gamma^* \). No other allocation \( \hat{\rho} \), subject to \( \hat{\rho}^{d(1)} = \gamma \), can achieve a faster rate of convergence of \( R_\theta(T) \) than \( \Gamma^* \).

**Proof:** Based on Lemma 1 and part (1) of Lemma 2, the allocation that maximizes the convergence of \( R_\theta(T) \) is given by \( \arg \max_{\rho} \min_{d \neq d^{(1)}} G^d(\rho^d) \), subject to \( \sum_{d} \rho^d = 1 \) and \( \rho^{d(1)} = \gamma \). The claim follows from the properties of \( G^d \) in Lemma 2.

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 \( \rho \). Theorem 1 states that \( \bar{q}_T \) converges to \( \rho \) 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), which we restate here in our notation. Lemma 4 shows that \( \bar{q}^d_t \) converges to the optimal share \( \rho^d \) as in Equation (A.1) if the assignment share is “self-correcting”, in the sense that the current assignment \( q^d_t \) is small whenever total assignment \( \bar{q}^d_t \) to \( d \) is too high. Lemma 5 shows that if \( \bar{q}^d_t \) 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 3 is a posterior consistency result for adaptive assignments.

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

\[
\sum_{t=1}^{\infty} q_t^d \cdot 1(q_t^d > \rho^d + \delta) < \infty
\]

for all \( d \neq d^{(1)} \) and all \( \delta > 0 \), with probability 1. Then \( q_t^d \to^p \rho^d \).

**Lemma 5 (Lemma 13 of Russo 2016)** Suppose that \( \bar{q}_t^{d^{(1)}} \to^p \gamma \) 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 \( t \)

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

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

**A.3. Proof of Theorem 7**

Recall that exploration sampling assigns a share \( q_t^d = \frac{p_t^d(1-p_t^d)}{\sum_{d'} p_t^{d'}(1-p_t^{d'})} \) of wave \( t \) 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 \( \theta^d \) assigns positive mass to the 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 $\inf_t p_t^d > 0$.

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 $p_t$, 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 $p_t$. The maximum of $q_t^d$ is therefore achieved at a corner of the simplex of possible values for $p_t$ given $p_t^d$. Each corner is such that $p_t^{d'}$ is equal to 0 for all but two values of $d'$ (one of them $d$). For any such $p_t$, we get $q_t^d = 1/2$. This implies $q_t^d \leq 1/2$ for all values of $p_t$ and all $d$.

In the reverse, again by concavity and symmetry of the denominator, the minimum of $q_t^d$ with respect to $\{p_t^{d'}\}_{d' \neq d}$, 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 treatment $d^{(1)}$, and therefore $\frac{p_t^{d(1)}}{p_t^{d(1)} + 1} \to^p 1/2$. We get that $q_t^{d(1)} \to^p 1/2$ in probability. The claim for the average across $t$, $\bar{q}_T^{d(1)}$, follows by the law of large numbers.

**Step 3:** The share of observations $\bar{q}_T^{d}$ assigned to each treatment
$d \neq d^{(1)}$ converges in probability to a non-random share $\rho^d$. $\rho^d$ is such that $-\frac{1}{NT} \log p^d_T \to^p \Gamma^*$ 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 $\rho^d$. Lemma 5 implies that along this subsequence, the posterior probability $p^d_t$ that $d$ is optimal has to go to zero exponentially faster than $\max_{d' \neq d^{(1)}} p^{d'}_t$ (we assume wlog $k > 2$, since the claims of Theorem 1 are immediate for $k = 2$). Note now that by definition of $q^d_t$

$$q^d_t \leq \frac{p^d_t \cdot (1 - p^d_t)}{\max_{d' \neq d^{(1)}} p^{d'}_t \cdot (1 - p^{d'}_t)} \leq 2 \frac{p^d_t}{\max_{d' \neq d^{(1)}} p^{d'}_t},$$

where the second inequality holds as long as $\max_{d' \neq d^{(1)}} p^{d'}_t \leq 1/2$. Since $p^d_t / \max_{d' \neq d^{(1)}} p^{d'}_t$ converges to 0 at an exponential rate, by Lemma 5, the same holds for $q^d_t$. Thus, along any subsequence where the share of observations allocated to $d$ up to time $t$ exceeds $\rho^d$ we have that $q^d_t$ goes to 0 sufficiently fast (i.e., at an exponential rate). By Lemma 4, we have that the share of observations assigned to $d$ has to converge to $\rho^d$. The second part of the claim follows from the definition of $\rho$ in Lemma 3.

Step 4: Expected policy regret converges to 0 at the same rate $\Gamma^*$, that is, $-\frac{1}{NT} \log R_\theta(T) \to \Gamma^*$. No algorithm with limit assignment shares $\hat{\rho} \neq \rho$ and $\hat{\rho}^{d^{(1)}} = 1/2$ exists for which $R_\theta(T)$ goes to 0 at a faster rate than $\Gamma^*$.

By Lemma 1 and Lemma 3. □

A.4. Proof of Proposition 7

To show this result, let us consider the posterior probabilities $p^d_2$ that $d$ is optimal after wave 1, for each $d$. Note first that $p^d_2 \to^p 1$ 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 $p_{d}/p_{d}^{(2)} \rightarrow^{p} 0$ for $d \notin \{d^{(1)}, d^{(2)}\}$: The posterior probability of the set $\Theta^{d}$ of values for $\theta$ 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 $\Theta^{d}$ to the true data generating $\theta$. Given uniqueness of $d^{(2)}$, this KL-divergence is strictly larger for any $d \notin \{d^{(1)}, d^{(2)}\}$ than for $d^{(2)}$, and $p_{d}/p_{d}^{(2)} \rightarrow^{p} 0$ follows.

1. From these conditions on $p_{2}$, the claim for Thompson sampling follows immediately.

2. To show the claim for exploration sampling, recall that $q_{d}^{2} = \frac{p_{d}^{(1)}(1 - p_{d}^{2})}{\sum_{d'} p_{d'}^{2}(1 - p_{d'}^{2})}$. Note that under these conditions on $p_{2}$, for $d \notin \{d^{(1)}, d^{(2)}\}$

$$0 \leq \frac{p_{d}^{2}}{1 - p_{d}^{(1)}} \leq \frac{p_{d}^{2}}{p_{d}^{(2)}} \rightarrow^{p} 0$$

$$\frac{p_{d}^{(2)}}{1 - p_{d}^{(1)}} = 1 - \sum_{d \notin \{d^{(1)}, d^{(2)}\}} \frac{p_{d}^{2}}{1 - p_{d}^{(1)}} \rightarrow^{p} 1, \text{ and thus}$$

$$q_{d}^{(1)} = \frac{1}{1 + \frac{p_{d}^{(1)}}{1 - p_{d}^{(2)}} \frac{p_{d}^{(1)}}{1 - p_{d}^{(2)}} + \sum_{d \notin \{d^{(1)}, d^{(2)}\}} \frac{1 - p_{d}^{(2)}}{1 - p_{d}^{(1)} p_{d}^{2}} \frac{1 - p_{d}^{(2)}}{1 - p_{d}^{(2)}}} \rightarrow^{p} \frac{1}{2},$$

$$q_{d}^{(2)} = \frac{1}{1 + \frac{p_{d}^{(1)}}{1 - p_{d}^{(2)}} \frac{p_{d}^{(1)}}{1 - p_{d}^{(2)}} + \sum_{d \notin \{d^{(1)}, d^{(2)}\}} \frac{p_{d}^{2}}{p_{d}^{(2)}(1 - p_{d}^{(2)})}} \rightarrow^{p} \frac{1}{2}. \quad \square$$
where $\gamma^* = \arg\max_g \Gamma_g^*$, so that the rate of convergence $\Gamma_g^*$ is the highest possible rate of convergence.

Define a hypothetical experiment $B$ that uses the second algorithm with limit shares $\rho_g^*$, but with a number of waves equal to $\lfloor \min(\gamma, 1 - \gamma) \cdot T \rfloor$ for any $T = 1, \ldots, \infty$. Denote regret for experiment $B$, as a function of $T$, by $\widetilde{R}_\theta(T)$. By Lemmas 1 and 2

$$\frac{1}{\lfloor \min(\gamma, 1 - \gamma) \cdot T \rfloor \cdot N} \log \widetilde{R}_\theta(T) \to \Gamma_{\gamma^*}^*.$$  

Since $\lim_{T \to \infty} \frac{\lfloor \min(\gamma, 1 - \gamma) \cdot T \rfloor}{T} = \min(\gamma, 1 - \gamma)$, it follows that

$$\Gamma' \equiv \lim_{T \to \infty} \frac{1}{TN} \log \widetilde{R}_\theta(T) = \min(\gamma, 1 - \gamma) \cdot \Gamma_{\gamma^*}^*.$$  

It remains to show that $\Gamma_{\gamma}^* \geq \Gamma'$. By assumption, the number of observations assigned to $d^{(1)}$ under experiment $A$, divided by the total number of observations $NT$, converges to $\gamma$ in probability. The number of observations assigned to $d^{(1)}$ under experiment $B$, divided by $NT$, converges to $\min(\gamma, 1 - \gamma) \cdot \gamma^* < \gamma$ in probability (since $0 < \gamma^* < 1$; otherwise $\Gamma_{\gamma^*}^* = 0$, which can not be optimal). By a symmetric argument, the number of observations assigned to all $d \neq d^{(1)}$ under experiment $A$, divided by $NT$, converges to $1 - \gamma$, while under $B$ it converges to $\min(\gamma, 1 - \gamma) \cdot (1 - \gamma^*) < 1 - \gamma$.

Therefore, with probability going to 1 as $T$ goes to $\infty$, under experiment $A$ more observations are assigned to treatment $d^{(1)}$, as well as to all other treatments $d \neq d^{(1)}$ combined, relative to experiment $B$. Furthermore, for $A$, the allocation among the treatments $d \neq d^{(1)}$ is such that it provides the largest possible value of $\Gamma_{\gamma}^*$. This, in combination with the fact that the rate of convergence of expected policy regret is monotonically increasing in the number of observations available for any treatment arm, yields the claim. \qed
A.6. Generalizing Theorem 1

Generalizing Theorem 1, Proposition 3 below shows that similar properties are enjoyed by any algorithm satisfying the following conditions:

1. The vector of assignment shares in wave \( t \), \( \mathbf{q}_t \), is given by a function \( \psi(p_t) \) of the vector \( p_t \).
2. \( \lim_{p \to e_d} \psi_d(p) = \gamma \), where \( e_d \) is the \( d \)'th unit vector and \( \gamma < 1 \).
3. \( q_d^t \leq C \cdot \frac{p_d^t}{\max_{d \neq d(1)} p_d^t} \), for some constant \( C \) and all \( d \neq d(1) \).
4. Each treatment is assigned infinitely often.

Conditions 2 and 3 are immediately verifiable properties of the function \( \psi \).

Condition 4 requires case-specific proofs.

Proposition 3 Consider any algorithm satisfying the conditions stated above, in the setting of Section 2, with fixed wave size \( N_t = N \geq 1 \). Assume that the optimal policy \( d(1) \) is unique and that \( \theta_d(1) < 1 \). As \( T \to \infty \), the following holds:

1. The share of observations \( \bar{q}_d^{T(1)} \) assigned to the best treatment converges in probability to \( \rho_d^{T(1)} \equiv \text{plim}_{T \to \infty} \bar{q}_d^{T(1)} = \gamma \).

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

3. Expected policy regret converges to 0 at the same rate \( \Gamma^* \), that is, \( -\frac{1}{NT} \log R_\theta(T) \to \Gamma^* \). No algorithm with limit assignment shares \( \hat{\rho}_\gamma \) with \( \hat{\rho}_\gamma^{d(1)} = \gamma \) exists for which \( R_\theta(T) \) goes to 0 at a faster rate than \( \Gamma^* \).

As an example of an algorithm satisfying these properties, consider treatment assignment using the assignment shares \( q_d^t = \gamma \), and \( q_d^t = \left[ (1 - \gamma)/(1 - p_d^t) \right] \cdot p_d^t \) for \( d \neq d^* \). Properties 1 and 2 are immediate. Property 3 holds with \( C = 1 \). Noting that \( q_d^t \) is bounded away from 0 when the same holds for \( p_d^t \), Property 4 holds by arguments exactly analogous to those in step 1 of the proof of Theorem 1.
The proof of Proposition 3 follows the same steps as our proof of Theorem 1.

**Step 1:** Each treatment is assigned infinitely often.
This holds by assumption.

**Step 2:** The share of observations \( \bar{q}_{T}^{d(1)} \) assigned to the best treatment converges in probability to \( \gamma \) as \( T \to \infty \).
Since each treatment is assigned infinitely often, we have that \( p_{T}^{d(1)} \to \rho 1 \) in probability by Lemma 6 and uniqueness of the optimal treatment \( d^{(1)} \), and therefore \( q_{T}^{d(1)} = \psi^{d(1)}(p_{T}) \to \rho \gamma \) in probability, by our assumption on \( \psi \). The claim for the average across \( t \), \( \bar{q}_{T}^{d(1)} \), 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 \( \rho_{d}^{\gamma} \), for all \( d \).
\( \rho_{d}^{\gamma} \) is such that \( -\frac{1}{NT} \log p_{T}^{d} \to \rho \Gamma^{*} \) 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 \( \rho_{d}^{\gamma} \). 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 \( \max_{d' \neq d^{(1)}} p_{T}^{d'} \) (we assume wlog \( k > 2 \), since the claims of Proposition 3 are immediate for \( k = 2 \)). Note now that by assumption \( q_{T}^{d} \leq C \cdot \frac{p_{T}^{d}}{\max_{d' \neq d^{(1)}} p_{T}^{d'}} \), for some constant \( C \) and all \( d \neq d^{(1)} \). Since \( p_{T}^{d} / \max_{d' \neq d^{(1)}} p_{T}^{d'} \) converges to 0 at an exponential rate, by Lemma 5, the same holds for \( q_{T}^{d} \). Thus, along any subsequence where the share of observations allocated to \( d \) up to time \( t \) exceeds \( \rho_{d}^{\gamma} \) we have that \( q_{T}^{d} \) goes to 0 sufficiently fast (i.e., at an exponential rate). By Lemma 4, we have that the share of observations assigned to \( d \) has to converge to \( \rho_{d}^{\gamma} \). The second part of the claim follows from the definition of \( \rho_{\gamma} \) in Lemma 3.

**Step 4:** Expected policy regret converges to 0 at the same rate, that is, \( -\frac{1}{NT} \log R_{\theta}(T) \to \Gamma^{*} \). No other assignment shares \( \hat{\rho}_{\gamma}^{d} \) exist for which \( \hat{\rho}_{\gamma}^{d(1)} = \gamma \) and \( R_{\theta}(T) \) goes to 0 at a faster rate than \( \Gamma^{*} \).
By Lemma 1 and Lemma 3. □