Supplement for: Adaptive treatment assignment in experiments for policy choice

Maximilian Kasy* Anja Sautmann†

August 24, 2020

1 Optimal treatment assignment

In section 2, we describe the experimental setup and make reference to the fact that the choice of treatment assignment $\boldsymbol{n}_t$ for each $t = 1, \ldots, T$ is a finite dynamic stochastic optimization problem that can be solved using backward induction.

The state at the end of wave $t-1$ is given by $(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1})$, and the action in t is given by $\boldsymbol{n}_t$. The transition between states is described by $\boldsymbol{m}_t = \boldsymbol{m}_{t-1} + \boldsymbol{n}_t$, $\boldsymbol{r}_t = \boldsymbol{r}_{t-1} + \boldsymbol{s}_t$. The success probabilities conditional on the choice of treatment assignment follow a Beta-Binomial distribution and are given by

$$P(s_t^d = s | \boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) = \binom{n_t^d}{s} \frac{B(\alpha_{t-1}^d + s, \beta_{t-1}^d + n_t^d - s)}{B(\alpha_{t-1}^d, \beta_{t-1}^d)}. \tag{1.1}$$

Denote by V_t the value function after completion of wave t, that is, expected welfare assuming that all future treatment assignment decisions will be optimal, and that the optimal policy is implemented after the experiment. V_t is a function of the state $(\boldsymbol{m}_t, \boldsymbol{r}_t)$. After the experiment is concluded, the value function is given by expected welfare for the optimal choice of policy, based on current beliefs:

$$V_T(\boldsymbol{m}_T, \boldsymbol{r}_T) = \max_d \left(E[\theta^d | \boldsymbol{m}_T, \boldsymbol{s}_T] - c^d \right) = \max_d \left(\tfrac{\alpha_0^d + r_T^d}{\alpha_0^d + \beta_0^d + m_T^d} - c^d \right). \tag{1.2}$$

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

$$U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) = \sum_{\boldsymbol{s}: \boldsymbol{s} \leq \boldsymbol{n}_t} P(\boldsymbol{s}_t = \boldsymbol{s} \,|\, \boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) V_t \left(\boldsymbol{m}_{t-1} + \boldsymbol{n}_t, \boldsymbol{r}_{t-1} + \boldsymbol{s} \right), \tag{1.3}$$

*Department of Economics, Oxford University, maximilian.kasy@economics.ox.ac.uk.
†World Bank, asautmann@worldbank.org.

where the probabilities for each vector of successes are given by Equation (1.1). Then the period t value function and the optimal treatment assignment satisfy

$$\begin{aligned} V_{t-1}(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}) &= \max_{\boldsymbol{n}_t:\, \sum_d n_t^d \leq N_t} U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) \\ \boldsymbol{n}_t^*(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}) &= \operatorname*{argmax}_{\boldsymbol{n}_t:\, \sum_d n_t^d \leq N_t} U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t). \end{aligned} \tag{1.4}$$

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

Computational complexity. One can solve for the optimal treatment assignment using backwards induction. This involves enumerating all possible actions and associated outcomes in each time period. With larger sample sizes $M_t = \sum_{t' \leq t} N_{t'}$ and a greater number of waves T and treatments k, however, solving for the optimal assignment quickly becomes infeasible, motivating our simpler exploration sampling approach.

We assume full memoization, where the value function is calculated and stored for every possible state, and then action values are calculated using backwards induction. This approach minimizes the growth of computational time in terms of the number of states and actions; cf. Erickson (2019) chapter 3. At the end of wave t, there are $\binom{M_t+k-1}{k-1} = O(M_t^{k-1})$ possible values $\boldsymbol{m}_t$, and for each $\boldsymbol{m}_t$ there are $\prod_d m_t^d = O(M_t^k)$ possible values of $\boldsymbol{r}_t$, so that the number of possible states at the end of wave t is of order $O(M_t^{2k-1})$.

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

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

$$\sum_{t=1}^{T-1} O\left((M_t N_{t+1})^{2k-1}\right) + O(M_T^{2k-1} k), \tag{1.5}$$

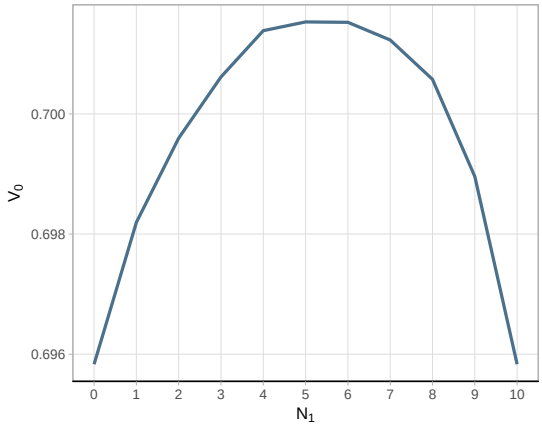
and the memory complexity is of order $\sum_{t=1}^{T} O\left(M_t^{2k-1}\right)$.

2 Optimal design in a simple example

In this section, we discuss optimal experimental designs in a simple example with two waves to show that the optimal assignment in wave 2 assigns more units to those treatments that performed better in wave 1.

Suppose we have ten experimental units that we can enroll in two waves. There are three

Figure 2.1: Dividing the sample across waves.



Notes: The graph shows the expected welfare V_0 (success rate of the final choice) as a function of the sample size N_1 in period 1, assuming a total sample size of 10 and three treatments, and a uniform prior over the three treatment success rates.

treatments. We impose a uniform prior for $\boldsymbol{\theta}$.

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

Figure 2.1 plots expected welfare as a function of the sample size N_1 in wave 1. The boundary cases $N_1 = 0$ and $N_1 = 10$ correspond to an experiment with only one wave. The figure shows that the optimal split assigns either five or six units to the first wave. Splitting the sample in this manner allows us to observe the outcomes from the first-wave assignment (e.g. of two units per treatment if $N_1 = 6$) and then assign treatments for optimal learning to the remaining units in the second wave.[1]

Assigning treatments in wave 2. Based on Figure 2.1, we set $N_1 = 6$. Due to the symmetric prior, it is optimal to assign two units to each of the three treatments in wave 1. Optimal assignment in wave 2, where $N_2 = 4$, depends on the outcomes of the first wave.

We explore several scenarios in Figure 2.2. This figure plots expected welfare for any second-wave treatment assignment in the simplex $n_2^1 + n_2^2 + n_2^3 = 4$, conditional on first-wave outcomes. For each scenario, the number of successes in each treatment in the first wave determines the prior for treatment assignments in the second wave. Our uniform prior for $\boldsymbol{\theta}$ implies a Beta posterior with $\alpha_1^d = 1 + s_1^d$ and $\beta_1^d = 1 + 2 - s_1^d$ for $s_1^d \in \{0, 1, 2\}$ we get. This Beta posterior has a mean of $(1 + s_1^d)/4$.

The four outcome scenarios we consider are $\boldsymbol{s}_1 = (1, 1, 1)$, $\boldsymbol{s}_1 = (1, 1, 2)$, $\boldsymbol{s}_1 = (1, 1, 0)$, and

[1] The welfare differences across alternative designs are relatively small in this setting, owing to the small number of units involved.

Table 1: Thompson shares and assignment shares for different Beta posteriors.

$\alpha_{1,2}$	$\beta_{1,2}$	α_3	β_3	$p_{1,2}$	p_3	$q_{1,2}$	q_3
2	2	2	2	0.3333	0.3333	0.3333	0.3333
2	2	3	1	0.1548	0.6905	0.2752	0.4496
2	2	1	3	0.4548	0.0905	0.4288	0.1424
3	1	1	3	0.4940	0.0120	0.4884	0.0232

$\boldsymbol{s}_1 = (2, 2, 0)$. In the first scenario, each treatment had one success and one failure, leading to a posterior that is again symmetric across treatments. In this scenario, shown in the top left of Figure 2.2, it is optimal to assign two units to either one of the three treatments, and one unit to each of the other two arms.

In the second scenario, treatment 3 performed better than treatments 1 and 2. In this scenario, shown in the top right of Figure 2.2, it is optimal to assign three units to treatment 3, and one unit to either one of the other two arms. In the third and fourth scenario, treatment 3 performed worse than treatments 1 and 2. In these scenarios, shown in the bottom part of Figure 2.2, it is optimal to assign no units to treatment 3, three units to either one of treatment 1 or 2, and one unit to the other treatment. Interestingly, this dominates (though not by much) the assignment of two units to each of treatment 1 and 2.

We can compare these with the exploration sampling assignment probabilities. Table 1 lists the Beta distribution parameters along with the Thompson shares and exploration shares for each scenario. The Thompson sampling share for the third treatment (which in each scenario has different numbers of successes from the first two treatments) is given by

$$\int_0^1 F(x, \alpha_{1,2}, \beta_{1,2})^2 \cdot f(x, \alpha_3, \beta_3) \mathrm{d}x,$$

where

$$f(x, \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad F(x, \alpha, \beta) = \frac{B(x; \alpha, \beta)}{B(\alpha, \beta)}.$$

Discussion. In each of these examples, the largest number of units is assigned to the treatment arms with the highest expected return. In addition, one unit is assigned to at least one close competitor. This reflects that more precise effect estimates for treatment arms with low expected return are less likely to affect the ultimate policy decision. The shift towards more successful treatments occurs even though our objective function does not assign any weight to the welfare of experimental units, because there is no exploitation motive. This property is mimicked by the exploration sampling algorithm.

For this small sample, there are also properties that exploration sampling does not replicate. In particular, an interesting feature is that a symmetric assignment is generally not optimal, even when two treatments have the same current prior. Exploration sampling then produces

equal shares for those two treatments. However, in the second to fourth scenario above, the prior distribution for treatments 1 and 2 is the same, but the optimal design assigns either more units to treatment 1 or to 2. This reflects a non-convexity in the value of information, due to the concave objective function $\max_d \left(E[\theta^d|\boldsymbol{m}_T, \boldsymbol{s}_T] - c^d\right)$. This situation is analogous to option pricing, where higher volatility can increase the value of a stock option which is only exercised for high profit realizations.

3 Details for the calibrated simulations

In section 5 of the paper, we use data from three real experiments to conduct calibrated simulations of the various algorithms we consider. Here we describe these experiments in more detail and show some additional graphs of the policy regret distributions.

The experiments used for calibration. Ashraf et al. (2010) conducted a field experiment with about 1,000 households in Lusaka, Zambia. During a door-to-door sale of Clorin, a water disinfectant, each participating household was offered to buy a bottle at a randomly chosen price, ranging from 300 to 800 Zambian Kwacha. The study varied the offer price as well as the actual purchase price and measured the ex-post uptake of Clorin for water disinfection at different purchase prices, in order to test for the presence of a sunk-cost effect. The outcome we consider here is the 'first stage', that is, whether the household agreed to buy the bottle of Clorin at the original offer price.

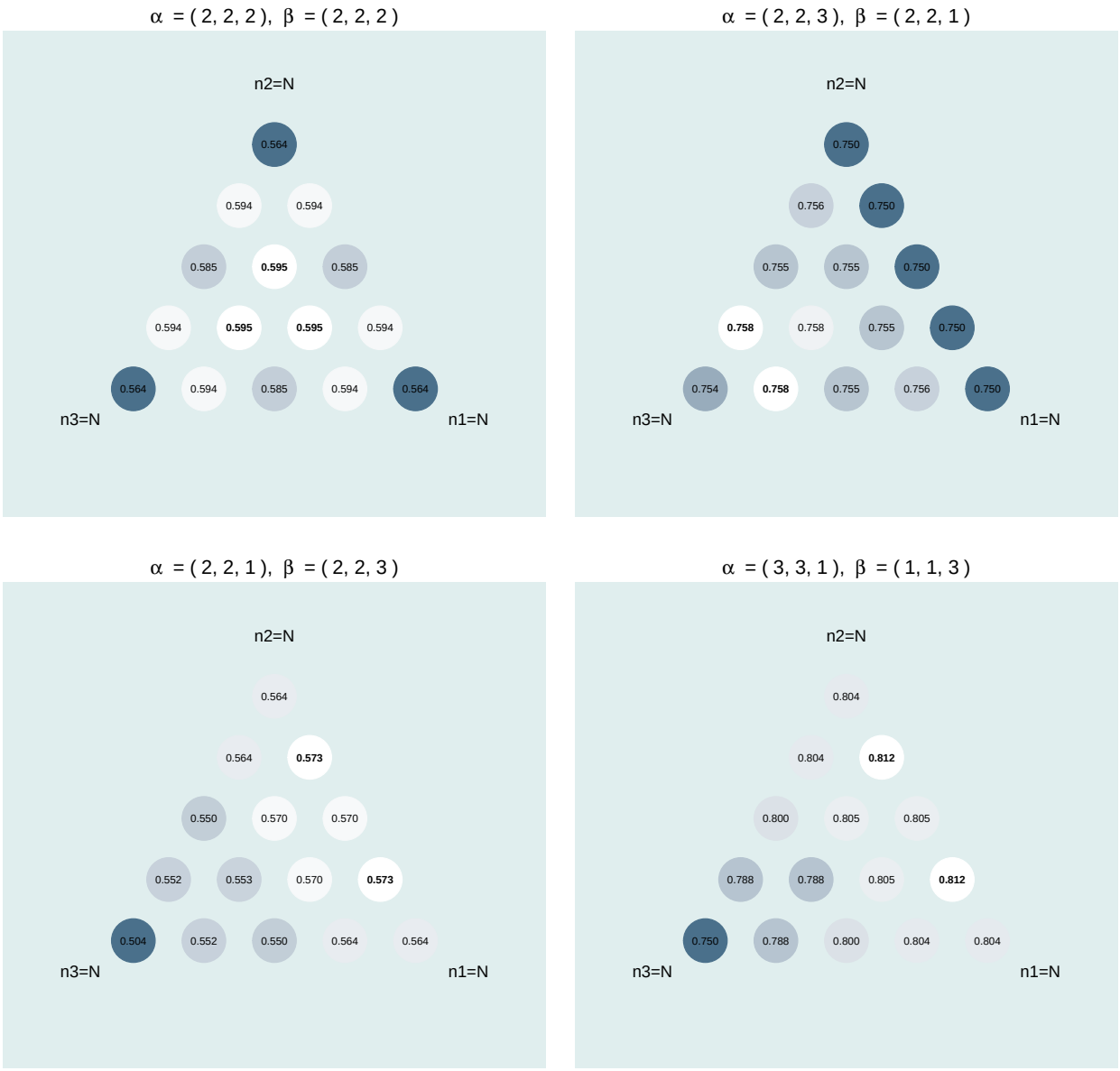
Bryan et al. (2014) conducted a field experiment in rural Bangladesh. Households were randomly assigned a cash or credit incentive of $8.50 (an amount covering round-trip travel), or an information treatment, conditional on a household member migrating during the 2008 monga (lean) season. The outcome we focus on is again take-up, i.e. whether at least one household member migrated (the first stage of the original paper).

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

Plots of simulation results. Figures 3.1 to 3.3 compare the distribution of regret between *non-adaptive* assignment and *exploration sampling* with probability mass functions (histograms) and quantile functions, for two, four, and ten experimental waves.

The uniformly lower quantile function for exploration sampling, relative to non-adaptive assignment, implies that its distribution of regret is first-order stochastically dominated. The integrated difference between the two quantile functions equals the decrease in average regret (increase in average welfare) that is gained from switching to exploration sampling.

Figure 2.2: Expected welfare as a function of treatment assignment



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

Figure 3.1: The distribution of policy regret (top) and regret quantiles (bottom) in Ashraf et al. (2010).

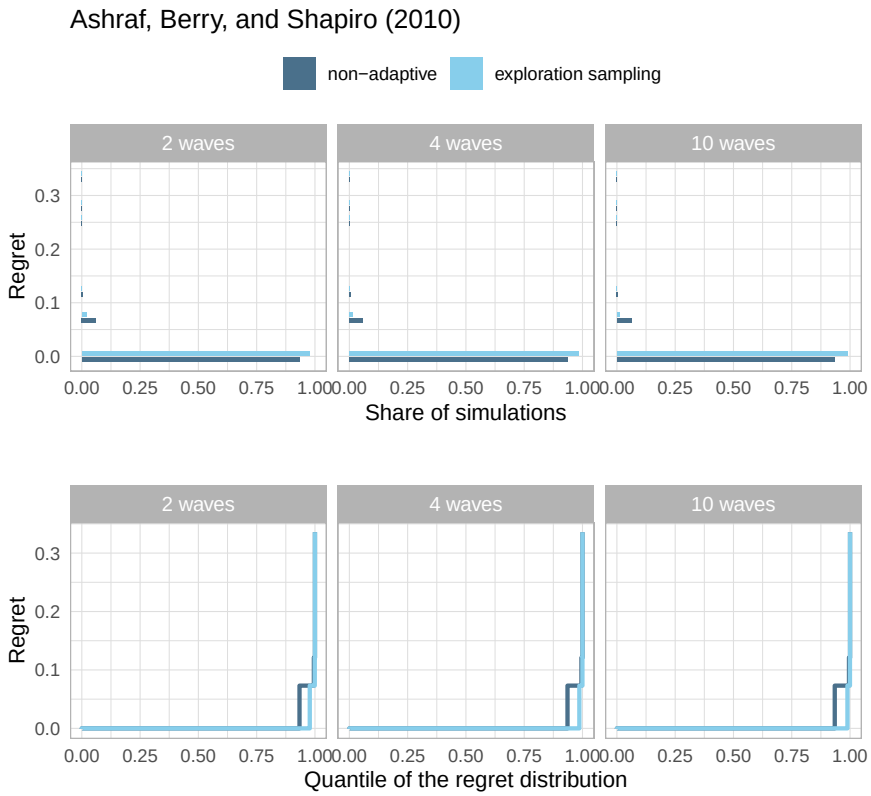


Figure 3.2: The distribution of policy regret (top) and regret quantiles (bottom) in Bryan et al. (2014).

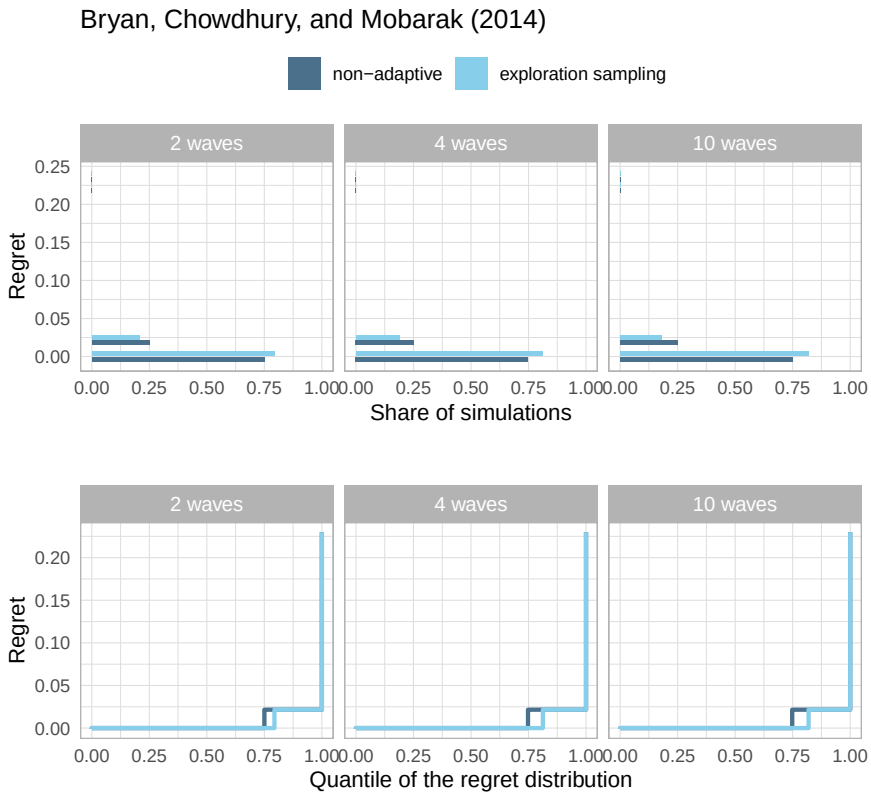
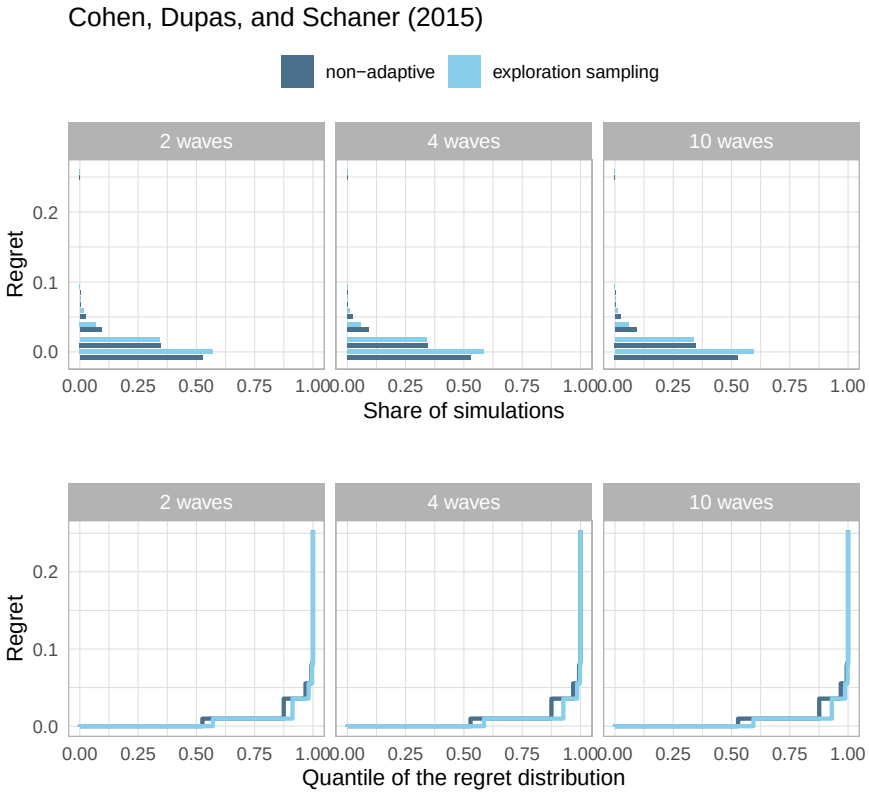


Figure 3.3: The distribution of policy regret (top) and regret quantiles (bottom) in Cohen et al. (2015).



References

Ashraf, N., Berry, J., and Shapiro, J. M. (2010). Can higher prices stimulate product use? Evidence from a field experiment in Zambia. *American Economic Review*, 100(5):2383–2413.

Bryan, G., Chowdhury, S., and Mobarak, A. M. (2014). Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh. *Econometrica*, 82(5):1671–1748.

Cohen, J., Dupas, P., and Schaner, S. (2015). Price subsidies, diagnostic tests, and targeting of malaria treatment: evidence from a randomized controlled trial. *American Economic Review*, 105(2):609–45.

Erickson, J. (2019). *Algorithms*.