

Adaptive Combinatorial Allocation

PRELIMINARY AND INCOMPLETE

Maximilian Kasy*

Alexander Teytelboym[†]

October 15, 2020

Abstract

We consider settings where an allocation has to be chosen repeatedly, returns are unknown but can be learned, and decisions are subject to constraints. Our model covers two-sided and one-sided matching, even with complex constraints. We propose an approach based on Thompson sampling. Our main result is a prior-independent finite-sample bound on the expected regret for this algorithm. Although the number of allocations grows exponentially in the number of participants, the bound does not depend on this number. We illustrate the performance of our algorithm using data on refugee resettlement in the United States.

1 Introduction

Adaptive experimentation uses information obtained in the course of an experiment in order to optimize the treatment assignment for later study participants.¹ For example, if job seekers arrive at a job center over time, a policymaker can use the outcomes of earlier job seekers in order to improve the assignment of labor market interventions for later participants (Caria et al., 2020). Using this information, adaptive experimentation can for instance be used to maximize the welfare of study participants, or to inform subsequent policy choices (Kasy and Sautmann, 2020).

In many policy settings, however, policymakers do not simply choose between a few interventions. Instead, they need to select an *entire allocation of resources* among participants. These resources are typically scarce, and feasible allocations can be described by combinatorial constraints. For example, if the policymaker wants to allocate students to classrooms when classroom composition affects student outcomes (Graham et al., 2010), she must ensure that all students are assigned to classrooms, that the capacity of classrooms is not exceeded, and that the allocation respect the demographic composition of students in the population. If the policymaker wants to allocate children to foster families when families impact the outcomes of the children, she needs to ensure that siblings are placed together and that foster homes are

*Department of Economics, University of Oxford, maximilian.kasy@economics.ox.ac.uk.

[†]Department of Economics, University of Oxford, alexander.teytelboym@economics.ox.ac.uk.

¹We thank Daniel Privitera and Manos Perdikakis for excellent research assistance.

close to schools and family homes (MacDonald, 2019; Robinson-Cortés, 2019). If the policy-maker wants to allocate tenants to social housing, she needs to ensure that housing matches the needs of tenants and respects waiting-list priorities (Thakral, 2016; Waldinger, 2018; van Dijk, 2019). If the policymaker wants to allocate combinations of therapies to different patients in order to overcome a disease, she needs to ensure that the therapies are actually available at the appropriate time and can be combined.

Combinatorial resource constraints might make adaptive experimentation more difficult relative to the unconstrained case, as considered in the multi-armed bandit literature, since the number of possible allocations can be vast. For example, the number of ways to allocate students to classrooms grows exponentially in the number of students. This might cause both computational difficulties (requiring optimization over a large discrete space), and statistical difficulties (the expected rewards for many different allocations have to be learned). We show, however, that remarkably despite these difficulties, performance close to the unconstrained case can be achieved. In this paper we consider an adaptive allocation policy extending the idea of Thompson sampling (Thompson, 1933). This policy is close to optimal for maximizing the outcomes of experimental participants in the presence of combinatorial resource constraints.

The general setting we consider is the following. The decision-maker has access to a finite number of *options* (e.g., matches) but is constrained to selecting only allocations (combinations of options; e.g., matchings) that satisfy the resource constraints (e.g., a one-to-one matching). Participants arrive in *batches* every period. The decision-maker selects an allocation and observes the outcome of each selected option. The outcome of each option results in a *reward*. The decision-maker’s objective is to maximize the expected cumulative rewards from all the options she picked over time; equivalently, the decision-maker aims to minimize expected *regret*, i.e., the expected difference between the optimal combination of options in each period. This type of setting has been called a *combinatorial semi-bandit* setting with *linear* rewards in the literature (Audibert et al., 2011). “Combinatorial” because the decision-maker can choose combinations of options; “semi-bandit” because the decision-maker can observe the outcomes of every match, not just of the entire matching; and “linear rewards” because the objective function is the sum of the rewards of all matches made.

Our main theoretical result is a bound on the worst-case regret obtained by using Thompson sampling in our setting. Thompson sampling is a classic heuristic for standard bandit problems; it requires that each action is picked with probability equal to the posterior probability that this action is optimal. Our theoretical result is appealing for three reasons. First, the worst-case expected regret does not depend on the batch size even though the number of possible actions (i.e., allocation) grows exponentially in the batch size. Second, our bound holds in finite-samples and does not rely on asymptotic approximations. Third, our bound is prior-independent allowing for arbitrary statistical dependence across match outcomes; this is particularly relevant for matching contexts.²

²This contrasts with existing bounds which require prior independence across options, as discussed below.

We apply our approach to the problem of matching resettled refugees to local communities in the United States (Bansak et al., 2018; Trapp et al., 2020). Our data cover the placement of all refugees by HIAS, an American resettlement agency, between 2014 and 2020. Our objective is to maximize the probability of employment of refugees in the first three months after their arrival. The allocation of refugees to local communities is subject to constraints. Local communities have a quota on the total number of refugees they can resettle in a given year, and refugees who come from the same family must be placed together. The implied optimization problem is equivalent to a “multiple knapsack problem” (Trapp et al., 2020; Delacrétaz et al., 2019).

Our paper is closely related to the literature on multi-armed bandit problems. Rather than attempting to characterize analytical solutions (e.g., Gittins (1979)), we focus on analyzing properties of the well-known probability matching heuristic due to Thompson (1933). Agrawal and Goyal (2012) and Kaufmann et al. (2012) have shown, for the fixed parameter case, that the asymptotic bound on expected regret of the Thompson algorithm in bandit settings matches the lower bound on regret for *any* bandit algorithm, which was derived by (Lai and Robbins, 1985). Wang and Chen (2018) provide a distribution-dependent regret bound for the Thompson algorithm in the *combinatorial semi-bandit* setting; our result is distribution-free. Adaptive experimentation using the Thompson algorithm has been proposed for applications such as drug trials (Berry, 2006), recommender systems (Kawale et al., 2015), and customer acquisition (Schwartz et al., 2017). More recently, adaptive experimentation has been deployed in field experiments in development contexts (Kasy and Sautmann, 2020; Caria et al., 2020).

The rest of the paper is organized as follows. Section 2 describes our combinatorial semi-bandit setting and the Thompson heuristic; Section 2.1 then discusses several examples covered by this general framework. Section 3 gives our main theoretical result and the intuition for its proof. Section 4 covers several considerations for implementation in practice, including the choice of model and prior, methods for sampling from the posterior, and statistical inference. Section 5 discusses calibrated simulations based on the motivating applications. Appendix A provides a brief review of information theory, which is need for the proof of our main result. All proofs can be found in Appendix B.

2 Setup

We denote all random variables with capital letters (e.g., A) and the realizations of random variables with lower-case letters (e.g., a).

Feasible options and actions The decision-maker has access to *options* (e.g., matches or projects) $j \in 1, \dots, J$, but only has sufficient resources to select $M \leq J$ of these. We denote by $\mathcal{A} \subseteq \{a \in \{0, 1\}^J : \|a\|_1 = M\}$ a collection of *feasible* combinations of options. \mathcal{A} is a strict subset if the decision-maker faces additional allocation constraints. The decision-maker’s *action* $a \in \mathcal{A}$ is a feasible combination of options.

Timing, potential outcomes, and observability The program takes places in a finite number of periods $t = 1, \dots, T$. In each period, there is a vector $Y_t \in [0, 1]^J$ of potential outcomes, where Y_{jt} is the potential outcome for option j in period t . The vectors Y_t are i.i.d. across periods. We denote the average potential outcome (or average structural function) for option j by Θ_j , that is, $\Theta_j = \mathbf{E}[Y_{jt}|\Theta]$. The decision-maker holds a prior belief over the vector $\Theta \in [0, 1]^J$, where we allow for arbitrary dependence of this prior across the options j .

In each period, the decision-maker chooses an action $A_t \in \{0, 1\}^J$. If the decision-maker chooses action a , they observe the outcomes of the chosen options j (the options j for which $a_j = 1$), i.e., the vector

$$Y_t(a) = (a_j \cdot Y_{jt} : j = 1, \dots, J). \quad (1)$$

We assume “stable unit treatment values” (i.e., no spillovers or interference) across options j , in the sense that Y_{jt} does not depend on the chosen action $a_{j't}$ for any j' . Note that this assumption is consistent with settings where Y_{jt} is itself the equilibrium outcome of interactions across multiple individuals comprising an option j , as is the case for example in the applications to peer effects or matching discussed below.

Given our assumption about observability, the *information* available at the beginning of period t is given by

$$\mathcal{F}_t = \{(A_{t'}, Y_{t'}(A_{t'})) : 1 \leq t' < t\}. \quad (2)$$

Throughout this paper, the subscript t on \mathbf{E}_t indicates that the expectation is evaluated under the posterior distribution $\mathbf{P}_t(\cdot) = \mathbf{P}(\cdot \mid \mathcal{F}_t)$, where \mathcal{F}_t is the information available at the beginning of period t . The decision-maker can choose their action A_t at the beginning of each period t based on the information \mathcal{F}_t , as well as possibly based on a randomization device that is statistically independent across periods and independent of the sequence of potential outcomes $(Y_t)_{t=1}^T$.

Objective and policy If the decision-maker chooses action a in period t , they receive a *reward* which is equal to $\langle a, Y_t \rangle$. Therefore, upon taking action a the decision-maker’s *expected reward* given Θ , which is the same across periods t , equals

$$R(a) = \mathbf{E}_t[\langle a, Y_t \rangle | \Theta] = \langle a, \Theta \rangle. \quad (3)$$

The decision-maker would like to maximize cumulative expected rewards,

$$\mathbf{E}_1 \left[\sum_{t=1}^T R(A_t) \right]. \quad (4)$$

The expectation in this expression is taken over the randomness in the choice of actions A_t , the sampling distribution of potential outcomes Y_t , and over the prior distribution of Θ . Denote by A^* a feasible action that maximizes the expected reward conditional on Θ (but not conditional

on the vector Y_t), that is,

$$A^* \in \operatorname{argmax}_{a \in \mathcal{A}} R(a) = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \Theta \rangle. \quad (5)$$

The decision-maker’s objective is equivalent to minimizing expected *regret* at T

$$\mathbf{E}_1 \left[\sum_{t=1}^T (R(A^*) - R(A_t)) \right]. \quad (6)$$

Solving this dynamic stochastic combinatorial optimization problem is computationally quite costly. Rather than solving it, we propose that the decision-maker adopt the following heuristic policy. In each period the decision-maker should take an action a from the feasible set \mathcal{A} according to the posterior probability that a is optimal, that is, for each $a \in \mathcal{A}$,

$$\mathbf{P}_t(A_t = a) = \mathbf{P}_t(A_t^* = a). \quad (7)$$

This assumption implies in particular that $\mathbf{E}_t[A_t] = \mathbf{E}_t[A^*]$. This heuristic approach is known as Thompson sampling, and was originally introduced by Thompson (1933) for treatment assignment in adaptive experiments.

2.1 Examples

In the following we discuss several examples that are covered by our general framework, and thus in particular by the regret bound provided in Theorem 1 below. These examples correspond to practically relevant policy problems. They also illustrate how various combinatorial allocation problems that have been studied in the literature fit into our framework, such as assignment to peers, one-to-one matching, many-to-one matching, knapsack problems, etc.

For each of these examples, several options might correspond to same underlying parameter, so that $\Theta_j = \Theta_{j'}$ with prior probability 1, for some j, j' . In the case of one-to-one matching, for instance, each matched pair corresponds to one option, but Θ_j is the same for all matched pairs j with the same observed covariates on both sides of the match.

Example 1 (Allocation of refugees to local communities). American refugee resettlement agencies need to make weekly decisions about the allocation of arriving refugee families to local communities. An action a is a matching of refugee families to local communities. The number of options J is the number of distinct matches between different family-locality pairs, and the batch size M is equal to the number of refugee families arriving in a given week. We will consider this example in greater detail in Section 5 below.

Example 2 (Foster parent allocation). Foster families are typically able to host several foster children at the same time (MacDonald, 2019; Robinson-Cortés, 2019). An action a is a *many-to-one matching* between families and children. The feasible actions a require that no family receives more children than it can host, that all siblings are matched to the same foster family,

and that children are hosted near their school and activities. The parameters Θ_j are again perfectly dependent across options j that are observationally identical, i.e., across matches of children and families with the same observed covariates.

Example 3 (Peer effects and classroom composition). Suppose that a policymaker would like to choose the gender composition of classrooms in order to maximize student performance (Graham et al., 2010). Assume students are of two types, boys and girls. Classrooms have a fixed number of students. An action a allocates (i.e., groups) the students into classrooms. Classroom identity does not matter, but the identity of peers does matter, for student outcomes. The number of options J is equal to the number of classroom-sized subsets of the set of all students. The batch size M is equal to the number of classrooms. If students are observationally indistinguishable from each other, except for gender, then the prior exhibits perfect dependence across classrooms with the same number of girls and boys.

3 Performance guarantee

We next state our main theoretical result. This result provides tight worst-case guarantees for the expected regret of Thompson sampling in our setup.

Theorem 1. *Under the assumptions of Section 2,*

$$\mathbf{E}_1 \left[\sum_{t=1}^T (R(A^*) - R(A_t)) \right] \leq \sqrt{\frac{1}{2} J T M \cdot [\log(\frac{J}{M}) + 1]}.$$

Discussion of Theorem 1 Several features of the regret bound in Theorem 1 are worth emphasizing. First, this bound is a finite sample bound. There is no large sample limit and no remainder term. Second, this bound does not depend on the prior distribution for Θ in any way. Furthermore, it allows for prior distributions with arbitrary statistical dependence across the components of Θ , as required by our motivating examples. Third, this bound implies that Thompson sampling in our setting achieves the efficient rate of convergence for regret: As shown by Audibert et al. (2014), the minimax regret in our setting grows at a rate of $\sqrt{J T M}$, up to logarithmic terms.

Theorem 1 bounds the worst case expected regret across all possible priors, summed across units. To get the worst case expected regret per unit, divide this expression by $T M$, which yields the bound $\sqrt{J \cdot [\log(\frac{J}{M}) + 1] / (2 T M)}$. This bound goes to 0 at a rate of 1 over the square root of the sample size, that is, at a rate of $1/\sqrt{T M}$. The theorem furthermore shows that this worst case expected regret grows, as a function of the number of possible options J , like \sqrt{J} (neglecting the logarithmic term). Remarkably, worst case regret does not grow in the batch size M . This is despite the fact that the setup of Section 2 allows for action sets of size $\binom{J}{M}$. For comparison, application of the worst case regret bound for Thompson sampling in bandits with dependent arms provided by Proposition 3 in Russo and Van Roy (2016) yields a

much larger bound which grows in proportion to $\sqrt{\binom{J}{M} \log \binom{J}{M}}$. Instead, the regret bound in Theorem 1 grows like that for a simple multiarmed bandit with J arms.

Intuition for the proof of Theorem 1 The proof of Theorem 1 is provided in Appendix B. This proof builds on several definitions and standard results from information theory which are reviewed in Appendix A. Here we just sketch some of the key steps in our proof.

First, we use Pinsker’s inequality in order to relate expected regret to the information about the optimal action A^* provided by observations, where information is measured by the KL-distance of posteriors and priors. Pinsker’s inequality implies, for Bernoulli random variables B and B' , that $(\mathbf{E}[B] - \mathbf{E}[B'])^2 \leq \frac{1}{2} D_{KL}(B, B')$. Lemma 1 applies Pinsker’s inequality to terms showing up in the definition of expected regret which are of the form $\mathbf{E}_t[\Theta_j | A_j^* = 1] - \mathbf{E}_t[\Theta_j]$. This use of Pinsker’s inequality is at the core of the proofs in Russo and Van Roy (2016).

Second, following some of the ideas introduced in Bubeck and Sellke (2020), Lemma 2 relates the KL-distance to the entropy of the events $A_j^* = 1$. The combination of these two Lemmas allows to bound the expected regret for option j in terms of the entropy reduction for the posterior of A_j^* .

Third and lastly, Lemma 3 shows that the total reduction of entropy across the options j , and across the time periods t , can be no more than the sum of the prior entropy for each of the events $A_j^* = 1$, which is bounded by $M \cdot [\log \binom{J}{M} + 1]$. The proof of Theorem 1 then combines these three Lemmas.

Relationship to the literature Our proof builds on the information theoretic approach pioneered by Russo and Van Roy (2016) (in particular Lemma 1 and 2 as well as Proposition 6 therein) and some variations of this approach proposed by Bubeck and Sellke (2020) (in particular Lemma 13). Despite the close relationship to their arguments, Theorem 1 differs from the bounds provided in these papers as follows. The closest result in Russo and Van Roy (2016) is their Proposition 6. Their result, however, requires statistical independence of the prior and posterior distribution for the components of Θ at all times t . By contrast, Theorem 1 above allows for arbitrary dependence. This is especially relevant for the matching setting, where independence in the prior distribution would be quite hard to justify.

The closest result in Bubeck and Sellke (2020) is their Theorem 21. Their result is asymptotic, rather than providing an exact finite sample bound. The main interest of Bubeck and Sellke (2020) is an asymptotic refinement of regret bounds that scales in the best achievable regret, allowing for the latter to converge to 0; this is something which our result does not aim to do.

4 Implementation of Thompson sampling for matching problems

4.1 Model and prior for matching settings

In order to achieve good performance in practice, our proposed procedure relies on specifying an appropriate model for the data generating process, and an appropriate prior distribution for the underlying parameters. We generally advocate for the use of default priors that are diffuse and symmetric across types, while incorporating reasonable assumptions about the dependency structure between different options j .

Table 1 proposes some variants of models and priors for matching settings, covering our leading motivating examples, including those used in our empirical applications. For each of these variants, we assume that the options j consist of two-sided matches between types u_j and types v_j . For each possible match (option), the potential outcomes Y_{jt} are drawn from some distribution with mean Θ_j . We need to specify this distribution of Y_{jt} , as well as a joint prior distribution of the parameters Θ_j across j .

Each of these models assumes that the option-effect Θ_j is determined by the sum of type-effects $\Gamma^u_{u_j}$ and $\Gamma^v_{v_j}$, plus an interaction effect $\Gamma^{uv}_{u_j, v_j}$. For continuous outcomes, we assume that Θ_j is directly given by this sum. For binary or discrete outcomes, we assume that Θ_j is given by the logit link function applied to this sum.

For the model for outcomes with discrete bounded support, the distribution of Y_{jt} is governed by the mean parameter Θ_j as well as a dispersion parameter m . The latter is necessary to allow for larger dispersions relative to a more restrictive Binomial model, which might put excessive weight on the information content of single observations.

4.2 Inference

In order to implement Thompson sampling, we need to sample from the posterior for Θ . This posterior is also relevant for statistical inference on parameter values. Such inference is often a secondary goal, in addition to the primary goal of maximizing participant outcomes. Such inference might be Bayesian, using the same posterior distributions that go into the assignment algorithm. Alternatively, such inference might be based on permutation tests, as described below.

Markov Chain Monte Carlo and Bayesian inference For hierarchical priors, such as those discussed in Section 4.1, posterior distributions are not available in closed form, in general. We can, however, sample from the posterior for Θ using Markov Chain Monte Carlo (MCMC) methods. Such MCMC methods only require us to specify the posterior up to a multiplicative constant (typically, up to the denominator of the posterior density, which is given by the marginal density of the observed data). MCMC methods are based on constructing a Markov

Table 1: Models and priors for matching

Continuous outcomes

$$\begin{aligned}
Y_{jt} &\sim N(\Theta_j, \sigma^2) \\
\Theta_j &= \Gamma^u_{u_j} + \Gamma^v_{v_j} + \Gamma^{uv}_{u_j, v_j} \\
\Gamma^u_{u_j} &\sim N(0, \tau_{\Gamma^u}^2), \quad \Gamma^v_{v_j} \sim N(0, \tau_{\Gamma^v}^2), \quad \Gamma^{uv}_{u_j, v_j} \sim N(\mu, \tau_{\Gamma^{uv}}^2),
\end{aligned}$$

Binary outcomes

$$\begin{aligned}
Y_{jt} &\sim \text{Bernoulli}(\Theta_j) \\
\Theta_j &= \frac{1}{1 + \exp(-(\Gamma^u_{u_j} + \Gamma^v_{v_j} + \Gamma^{uv}_{u_j, v_j}))} \\
\Gamma^u_{u_j} &\sim N(0, \tau_{\Gamma^u}^2), \quad \Gamma^v_{v_j} \sim N(0, \tau_{\Gamma^v}^2), \quad \Gamma^{uv}_{u_j, v_j} \sim N(\mu, \tau_{\Gamma^{uv}}^2),
\end{aligned}$$

Discrete outcomes with bounded support $\{0, \dots, \bar{y}\}$

$$\begin{aligned}
Y_{jt} &\sim \text{Beta-Binomial}(\alpha_j, \beta_j, \bar{y}) \\
A_j &= m \cdot \Theta_j, \quad B_j = m \cdot (1 - \Theta_j) \\
\Theta_j &= \frac{1}{1 + \exp(-(\Gamma^u_{u_j} + \Gamma^v_{v_j} + \Gamma^{uv}_{u_j, v_j}))} \\
\Gamma^u_{u_j} &\sim N(0, \tau_{\Gamma^u}^2), \quad \Gamma^v_{v_j} \sim N(0, \tau_{\Gamma^v}^2), \quad \Gamma^{uv}_{u_j, v_j} \sim N(\mu, \tau_{\Gamma^{uv}}^2),
\end{aligned}$$

Notes: For each of these cases we assume that the components of $\Gamma^u, \Gamma^v, \Gamma^{uv}$ are mutually independent given the hyper-parameters. The hyper-parameters are given by $\sigma^2, \tau_{\Gamma^u}^2, \tau_{\Gamma^v}^2, \tau_{\Gamma^{uv}}^2$ and μ for continuous outcomes, by $\tau_{\Gamma^u}^2, \tau_{\Gamma^v}^2, \tau_{\Gamma^{uv}}^2$ and μ for binary outcomes, and by $\tau_{\Gamma^u}^2, \tau_{\Gamma^v}^2, \tau_{\Gamma^{uv}}^2$ and μ for discrete outcomes with bounded support. We propose to use some diffuse prior for these hyper-parameters.

Chain which converges to an ergodic distribution that is given by the posterior of interest. There are various ways of constructing such Markov Chains; one of them is Hamiltonian Monte Carlo. In our applications, we sample from the posterior using Hamiltonian Monte Carlo as implemented in the software STAN (Carpenter et al., 2017).

Let $\hat{\Theta}_t$ be a draw from the posterior given \mathcal{F}_t generated by MCMC, after a sufficiently long warm-up period. Choose

$$A_t = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \hat{\Theta}_t \rangle. \quad (8)$$

Then A_t follows the distribution required for Thompson sampling, that is, it satisfies Equation (7).

In order to form $1 - \alpha$ credible sets for the parameters Θ_j given the history \mathcal{F}_t , sample a large number of draws $\hat{\Theta}_t$ from the posterior, and form a credible interval based on the $\alpha/2$ and $1 - \alpha/2$ quantiles of $\hat{\Theta}_{j,t}$ across these draws.

Randomization inference An alternative to Bayesian inference is randomization (permutation) inference. In the context of treatment effect estimation, randomization inference can be used to test the null hypothesis that treatment does not affect any outcome, so that for instance $Y_i^1 = Y_i^0$ for all units i and treatment value 0, 1.

In the context of our setting, we need to modify this null hypothesis. Permutation inference requires that we specify the counterfactual outcome vector $Y_t^0(a)$ for any counterfactual action $a \in \mathcal{A}$ under the null hypothesis H^0 , given knowledge of $Y_t(A_t)$ for the realized action A_t . In many cases of interest, there might be more than one plausible way to specify such a null hypothesis and the corresponding counterfactual outcome vectors.

To illustrate, consider the case of one-to-one matching (of refugees to local communities, say), where each option j corresponds to a match of a refugee family to a local community. We could formalize the null hypothesis that “the matching does not matter” in two different ways. We could consider the hypothesis that refugee outcomes are the same, no matter which community they are allocated to. Or we could consider the hypothesis that outcomes in a community are the same, no matter which refugees are allocated to be there.

Given some specification of counterfactual outcomes, we can sample counterfactual histories $\tilde{\mathcal{F}}_t$ by re-running the Thompson sampling algorithm iteratively. In each period s , draw $\tilde{\Theta}_{t'}$ and the corresponding $\tilde{A}_{t'}$ from the posterior given $\tilde{\mathcal{F}}_{t'}$. Impute a counter-factual outcome vector $Y_{t'}^0(\tilde{A}_{t'})$, based on the null hypothesis to be tested. Update the history $\tilde{\mathcal{F}}_{t'}$ by adding $\tilde{A}_{t'}, Y_{t'}^0(\tilde{A}_{t'})$, and iterate for the next period. Once $t' = t$, calculate a realization of the test-statistic as a function of $\tilde{\mathcal{F}}_t$. Repeat this process to generate a sampling distribution of the test-statistic, and corresponding critical values and p-values for testing the null hypothesis under consideration.

5 Applications

5.1 Refugee resettlement

The United States has historically been the world’s largest destination of *resettled* refugees, with 78,340 admitted in 2016. There is a lot of evidence that the initial match between refugees and local communities dramatically affects the socioeconomic outcomes of refugees (Bansak et al., 2018; Trapp et al., 2020). However, local community capacities are tightly regulated by the US government. As a result, one US resettlement agency (HIAS) optimizes the placement of the resettled refugees using its recommendation system called Annie MOORE. However, Annie’s estimates of refugee employment are static and come from a LASSO regression run annually (Trapp et al., 2020).

Here, we draw on the data used by Annie MOORE in order to run calibrated simulations for our proposed procedure subject to realistic constraints, hoping to inform actual refugee placement by Annie MOORE in the future.

Data Our data covers all refugees resettled by HIAS between October 2005 and September 2020. There were a total of 37,149 refugees that constitute 15,523 cases (i.e., families). We focus on the employment of primary applicants in each family. For each primary applicant in the family, we observe three binary variables: whether the applicant is of prime working age (25-54), their gender, and whether they are English-speaking. We also observe the family size, and the affiliate where the family was resettled. We furthermore observe whether the primary applicant had any US ties. Applicants with US ties (e.g., US resident friends or family) are automatically resettled to the community where their US ties reside. Applicants without US ties can be resettled to any of the communities where HIAS operates. Finally, we observe whether or not the primary applicant was employed within 90 days of arrival. This is a key metric used by the State Department to assess the performance of American resettlement agencies.

There are 57 affiliates in our data. We drop any affiliate with fewer than 150 resettled cases over the whole period under consideration, leaving us with 17 affiliates.³ All affiliates are anonymized. Based on the available observables, we classify refugees into 8 “types” u , while treating each of the 17 affiliates as a separate “type” v . This means that there are $8 \cdot 17 = 136$ parameters (probabilities of finding employment) that we might wish to learn.

As noted above, the affiliates have a limited capacity in hosting refugees, where capacity is the total number of resettled people (primary applicants and their family members). The capacities of affiliates can sometimes change throughout the course of the year. For our simulations, we conservatively set the available annual capacities to be 110% of the total number of refugees without US ties actually resettled to each affiliate in a given year.⁴ The monthly

³In their analysis, Trapp et al. (2020) also pool some affiliates because of small numbers of observations.

⁴Resettlement agencies are allowed to exceed their official capacity by 10% without further approval; however, in practice, agencies occasionally have to seek approval for capacity extensions and reductions because of stochastic refugee flows. Dynamic capacity management is beyond the scope of this paper.

quota for each affiliate is therefore 110% of its annual capacity divided by 12.

Simulation design Our simulated matching process works as follows. For each month t in the available data, we consider all the refugees who were resettled by HIAS in this month. We match the refugees with US ties to their actual affiliates. For all the refugees without US ties, we match them to affiliates using the Thompson algorithm discussed above. This matching has to satisfy the capacity constraints of affiliates described above, and the varying sizes of refugee families. Solving for the optimal matching for a given draw $\hat{\Theta}$ from the posterior defines a so-called multiple knapsack problem. Any leftover capacity and any unmatched refugees without US ties are carried over to the next month.

We then impute counterfactual outcomes for the refugees allocated using the Thompson algorithm, based on calibrated parameter values Θ_0 . These calibrated parameter values are set to the posterior mean for the Bayesian hierarchical model for binary outcomes described in Section 4.1.

Results WORK IN PROGRESS.

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.
- Audibert, J.-Y., Bubeck, S., and Lugosi, G. (2011). Minimax policies for combinatorial prediction games. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 107–132.
- Audibert, J.-Y., Bubeck, S., and Lugosi, G. (2014). Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45.
- Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., and Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science*, 359(6373):325–329.
- Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery*, 5(1):27–36.
- Bubeck, S. and Sellke, M. (2020). First-order bayesian regret analysis of thompson sampling. In *Algorithmic Learning Theory*, pages 196–233.
- Caria, S., Gordon, G., Kasy, M., Osman, S., Quinn, S., and Teytelboym, A. (2020). Adaptive treatment assignment in experiments for policy choice. Mimeo.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).
- Delacrétaz, D., Kominers, S. D., Teytelboym, A., et al. (2019). Matching mechanisms for refugee resettlement. Technical report.
- Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164.
- Graham, B. S., Imbens, G. W., and Ridder, G. (2010). Measuring the effects of segregation in the presence of social spillovers: a nonparametric approach. Technical report, National Bureau of Economic Research.
- Kasy, M. and Sautmann, A. (2020). Adaptive treatment assignment in experiments for policy choice. *Econometrica*, forthcoming.
- Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer.

- Kawale, J., Bui, H. H., Kveton, B., Tran-Thanh, L., and Chawla, S. (2015). Efficient thompson sampling for online matrix-factorization recommendation. In *Advances in neural information processing systems*, pages 1297–1305.
- Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.
- MacDonald, D. E. (2019). Foster care: A dynamic matching approach. Technical report, Mimeo.
- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.
- Robinson-Cortés, A. (2019). Who gets placed where and why? an empirical framework for foster care placement.
- Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471.
- Schwartz, E. M., Bradlow, E. T., and Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522.
- Thakral, N. (2016). The public-housing allocation problem: Theory and evidence from pittsburgh. Technical report, Harvard University.
- 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.
- Trapp, A. C., Teytelboym, A., Martinello, A., Andersson, T., Ahani, N., et al. (2020). Placement optimization in refugee resettlement. *Operations Research*.
- van Dijk, W. (2019). The socio-economic consequences of housing assistance. Technical report, University of Chicago Kenneth C. Griffin Department of Economics.
- Waldinger, D. (2018). Targeting in-kind transfers through market design: A revealed preference analysis of public housing allocation. Technical report, New York University.
- Wang, S. and Chen, W. (2018). Thompson sampling for combinatorial semi-bandits. *arXiv preprint arXiv:1803.04623*.

A A brief review of information theory

In this section we review some basic definitions and facts about entropy, mutual information, and KL-divergence. For further background, see MacKay (2003) (in particular chapter 8), as well as Section 3 in Russo and Van Roy (2016). For our purposes, it is enough to restrict attention to the Bernoulli case, so that we can introduce the following definitions in elementary form. Let A be a Bernoulli random variable with expectation p , and let A' be a Bernoulli random variable with expectation q . We overload notation by allowing the arguments A and p to be used interchangeably.

- **Entropy:**

$$H(A) = H(p) = -[p \log(p) + (1-p) \log(1-p)]. \quad (\text{A1})$$

- **KL divergence:**

$$D_{KL}(A, A') = D_{KL}(p, q) = p \log\left(\frac{p}{q}\right) + (1-p) \log\left(\frac{1-p}{1-q}\right). \quad (\text{A2})$$

- **Pinsker's inequality:**

$$(\mathbf{E}[A] - \mathbf{E}[A'])^2 = |p - q| \leq \frac{1}{2} D_{KL}(p, q) = \frac{1}{2} D_{KL}(A, A'). \quad (\text{A3})$$

- **Mutual information** as expected divergence of the posterior:

For any random variable or vector F , let $p(f) = \mathbf{E}[A|F = f]$. Then

$$I(A; F) = \mathbf{E}[D_{KL}(p(F), p)]. \quad (\text{A4})$$

- **Conditional entropy:**

$$H(A|F) = \mathbf{E}[H(p(F))]. \quad (\text{A5})$$

- **Entropy reduction** form of mutual information:

$$I(A; F) = H(A) - H(A|F). \quad (\text{A6})$$

- **Data processing inequality:** For any transformation $g(F)$ of a random variable or vector F ,

$$I(A; g(F)) \leq I(A; F). \quad (\text{A7})$$

- **Chain rule** of mutual information:

$$I(A; (F, G)) = I(A; F) + I(A; G|F). \quad (\text{A8})$$

B Proofs

For ease of reference, we begin by restating our notation and assumptions.

$Y_t, \Theta, A_t \in \mathbb{R}^J$	Outcome, parameter, and action vectors
$A_t \in \mathcal{A} \subseteq \{a \in \{0, 1\}^J : \ a\ _1 = M\}$	Feasible allocations and batch size
$Y_{jt} \in [0, 1]$	Bounded outcomes
$\Theta = \mathbf{E}_t[Y_t \Theta]$	Parameters are expectation of outcomes
$\bar{\Theta}_t = \mathbf{E}_t[\Theta] = \mathbf{E}_t[Y_t]$	Prior expectation of the parameter (at t)
$R(a) = \mathbf{E}_t[\langle a, Y_t \rangle \Theta] = \langle a, \Theta \rangle$	Linear (combinatorial) expected rewards
$Y_t(a) = (a_j \cdot Y_{jt} : j = 1, \dots, J)$	Observable outcomes (semi bandit)
$A^* \in \operatorname{argmax}_{a \in \mathcal{A}} R(a) = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \Theta \rangle$	Optimal action
$\bar{\Theta}_{jt}^* = \mathbf{E}_t[\Theta_j A_j^* = 1] = \mathbf{E}_t[Y_{jt} A_j^* = 1]$	Conditional expectation of parameters
$p_t = \mathbf{E}_t[A^*]$	Expected optimal action

For Thompson sampling we have that A_t has the same distribution as A^* , and therefore

$$\mathbf{E}_t[A_t] = \mathbf{E}_t[A^*] = p_t.$$

We next prove three preliminary Lemmas, before combining them in the proof of Theorem 1 itself.

Lemma 1 (Bounding regret by the component-wise information).

$$\mathbf{E}_t[R(A^*) - R(A_t)] \leq \sqrt{\frac{J}{2} \cdot \sum_{j=1}^J p_{jt}^2 \cdot D_{KL}(\bar{\Theta}_{jt}^*, \bar{\Theta}_{jt})}.$$

Proof of Lemma 1:

$$\mathbf{E}_t[R(A^*) - R(A_t)] = \mathbf{E}_t[\langle A^* - A_t, \Theta \rangle] \tag{B1}$$

$$= \langle p_t, \bar{\Theta}^* \rangle - \langle p_t, \bar{\Theta}_t \rangle \tag{B2}$$

$$\leq \sqrt{J \cdot \sum_{j=1}^J p_{jt}^2 \cdot (\bar{\Theta}_{jt}^* - \bar{\Theta}_{jt})^2} \tag{B3}$$

$$\leq \sqrt{\frac{J}{2} \cdot \sum_{j=1}^J p_{jt}^2 \cdot D_{KL}(\bar{\Theta}_{jt}^*, \bar{\Theta}_{jt})} \tag{B4}$$

These steps hold for the following reasons.

(B1) By definition of R .

(B2) By splitting the inner product, and using (i) iterated expectations, conditioning on $A_j^* = 1$ for each component j in turn, and (ii) independence of A_t and Θ_t and the definition of Thompson sampling.

(B3) By Cauchy Schwarz (for the inner product with a J -vector of 1s).

(B4) By Pinsker's inequality, applied to Bernoulli random variables with expectation $\bar{\Theta}_{jt}^*, \bar{\Theta}_{jt}$.

□

Lemma 2 (Divergence and component-wise information gain).

$$p_{jt}^2 \cdot D_{KL}(\bar{\Theta}_{jt}^*, \bar{\Theta}_{jt}) \leq I_t(A_j^*; Y_t(A_t), A_t).$$

Proof of Lemma 2:

For the purpose of this proof, construct a Bernoulli random variable \tilde{Y}_{jt} with expectation Y_{jt} , independently of everything else. Note that $\mathbf{E}_t[\tilde{Y}_{jt}] = \bar{\Theta}_{jt}$. $D_{KL}(\bar{\Theta}_{jt}^*, \bar{\Theta}_{jt})$ can be interpreted as the KL-divergence between the distribution of \tilde{Y}_{jt} conditional on $A_j^* = 1$ and the (unconditional) distribution of \tilde{Y}_{jt} . Taking the expectation over A_j^* of the KL-divergence yields the mutual information between A_j^* and \tilde{Y}_{jt} , $I_t(A_j^*; \tilde{Y}_{jt})$:

$$\begin{aligned} I_t(A_j^*; \tilde{Y}_{jt}) &= p_{jt} \cdot D_{KL}(\mathbf{E}_t[\Theta_{jt}|A_j^* = 1], \bar{\Theta}_{jt}) \\ &\quad + (1 - p_{jt}) \cdot D_{KL}(\mathbf{E}_t[\Theta_{jt}|A_j^* = 0], \bar{\Theta}_{jt}), \end{aligned} \tag{B5}$$

and thus

$$p_{jt}^2 \cdot D_{KL}(\bar{\Theta}_{jt}^*, \bar{\Theta}_{jt}) \leq p_{jt} \cdot I_t(A_j^*; \tilde{Y}_{jt}) \tag{B6}$$

$$\leq p_{jt} \cdot I_t(A_j^*; Y_{jt}) \tag{B7}$$

$$= I_t(A_j^*; A_{jt} \cdot Y_{jt}, A_{jt}) \tag{B8}$$

$$\leq I_t(A_j^*; Y_t(A_t), A_t). \tag{B9}$$

These steps hold for the following reasons.

(B6) Because the second term in Equation (B5) is non-negative.

(B7) By the data-processing inequality, applied to the mapping from Y_{jt} to \tilde{Y}_{jt} .

(B8) By the law of iterated expectations, applied to $I_t(A_j^*; A_{jt} \cdot Y_{jt}, A_{jt})$, averaging over the distribution of A_{jt} (under Thompson sampling).

(B9) By the data processing inequality, again.

□

Lemma 3 (Bounding the sum of component-wise information).

$$\sum_{t=1}^T \sum_{j=1}^J I_t(A_j^*; Y_t(A_t), A_t) \leq M \cdot \left[\log \left(\frac{J}{M} \right) + 1 \right]$$

Proof of Lemma 3:

$$\sum_{t=1}^T \sum_{j=1}^J I_t(A_j^*; Y_t(A_t), A_t) = \sum_{j=1}^J I_1(A_j^*; (Y_t(A_t), A_t : t = 1, \dots, T)) \quad (\text{B10})$$

$$\leq \sum_{j=1}^J H_1(A_j^*) \quad (\text{B11})$$

$$= - \sum_{j=1}^J [p_{j,1} \log(p_{j,1}) + (1 - p_{j,1}) \log(1 - p_{j,1})] \quad (\text{B12})$$

$$\leq J \cdot \left(\frac{M}{J} \log \left(\frac{J}{M} \right) + \left(\frac{J-M}{J} \right) \log \left(\frac{J}{J-M} \right) \right) \quad (\text{B13})$$

$$\leq M \cdot \left[\log \left(\frac{J}{M} \right) + 1 \right] \quad (\text{B14})$$

These steps hold for the following reasons.

(B10) The chain rule of mutual information.

(B11) The entropy reduction form of mutual information and the non-negativity of (conditional) entropy.

(B12) The definition of entropy for A_j^* .

(B13) Jensen's inequality.

(B14) The inequality $\log(1+x) \leq x$ for $x = m/(d-m)$.

□

Proof of Theorem 1:

$$\mathbf{E}_1 \left[\sum_{t=1}^T (R(A^*) - R(A_t)) \right] = \mathbf{E}_1 \left[\sum_{t=1}^T \mathbf{E}_t [R(A^*) - R(A_t)] \right] \quad (\text{B15})$$

$$\leq \mathbf{E}_1 \left[\sum_{t=1}^T \sqrt{\frac{J}{2} \sum_{j=1}^J I_t(A_j^*; Y_t(A_t), A_t)} \right] \quad (\text{B16})$$

$$\leq \sqrt{\frac{1}{2} J T \mathbf{E}_1 \left[\sum_{t=1}^T \sum_{j=1}^J I_t(A_j^*; Y_t(A_t), A_t) \right]} \quad (\text{B17})$$

$$\leq \sqrt{\frac{1}{2} J T M \cdot [\log \left(\frac{J}{M} \right) + 1]}. \quad (\text{B18})$$

These steps hold for the following reasons.

(B15) The law of iterated expectations.

(B16) Lemma 1.

(B17) Cauchy-Schwarz for the inner product with a T -vector of 1s.

(B18) Lemma 3.

□