Learning by Matching

Maximilian Kasy[†] and Alexander Teytelboym[‡]

[†]*Department of Economics, University of Oxford.*
E-mail: maximilian.kasy@economics.ox.ac.uk
E-mail: alexander.teytelboym@economics.ox.ac.uk

Summary We consider an experimental setting in which a matching of resources to participants has to be chosen repeatedly and returns from the individual chosen matches are unknown but can be learned. Our setting covers two-sided and one-sided matching with (potentially complex) capacity constraints, such as refugee resettlement, social housing allocation, and foster care. We propose a variant of the Thompson sampling algorithm to solve such adaptive combinatorial allocation problems. We give a tight, prior-independent, finite-sample bound on the expected regret for this algorithm. Although the number of allocations grows exponentially in the number of matches, our bound does not. In simulations based on refugee resettlement data using a Bayesian hierarchical model, we find that the algorithm achieves half of the employment gains (relative to the status quo) that could be obtained in an optimal matching based on perfect knowledge of employment probabilities.

Keywords: Matching, experimental design, bandits, Bayesian modeling, optimal policy, refugees.

1. INTRODUCTION

Adaptive experimentation uses information obtained in the course of an experiment in order to optimize the treatment assignment for later study participants. 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). Building on the large literature on multi-armed bandits, adaptive experimentation has been used to maximize the welfare of study participants (Berry, 2006) and to inform subsequent policy choices (Kasy and Sautmann, 2021).

In many policy settings, however, policymakers do not simply choose between a few interventions. Instead, they need to select an entire allocation of resources—which we call a *matching*—among participants. These resources are typically scarce, and feasible matchings can be subject by combinatorial constraints. Moreover, returns from the different matchings are unknown, but can be learned. Our motivating example of such an allocation problem is refugee resettlement, where a resettlement agency needs to match arriving refugee families to hosting communities while trying to maximize the employment outcomes of refugees.

There are many other applications with a similar structure. 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 respects the demographic composition of students in the population. If the policymaker wants to match 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 close to schools and family homes (MacDonald, 2019; Robinson-Cortés, 2019). If the policymaker wants to match 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 make adaptive experimentation more difficult relative to the unconstrained case (which is typically considered in the multi-armed bandit literature), since the number of possible matchings 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 matchings have to be learned). We show that, remarkably, despite these difficulties, learning performance close to the case without combinatorial constraints can be achieved. In this paper we consider an adaptive allocation policy extending the idea of Thompson sampling (Thompson, 1933). 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 characterization implies that this policy is close to optimal for maximizing the outcomes of experimental participants in matching problems with combinatorial constraints.

Setup We consider the following experimental setting. The decision-maker has access to a finite number of *matches* but is constrained to selecting only *matchings* (i.e., combinations of matches) that satisfy the resource constraints (e.g., a one-to-one matching). Participants arrive in *batches* every period. The decision-maker selects a matching and observes the outcome of each selected match. The outcome of each match results in a *reward*. The decision-maker's objective is to maximize the expected cumulative rewards from all the matches she picked over time; equivalently, the decision-maker aims to minimize expected *regret*, i.e., the expected difference relative to the reward for the optimal matching in each period. The decision-maker faces a trade-off between selecting a myopically optimal matching which benefits the current batch ("exploitation") and experimenting by trying another matching which helps the decision-maker learn about the rewards from different matches thereby improving future allocations ("exploration"). Such a setting is sometimes referred to as a *combinatorial semi-bandit* setting with *linear* rewards (Audibert et al., 2014). "Combinatorial" because the decision-maker can choose combinations of matches; "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 when using Thompson sampling in our setting. 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., matchings) 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 and allows for arbitrary prior dependence of the expected outcomes of different matches.

Application We apply our approach to the problem of matching resettled refugees to local communities in the United States (Bansak et al., 2018; Ahani et al., 2021). Our data cover the placement of all refugees by HIAS, an American resettlement agency, between 2011 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 capacity constraints. Local communities have a quota on the total number of refugees they can resettle in a given year and placement decisions are made in batches at regular intervals. In our simulations using a Bayesian model, we can optimize employment for each batch of arriving refugees, given a draw of parameters from the posterior, via linear programming (Bansak et al., 2018). We find that our Thompson sampling algorithm achieves half of the employment gains delivered by an oracle-optimal matching. However, there is substantial

redistribution in employment rates across communities and across refugees with different characteristics.

Literature 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). 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 acquistion (Schwartz et al., 2017). More recently, adaptive experimentation has been deployed in field experiments in development contexts (Caria et al., 2020; Kasy and Sautmann, 2021). 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). Wager and Xu (2021) derive characterizations of Thompson sampling based on local-to-zero asymptotics. The closest setting to ours is discussed by Wang and Chen (2018) who provide a distribution-dependent regret bound for the Thompson algorithm in the combinatorial semi-bandit setting; in contrast, our result is distribution-free. Other work has studied adversarial combinatorial (semi-) bandits (Audibert et al., 2014) where the outcomes are assumed to be chosen by an adversary and algorithm performance is compared to the best constant policy; and looked at algorithm performance for the upper tail of regret (Audibert et al., 2009).

The proof of our main theorem builds on the information-theoretic approach pioneered by Russo and Van Roy (2016) (in particular their Lemmata 1 and 2 and Proposition 6), as well as on the componentwise entropy approach introduced by Bubeck and Sellke (2020). While the core ideas of our proof are present in these papers, our main theorem provides a bound not stated there. 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 the parameter vector for all time periods. By contrast, our main result allows for arbitrary dependence. This dependence 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. 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.

Also related is the analysis of Zimmert and Lattimore (2019) (see also Lattimore and Gyorgy 2021, and Lattimore and Szepesvári 2020, especially chapter 30), who build on the work of Russo and Van Roy (2016) to derive adversarial regret bounds for online mirror descent algorithms. Their approach covers general linear partial monitoring games, which include (adversarial) semi-bandits as a special case. They draw connections between (modified) Thompson sampling and online stochastic mirror descent. Lastly, Perrault et al. (2020) also consider the stochastic semi-bandit framework, as we do. They provide asymptotic regret bounds for fixed parameter values, in the tradition of Agrawal and Goyal (2012). They allow for statistical dependency between outcomes across matches, while requiring prior independence across matches; this contrasts with our substantively motivated focus on allowing prior dependence across matches.

Roadmap 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 as well as methods for sampling from the posterior. Section 5 discusses calibrated simulations based on our refugee resettlement application. Section 6 concludes. Appendix A provides a brief review of information theory,

which is needed for the proof of our main result. All proofs can be found in Appendix B. Appendix C discusses randomization inference. Appendix D contains additional empirical results.

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 matches and actions The decision-maker has access to possible *matches* $j = 1,\dots,J$, but only has sufficient resources to select M J of these. We denote by A $\{a = (a_1,\dots,a_J) : \|a\|_1 = M\}$ a collection of *matchings*, i.e., *feasible* combinations of matches.[1] A is a strict subset if the decision-maker faces additional allocation constraints. The decision-maker's *action* $a \in A$ is a matching.

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 match j in period t.[2] The vectors Y_t are i.i.d. across periods. We denote the average potential outcome (or *average structural function*) for match 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 matches 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 matches j (the matches j for which $a_j = 1$), that is, the vector

$$Y_t(a) = (a_j \quad Y_{jt} : \ j = 1,\dots,J). \tag{2.1}$$

We assume "stable unit treatment values" (i.e., no spillovers or interference) across matches 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 a match 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

$$F_t = \{(A_{t'}, Y_{t'}(A_{t'})) : \ 1 \quad t' < t\}. \tag{2.2}$$

Throughout this paper, the subscript t on $\mathbf{E}_t$ indicates that the expectation is evaluated under the posterior distribution $\mathbf{P}_t(\ ) = \mathbf{P}(\ | F_t)$, where 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 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

$$R(a) = \mathbf{E}_t[\langle a, Y_t \rangle | \theta] = \langle a, \theta \rangle. \tag{2.3}$$

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

$$\mathbf{E}_1\left[\sum_{t=1}^{} R(A_t) \ :\right. \tag{2.4}$$

[1] More generally, matches can be thought of as "options" and matchings as "allocations".

[2] To avoid interfering superscripts, use subscript jt to denote match j in period t throughout the paper.

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,

$$A^* \in \operatorname*{argmax}_{a \in \mathcal{A}} R(a) = \operatorname*{argmax}_{a \in \mathcal{A}} \langle a, \Theta \rangle. \tag{2.5}$$

Therefore, A^* is an oracle-optimal action. The decision-maker's objective is equivalent to minimizing expected *regret* at T

$$\mathbf{E}_1 \left[\sum_{t=1}^{T} (R(A^*) - R(A_t)) \right]. \tag{2.6}$$

Expected regret is the difference between the cumulative expected rewards from the oracle-optimal action (which is based on perfect knowledge of Θ) and the cumulative expected rewards from the actions actually taken by the decision-maker. 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 $a \in \mathcal{A}$,

$$\mathbf{P}_t(A_t = a) = \mathbf{P}_t(A_t^* = a). \tag{2.7}$$

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 3.1 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 matches might correspond to the 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 match, but Θ_j is the same for all matched pairs j with the same observed covariates on both sides of the match.

Example 2.1. (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 matches 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 2.2. (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 matches j that are observationally identical, i.e., across matches of children and families with the same observed covariates.*

Example 2.3. (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 matches 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.*

Example 2.4. (Therapy combinations) *Many diseases, such as cancers, are best treated by a combination of therapies rather than by a single therapy (Mokhtari et al., 2017). The policymaker wishes to maximize a health-related objective, such as survival, but is constrained in the the total amount of each therapy that is available to arriving patients. The number of matches J is therefore the number of distinct matches between different patients and combinations of therapies (some therapy combination might be incompatible). An action a is then a feasible* many-to-many matching *between therapies and patients.*

3. PERFORMANCE GUARANTEE

We now state our main theoretical result which provides a tight worst-case guarantee for the expected regret of Thompson sampling in our setup.

Theorem 3.1. *Under the assumptions of Section 2,*

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

Discussion of Theorem 3.1 Several features of the regret bound in Theorem 3.1 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 $\sqrt{JTM}$, up to logarithmic terms.

Theorem 3.1 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 TM, which yields the bound $\sqrt{\frac{J}{2TM}\left(\log\left(\frac{J}{M}\right)+1\right)}$. This bound goes to 0 at a rate of 1 over the square root of the sample size, that is, at a rate of $1/\sqrt{TM}$. The theorem furthermore shows that this worst case expected regret grows, as a function of the number of possible matches 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 3.1 grows like that for a simple multiarmed bandit with J arms.

Intuition for the proof of Theorem 3.1 The proof of Theorem 3.1 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 $(\mathbf{E}[B] \quad \mathbf{E}[B'])^2 \quad \frac{1}{2}D_{KL}(B;B')$: Lemma B.1 applies Pinsker's inequality to terms showing up in the definition of expected regret which are of the form $\mathbf{E}_t[\Theta_j | A_j^* = 1] \quad \mathbf{E}_t[\Theta_j]$. 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 B.2 relates the KL-distance to the entropy of the events $A_j^* = 1$. The combination of these two lemmata allows us to bound the expected regret for match j in terms of the entropy reduction for the posterior of A_j^*.

Third and lastly, Lemma B.3 shows that the total reduction of entropy across the matches 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 $M \quad \log \frac{J}{M} + 1$. The proof of Theorem 3.1 then combines these three Lemmata.

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 matches j.

Table 1 proposes some variants of models and priors for matching settings, covering our leading motivating examples, including those used in our empirical application. For each of these variants, we assume that the matches j consist of two-sided matches between types u_j and types v_j. For each possible match, 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 match-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. Sampling from the posterior

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 in Appendix C.

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 Chain which converges to an ergodic

Table 1. Models and priors for matching

Continuous outcomes

$$
\begin{aligned}
&Y_{jt} \quad N(\Theta_j;\ ^2)\\
&\Theta_j = \Gamma^u_{u_j} + \Gamma^v_{v_j} + \Gamma^{uv}_{u_j;v_j}\\
&\Gamma^u_{u_j} \quad N(0;\ ^2_u); \quad \Gamma^v_{v_j} \quad N(0;\ ^2_v); \quad \Gamma^{uv}_{u_j;v_j} \quad N(\ ;\ ^2_{uv});
\end{aligned}
$$

Binary outcomes

$$
\begin{aligned}
&Y_{jt} \quad \text{Bernoulli}(\Theta_j)\\
&\Theta_j = \frac{1}{1+\exp\quad \Gamma^u_{u_j} + \Gamma^v_{v_j} + \Gamma^{uv}_{u_j;v_j}}\\
&\Gamma^u_{u_j} \quad N(0;\ ^2_u); \quad \Gamma^v_{v_j} \quad N(0;\ ^2_v); \quad \Gamma^{uv}_{u_j;v_j} \quad N(\ ;\ ^2_{uv});
\end{aligned}
$$

Discrete outcomes with bounded support $f0;:::;\bar{y}g$

$$
\begin{aligned}
&Y_{jt} \quad \text{Beta-Binomial}(\ _j;\ _j;\bar{y})\\
&A_j = m\ \ \Theta_j; \quad B_j = m\ \ (1\quad \Theta_j)\\
&\Theta_j = \frac{1}{1+\exp\quad \Gamma^u_{u_j} + \Gamma^v_{v_j} + \Gamma^{uv}_{u_j;v_j}}\\
&\Gamma^u_{u_j} \quad N(0;\ ^2_u); \quad \Gamma^v_{v_j} \quad N(0;\ ^2_v); \quad \Gamma^{uv}_{u_j;v_j} \quad N(\ ;\ ^2_{uv});
\end{aligned}
$$

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

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 F_t, generated by MCMC, after a sufficiently long warm-up period. Choose

$$A_t = \operatorname*{argmax}_{a\in\mathcal{A}} \hbar a; \hat{\Theta}_t i: \tag{4.8}$$

Then A_t follows the distribution required for Thompson sampling, that is, it satisfies Equation (2.7).

In order to form 1 credible sets for the parameters Θ_j given the history F_t, one can sample a large number of draws $\hat{\Theta}_t$ from the posterior, and form a credible interval based on the =2 and 1 =2 quantiles of $\hat{\Theta}_{jt}$ across these draws.

5. APPLICATION: REFUGEE RESETTLEMENT

The United States has historically been the world's largest destination of resettled refugees.[3] President Joe Biden has pledged to resettled 125,000 refugee in the fiscal year starting from October 2021. There is substantial empirical evidence that the initial match between refugees and local communities dramatically affects the socioeconomic outcomes of refugees (Bansak et al., 2018; Ahani et al., 2021). However, local community capacities are tightly regulated by the U.S. government. As a result, HIAS, one of nine U.S. resettlement agencies, 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 (Ahani et al., 2021). We draw on the data used by *Annie*™ in order to run calibrated simulations for our proposed algorithm subject to simple capacity constraints (in the spirit of Bansak et al. 2018) with the view to informing actual refugee placement by *Annie*™ in the future.

Data Our data covers all refugees resettled by HIAS between January 2011 and December 2019. Refugee families are resettled to local communities where HIAS runs their *affiliates*. For each primary applicant in the arriving refugee 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 whether the primary applicant had any U.S. ties. Applicants with U.S. ties (e.g., U.S. resident friends or family) are automatically resettled to the affiliate where their U.S. ties reside. Applicants with no U.S. ties (NUST) can be resettled to any of the affiliates run by HIAS. Finally, we can observe which affiliate each refugee family was resettled to and whether or not the primary applicant was employed within 90 days of arrival. This is a key metric used by the U.S. Department of State to assess the performance of American resettlement agencies. Based on the available observables, we classify refugees into 8 "types" u.

There are 57 affiliates in our raw data. We drop any affiliate with fewer than 150 resettled cases over the whole period under consideration, leaving us with 17 affiliates and 2441 refugee families without U.S. ties.[4] All affiliates are anonymized. We treat each of the 17 affiliates as a separate "type" v. This means that there are $8 \quad 17 = 136$ parameters (probabilities of finding employment) that we might wish to learn. Affiliates have a limited capacity in hosting refugees. The annual capacity is suggested by the resettlement agencies and approved by the U.S. Department of State, but the capacities can sometimes change throughout the course of the year.

For our simulations, we conservatively set the available annual capacities to be the total number of refugees families without U.S. ties actually resettled to each affiliate in a given year.[5]

As soon as a refugee family is allocated to a resettlement agency, the agency is responsible for allocating the family to an affiliate. Refugee families typically arrive to the U.S. between 3–6 months after being allocated. We therefore set the batch frequency to quarterly. The quarterly quota for each affiliate is therefore equal to the number of NUST arrivals for that affiliate in that quarter.

Model We fit the hierarchical Bayesian model for binary outcomes described in Table 1 to these data, and set Θ_0 to the posterior mean for this model, as described in Section 4.1.

[3] The resettlement process which bene ts only a small fraction of the world's 25 million refugees is highly regulated and well organized. Many people, of course, also seek asylum by arriving irregularly.

[4] In their analysis, Ahani et al. (2021) also pool some a liates because of small numbers of observations.

[5] In practice, the quotas apply to the total number of *refugees* rather than *families* resettled in each a liate (Ahani et al., 2021) and resettlement agencies are allowed to exceed their o cial capacity by 10% without further approval. Moreover, there are feasibility constraints on which refugees can be placed in which a liate (e.g., not all a liates can host single-parent refugee families). Since our application is illustrative, we abstract away from these practical issues (see also Ahani et al. (2021) for a discussion of dynamic quota management).

In the simulations described next, we sample counterfactual outcomes for refugees allocated using the Thompson algorithm based on the estimated parameter values Θ_0.

This model assumes that potential outcome distributions are stationary. Figure 4 in Appendix D shows that employment probabilities of different refugee types are indeed approximately stationary: The observed employment rate in each year closely tracks the estimated employment rate, which is based on the parameters Θ_0 and the actual demographics of arrivals and their allocation to affiliates.

Simulation design Our simulated matching process works as follows. We use calendar year 2011 as "burn-in" period for the Thompson algorithm and start rematching in January 2012.

For each quarter t in the available data, we consider all the refugees who were resettled by HIAS in that quarter. For example, we match all refugees who arrived between 1 October and 1 January to their affiliates on 1 October because of the lags between matching and arrival. Because employment is measured after 90 days, when we match refugees in period t, we only have the employment information for refugee families who arrived up to and including quarter $t \quad 2$.[6]

We match refugees with U.S. ties to their actual affiliates. For all the refugees without U.S. ties, we match them to affiliates using the Thompson algorithm. This matching has to satisfy the capacity constraints of affiliates described above. We can solve for the optimal matching for a given draw $\hat{\Theta}$ from the posterior using linear programming (Bansak et al., 2018). Since we set total quarterly quota of the affiliates is equal to the number of refugees arriving that quarter and each refugee family has a weakly positive employment probability in every affiliate, the optimal matching assigns every family to some affiliate in its arrival quarter.

Results Figure 1 summarizes the key takeaway message from the simulations: there are substantial gains from adaptive matching in refugee resettlement. The Thompson algorithm is able to obtain around half of the gains from oracle-optimal matchings, i.e., the optimal matchings obtained with the full knowledge of Θ_0 for each refugee type. Figure 1 shows that the oracle-optimal matching boosts employment by around 5 percentage points compared to the actual assignment, from around 40 percent to around 45 percent. However, adaptive matching alone can lift employment rates by 2-3 percentage points compared to the actual assignment. Learning happens quickly and the gains can be seen starting in the first year.

Figure 2 shows the simulated trajectories across the eight refugee types and reveals substantial redistribution of employment across types compared to the actual assignment. Working-age men, who constitute the most common household-head types, experience substantial gains from adaptive matching (and from oracle-optimal matching) while households with non-working-age, non-English-speaking women as primary applicants lose out. This illustrates that maximization of overall employment rates might not lead to an increase in employment rates for each subgroup.

Figure 3 in Appendix C shows that using the Thompson algorithm increases employment rates in 13 out of 17 affiliates (in some cases substantially). These gains align with those obtained by oracle-optimal matching.

[6] For example on 1 October, we observe all employment of all the refugees who arrived up to the April-June quarter but we do not yet observe the employment of refugees who arrived in the July-September quarter.

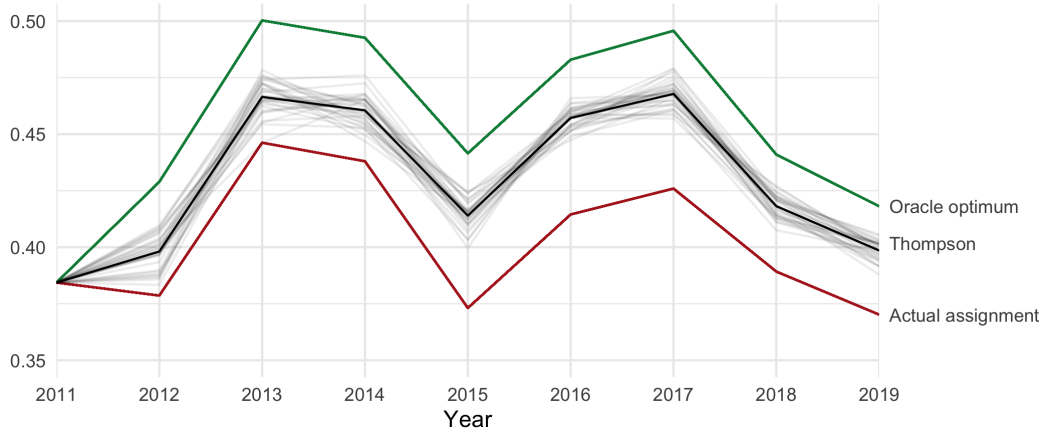

Figure 1. Simulated expected employment rates by year.

Notes: Simulations based on refugee resettlement data described in Section 5. Grey lines: 32 simulation runs of the Thompson algorithm. Black line: average of the 32 simulation runs of the Thompson algorithm. Red line: Expected employment based on the actual assignment of refugees to locations. Green line: Expected employment for the optimal assignment given knowledge of $_0$, subject to actual capacity constraints.

6. CONCLUSION

In many policy choice problems the policymaker is required to match many resources to many participants in each period. Since the number of possible matchings available to the policymaker can be vast, it is not clear whether exploration can take place quickly enough to improve welfare. We derive a tight, finite-sample, prior-independent regret bound for the Thompson algorithm in such a combinatorial semi-bandit setting that does not depend on the batch size. We test how our algorithm could increase employment rates of refugees resettled in the U.S. Our simulations suggest that the Thompson algorithm would be able to achieve substantial and persistent employment gains for refugees of different characteristics.

In our setting, which allows for arbitrary statistical dependence of the prior across matches, Thompson sampling achieves the efficient rate of convergence for regret. Of course, in many settings there might be additional structure on the parameters that would allow one to derive tighter bounds. For example, in the refugee resettlement context there might be refugees who are observationally identical in a given batch therefore their parameters would be perfectly correlated. We leave further improvements of our bound in specific settings for future research.

ACKNOWLEDGEMENTS

We thank Daniel Privitera and Manos Perdikakis for excellent research assistance. Teytelboym was supported by Economic and Social Research Council Grant ES/R007470/1.

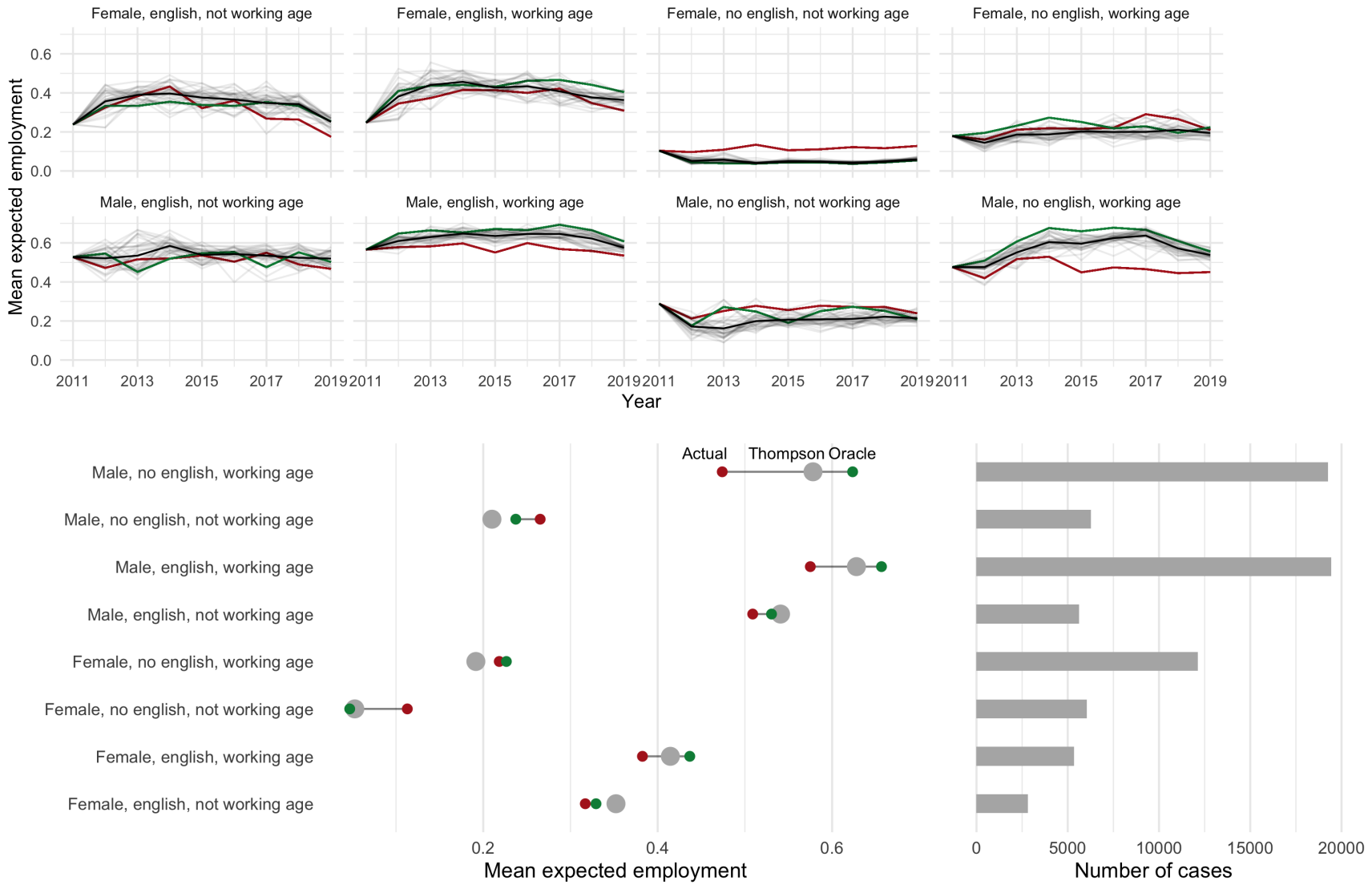

Figure 2. Simulated expected employment rates by year and by type.

Notes: Simulations based on refugee resettlement data described in Section 5. The plot on top shows annual averages across types. The bottom left shows average employment by type for the whole period. The bottom right shows the distribution across types. Grey lines: 32 simulation runs of the Thompson algorithm. Black lines / grey dots: average of the 32 simulation runs of the Thompson algorithm. Red: Expected employment based on the actual assignment of refugees to locations. Green: Expected employment for the optimal assignment given knowledge of $_0$, subject to actual capacity constraints.

REFERENCES

Agrawal, S. and N. Goyal (2012). Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, Volume 23, pp. 39.1–39.26.

Ahani, N., T. Andersson, A. Martinello, A. Teytelboym, and A. C. Trapp (2021, forthcoming). Placement optimization in refugee resettlement. *Operations Research*.

Ahani, N., P. Ǵolz, A. Procaccia, A. Teytelboym, and A. C. Trapp (2021). Dynamic placement in refugee resettlement. In *Economics and Computation, EC'21*.

Audibert, J.-Y., S. Bubeck, and G. Lugosi (2014). Regret in online combinatorial optimization. *Mathematics of Operations Research 39*(1), 31–45.

Audibert, J.-Y., R. Munos, and C. Szepesvári (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science 410*(19), 1876–1902.

Bansak, K., J. Ferwerda, J. Hainmueller, A. Dillon, D. Hangartner, D. Lawrence, and J. Weinstein (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 M. Sellke (2020). First-order Bayesian regret analysis of Thompson sampling. In *Algorithmic Learning Theory*, pp. 196–233.

Caria, S., G. Gordon, M. Kasy, S. Osman, S. Quinn, and A. Teytelboym (2020). An adaptive targeted field experiment: Job search assistance for refugees in Jordan. Working Paper 8535, CESifo.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of Statistical Software 76*(1).

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., G. W. Imbens, and G. Ridder (2010). Measuring the effects of segregation in the presence of social spillovers: a nonparametric approach. Working Paper 16499, National Bureau of Economic Research.

Kasy, M. and A. Sautmann (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica 89*(1), 113–132.

Kaufmann, E., N. Korda, and R. Munos (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pp. 199–213.

Kawale, J., H. H. Bui, B. Kveton, L. Tran-Thanh, and S. Chawla (2015). Efficient Thompson sampling for online matrix-factorization recommendation. In *Advances in Neural Information Processing Systems*, pp. 1297–1305.

Lai, T. L. and H. Robbins (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics 6*(1), 4–22.

Lattimore, T. and A. Gyorgy (2021). Mirror descent and the information ratio. In *Conference on Learning Theory*, pp. 2965–2992. PMLR.

Lattimore, T. and C. Szepesvári (2020). *Bandit algorithms*. Cambridge University Press.

MacDonald, D. E. (2019). Foster care: A dynamic matching approach.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Mokhtari, R. B., T. S. Homayouni, N. Baluch, E. Morgatskaya, S. Kumar, B. Das, and H. Yeger (2017). Combination therapy in combating cancer. *Oncotarget 8*(23), 38022.

Perrault, P., E. Boursier, V. Perchet, and M. Valko (2020). Statistical efficiency of thompson sampling for combinatorial semi-bandits. *arXiv preprint arXiv:2006.06613*.

Robinson-Cortés, A. (2019). Who gets placed where and why? An empirical framework for foster care placement.

Russo, D. and B. Van Roy (2016). An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research 17*(1), 2442–2471.

Schwartz, E. M., E. T. Bradlow, and P. S. Fader (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.

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.

van Dijk, W. (2019). The socio-economic consequences of housing assistance.

Wager, S. and K. Xu (2021). Diffusion asymptotics for sequential experiments. *arXiv preprint arXiv:2101.09855*.

Waldinger, D. (2018). Targeting in-kind transfers through market design: A revealed preference analysis of public housing allocation.

Wang, S. and W. Chen (2018). Thompson sampling for combinatorial semi-bandits. Preprint 1803.04623, arXiv.

Zimmert, J. and T. Lattimore (2019). Connections between mirror descent, thompson sampling and the information ratio. *arXiv preprint arXiv:1905.11817*.

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:

$$H(A) = H(p) = \quad [p\log(p) + (1 \quad p)\log(1 \quad p)]: \tag{A1}$$

KL divergence:

$$D_{KL}(A; A') = D_{KL}(p; q) = p\log \ \tfrac{p}{q} \ + (1 \quad p)\log \ \tfrac{1-p}{1-q} \ : \tag{A2}$$

Pinsker's inequality:

$$(\mathbf{E}[A] \quad \mathbf{E}[A'])^2 = jp \quad qj \quad \tfrac{1}{2}D_{KL}(p; q) = \tfrac{1}{2}D_{KL}(A; A'): \tag{A3}$$

Mutual information as expected divergence of the posterior:
For any random variable or vector F, let $p(f) = \mathbf{E}[AjF = f]$. Then

$$I(A; F) = \mathbf{E}[D_{KL}(p(F); p)]: \tag{A4}$$

Conditional entropy:

$$H(AjF) = E\left[H(p(F))\right]: \tag{A5}$$

Entropy reduction form of mutual information:

$$I(A; F) = H(A) \quad H(AjF): \tag{A6}$$

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

$$I(A; g(F)) \quad I(A; F): \tag{A7}$$

Chain rule of mutual information:

$$I(A; (F; G)) = I(A; F) + I(A; GjF): \tag{A8}$$

B. PROOFS

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

$Y_t; \Theta; A_t \ 2 \ \mathbb{R}^J$	Outcome, parameter, and action vectors
$A_t \ 2 \ A \quad fa \ 2 \ f0; 1g^J \ : \ kak_1 = Mg$	Feasible allocations and batch size
$Y_{jt} \ 2 \ [0; 1]$	Bounded outcomes
$\Theta = \mathbf{E}_t[Y_t j\Theta]$	Parameters are expectation of outcomes
$\bar{\Theta}_t = \mathbf{E}_t[\Theta] = \mathbf{E}_t[Y_t]$	Prior expectation of the parameter (at t)
$R(a) = \mathbf{E}_t[ha; Y_t ij\Theta] = ha; \Theta i$	Linear (combinatorial) expected rewards
$Y_t(a) = (a_j \quad Y_{jt} : \ j = 1; : : : ; J)$	Observable outcomes (semi bandit)
$A^* \ 2 \ \operatorname{argmax}_{a \in \mathcal{A}} \ R(a) = \operatorname{argmax}_{a \in \mathcal{A}} ha; \Theta i$	Optimal action
$\bar{\Theta}^*_{jt} = \mathbf{E}_t[\Theta_j jA^*_j = 1] = \mathbf{E}_t[Y_{jt} jA^*_j = 1]$	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 Lemmata, before combining them in the proof of Theorem 3.1 itself.

LEMMA B.1. (BOUNDING REGRET BY THE COMPONENT-WISE INFORMATION)

$$\mathbf{E}_t[R(A^*) \quad R(A_t)] \quad \sqrt{\frac{J}{2}\sum_{j=1} p_{jt}^2 \quad D_{KL} \quad \bar{\Theta}^*_{jt}; \bar{\Theta}_{jt}} \ :$$

Proof of Lemma B.1:

$$\begin{aligned}
\mathbf{E}_t[R(A^*) \quad R(A_t)] &= \mathbf{E}_t[hA^* \quad A_t; \Theta i] && \text{(B1)}\\
&= hp_t; \bar{\Theta}^* i \quad hp_t, \bar{\Theta}_t i && \text{(B2)}\\
&\sqrt{J\sum_{j=1} p_{jt}^2 \quad \bar{\Theta}^*_{jt} \quad \bar{\Theta}_{jt}}^{\,2} && \text{(B3)}\\
&\sqrt{\frac{J}{2}\sum_{j=1} p_{jt}^2 \quad D_{KL} \quad \bar{\Theta}^*_{jt}; \bar{\Theta}_{jt}} && \text{(B4)}
\end{aligned}$$

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 $\bar{\Theta}^*_{jt}; \bar{\Theta}_{jt}$.
2

LEMMA B.2. (DIVERGENCE AND COMPONENT-WISE INFORMATION GAIN)

$$p_{jt}^2 \quad D_{KL} \quad \bar{\Theta}^*_{jt}; \bar{\Theta}_{jt} \qquad I_t(A_j^*; Y_t(A_t); A_t):$$

Proof of Lemma B.2:
For the purpose of this proof, construct a Bernoulli random variable $\mathbb{Y}_{jt}$ with expectation Y_{jt}, independently of everything else. Note that $\mathbf{E}_t[\mathbb{Y}_{jt}] = \bar{\Theta}_{jt}$. $D_{KL} \quad \bar{\Theta}^*_{jt}; \bar{\Theta}_{jt}$ can be interpreted as the KL-divergence between the distribution of $\mathbb{Y}_{jt}$ conditional on $A_j^* = 1$ and the (unconditional) distribution of $\mathbb{Y}_{jt}$. Taking the expectation over A_j^* of the KL-divergence yields the mutual information between A_j^* and $\mathbb{Y}_{jt}$, $I_t(A_j^*; \mathbb{Y}_{jt})$:

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

and thus

$$\begin{aligned}
p_{jt}^2 \quad D_{KL} \quad \bar{\Theta}^*_{jt}; \bar{\Theta}_{jt} \qquad & p_{jt} \quad I_t(A_j^*; \mathbb{Y}_{jt}) && \text{(B6)}\\
& p_{jt} \quad I_t(A_j^*; Y_{jt}) && \text{(B7)}\\
& = I_t(A_j^*; A_{jt} \quad Y_{jt}; A_{jt}) && \text{(B8)}\\
& \quad I_t(A_j^*; Y_t(A_t); A_t): && \text{(B9)}
\end{aligned}$$

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 $\hat{Y}_{jt}$.

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

(B9) By the data processing inequality, again. ∎

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

$$\sum_{t=1}^{T} \sum_{j=1}^{J} I_t(A_j^*; Y_t(A_t); A_t) \leq M \log \tfrac{J}{M} + 1$$

Proof of Lemma B.3:

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

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

$$= \sum_{j=1}^{J} [p_{j;1} \log(p_{j;1}) + (1 - p_{j;1}) \log(1 - p_{j;1})] \tag{B12}$$

$$J \left[\tfrac{M}{J} \log \tfrac{J}{M} + \tfrac{J-M}{J} \log \tfrac{J}{J-M} \right] \tag{B13}$$

$$M \log \tfrac{J}{M} + 1 \tag{B14}$$

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 3.1:

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

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

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

$$\sqrt{\frac{1}{2} JTM \log \tfrac{J}{M} + 1} \,. \tag{B18}$$

These steps hold for the following reasons.

(B15) The law of iterated expectations.
(B16) Lemma B.1.
(B17) Cauchy-Schwarz for the inner product with a T-vector of 1s.
(B18) Lemma B.3. *2*

C. RANDOMIZATION INFERENCE

An alternative to Bayesian inference discussed in Section 4.2 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 $a \in A$ 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 many-to-one matching (of refugees to local communities, say), where each match 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 $\tilde{F}_t$ by re-running the Thompson sampling algorithm iteratively. In each period s, draw $\tilde{\Theta}_{t^s}$ and the corresponding $\tilde{A}_{t^s}$ from the posterior given $\tilde{F}_{t^s}$. Impute a counter-factual outcome vector $Y_{t^s}^0(\tilde{A}_{t^s})$, based on the null hypothesis to be tested. Update the history F_{t^s} by adding $\tilde{A}_{t^s}; Y_{t^s}^0(\tilde{A}_{t^s})$, and iterate for the next period. Once $t' = t$, calculate a realization of the test-statistic as a function of $\tilde{F}_t$. 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.

D. ADDITIONAL EMPIRICAL RESULTS