

## Matching with semi-bandits

MAXIMILIAN KASY AND ALEXANDER TEYTELBOYM

*Department of Economics, University of Oxford, UK.*

Email: [maximilian.kasy@economics.ox.ac.uk](mailto:maximilian.kasy@economics.ox.ac.uk)

Email: [alexander.teytelboym@economics.ox.ac.uk](mailto:alexander.teytelboym@economics.ox.ac.uk)

First version received: 8 September 2021; final version accepted: 8 April 2022.

**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 modelling, optimal policy, refugees.*

**JEL codes:** *C44, C78, J6.*

### 1. INTRODUCTION

Adaptive experimentation uses information obtained in the course of an experiment in order to optimise the treatment assignment for later study participants. For example, if job seekers arrive at a job centre over time, a policymaker can use the outcomes of earlier job seekers in order to improve the assignment of labour market interventions for later participants (Caria et al., 2020). Building on the large literature on multi-armed bandits, adaptive experimentation has been used to maximise 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 maximise 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), he/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, he/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, he/she needs to ensure that housing matches the needs of tenants and respects waiting-list priorities (Thakral, 2016; Waldinger, 2021; van Dijk, 2019). If the policymaker wants to allocate combinations of therapies to different patients in order to overcome a disease, he/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 optimisation 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 characterisation implies that this policy is close to optimal for maximising 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 maximise the expected cumulative rewards from all the matches he/she picked over time; equivalently, the decision-maker aims to minimise 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.<sup>1</sup>

Our main theoretical result is a prior-independent bound on the Bayesian regret obtained when using Thompson sampling in our setting. Our theoretical result is appealing for three reasons. First, the bound on Bayesian 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

<sup>1</sup> Combinatorial semi-bandits with linear rewards differ from two other important settings: linear bandit models have the same structure of rewards, but less observability; only the total reward of the matching is observed, rather than the rewards of individual matches. Correspondingly, linear bandits are 'harder'. Bandit models for individual matches, on the other hand, share the same structure of rewards and observability, but a larger action space; matches can be chosen independently, with no cross-constraints on the full matching. Correspondingly, such bandits are 'easier'.

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 maximise 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 optimise 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. We also discuss simulations comparing Thompson sampling using a prior with independence across components to Thompson sampling using our preferred hierarchical prior. We find that incorporating proper structure into the prior greatly improves performance, especially in initial periods.

**Literature** Our paper is closely related to the literature on multi-armed bandit problems. Rather than attempting to characterise analytical solutions (e.g., Gittins, 1979), we focus on analysing 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 acquisition (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 characterisations 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. Wang and Chen (2018) focus on the case of priors that are independent across arms. They also seek to generalise their result to the case of dependent arms (which is the focus of the present paper) and provide some generalisations to the case of matroid bandits. By contrast, our result allows for arbitrary dependence. 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).

An alternative to Thompson sampling is the upper confidence bound (UCB) algorithm, which has been proposed in the context of combinatorial semi-bandits by Gai et al. (2012) and analysed further by Chen et al. (2013). Chen et al. (2013) allow for nonlinear aggregation of options and approximate solvers; our setup is restricted to linear aggregation. They prove a distribution-dependent rate for regret as a function of the number of periods. Building on this work, Kveton et al. (2015) provide a more exhaustive characterisation of the UCB algorithm for linear combinatorial semi-bandits. They provide both distribution-dependent and worst-case regret bounds

for UCB, as well as lower bounds for any algorithm.<sup>2</sup> Their worst-case regret bound for UCB, in particular, is of the same order as our Bayesian regret bound for Thompson sampling (square root of the total number of matches, up to logarithmic terms). Note, however, that worst-case and Bayesian regret bounds are different objects; their relationship is discussed in detail at the end of Section 3 below.

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.

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 component-wise 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.

We also draw on the literature on matching in economics. Matching allocations can be evaluated in terms of participant preferences, or in terms of observable outcomes; our setting is of the latter form. There are many examples in the literature where a designer needs to pick an allocation to maximise observed outcomes, including matching of kidney disease patients to donors (where the objective might be to maximise the number of transplants and the designer needs to estimate the likelihood of a rejected transplant, see, e.g., Biró et al., 2021), housing allocation (where the objective might be to maximise the total number of people provided with social housing and the designer needs to estimate the likelihood of family accepting a house, see, e.g., Arnosti and Shi, 2020), and car-sharing (where the objective might be to maximise the total number of completed trips or the revenue for a platform, and the designer needs to estimate supply and demand in each location, see, e.g., Azevedo and Weyl, 2016). For some matching settings, such as kidney exchange and social housing, there are many theoretical papers using preference elicitation, but objective-maximisation approaches are used in practice. Refugee resettlement matching systems are another such example that we discuss in this paper. Variations of the adaptive matching setting have recently been considered by Johari et al. (2021) (with shadow prices rather than combinatorial constraints) and by Jagadeesan et al. (2021) (for stable matchings, rather than observed outcomes).

<sup>2</sup> Distribution-dependent bounds are a function of the parameter vector; worst-case bounds hold for any parameter vector.

**Roadmap** The rest of the paper is organised 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 randomisation inference. Appendix D contains additional empirical results.

## 2. SETUP

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

**Feasible matches and actions** The decision-maker has access to possible *matches*  $j \in 1, \dots, J$ , 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 *matchings*, i.e., *feasible* combinations of matches.<sup>3</sup>  $\mathcal{A}$  is a strict subset if the decision-maker faces additional allocation constraints. The decision-maker's *action*  $a \in \mathcal{A}$  is a matching.

**Timing, potential outcomes, and observability** The experiment takes place 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  $t$ . The vectors  $Y_t$  are i.i.d. across periods. We denote the average potential outcome (or *average structural function*) for match  $j$  by  $\Theta_j$ , 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  $a$ , he/she observes the outcomes of the chosen matches  $j$  (the matches  $j$  for which  $a_j = 1$ ), that is, the vector

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

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

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

Throughout this paper, the subscript  $t$  on  $\mathbf{E}_t$  indicates that the expectation is evaluated under the posterior distribution  $\mathbf{P}_t(\cdot) = \mathbf{P}(\cdot | \mathcal{F}_t)$ , 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 randomisation device that is statistically independent across periods and independent of the sequence of potential outcomes  $(Y_t)_{t=1}^T$ .

<sup>3</sup> More generally, matches can be thought of as ‘options’ and matchings as ‘allocations’.

**Objective and policy** If the decision-maker chooses action  $a$  in period  $t$ , he/she receives a reward which is equal to  $\langle a, Y_t \rangle$ . Therefore, upon taking action  $a$  the decision-maker's *expected reward* given  $\Theta$ , which is the same across periods  $t$ , equals

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

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

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

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  $\Theta$ . Denote by  $A^*$  a feasible action that maximises the expected reward conditional on  $\Theta$  (but not conditional on the vector  $Y_t$ ), that is,

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

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

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

Expected regret is the difference between the cumulative expected rewards from the oracle-optimal action (which is based on perfect knowledge of  $\Theta$ ) and the cumulative expected rewards from the actions actually taken by the decision-maker. Solving this dynamic stochastic combinatorial optimisation problem is computationally quite costly. Rather than solving it, we propose that the decision-maker adopts 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  $a \in \mathcal{A}$ ,

$$\mathbf{P}_t(A_t = a) = \mathbf{P}_t(A_t^* = a). \quad (2.1)$$

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  $\Theta_j$  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.*

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  $\Theta_j$  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 policy-maker would like to choose the gender composition of classrooms in order to maximise 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 maximise a health-related objective, such as survival, but is constrained in 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 prior-independent guarantee for the Bayesian regret of Thompson sampling in our setup.

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

$$\mathbf{E}_0 \left[ \sum_{t=1}^T (R(A^*) - R(A_t)) \right] \leq \sqrt{\frac{1}{2} J T M \cdot \left[ \log \left( \frac{J}{M} \right) + 1 \right]}.$$

**Discussion of Theorem 3.1** Several features of the regret bound in Theorem 3.1 are worth emphasising. 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  $\Theta$  in any way. Furthermore, it allows for prior distributions with arbitrary statistical dependence across the components of  $\Theta$ , 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  $\sqrt{J \cdot \left[\log\left(\frac{J}{M}\right) + 1\right]} / (2TM)$ . 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  $\sqrt{\binom{J}{M} \log\left(\frac{J}{M}\right)}$ . Instead, the regret bound in Theorem 3.1 grows like that for a simple multi-armed 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  $(\mathbf{E}[B] - \mathbf{E}[B'])^2 \leq \frac{1}{2} D_{KL}(B, B')$ . Lemma B.1 applies Pinsker’s inequality to terms showing up in the definition of expected regret which are of the form  $\mathbf{E}_t[\Theta_j | A_j^* = 1] - \mathbf{E}_t[\Theta_j]$ . 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  $M \cdot \left[\log\left(\frac{J}{M}\right) + 1\right]$ . The proof of Theorem 3.1 then combines these three lemmata.

**Bayesian and worst-case regret bounds** An interesting implication of the bound on Bayesian regret provided by Theorem 3.1 is a corresponding bound on worst-case stochastic regret—a frequentist object which does not depend on the chosen prior.<sup>4</sup> To see this, let  $\mathcal{R}(\mathbf{P}_0, \pi) = \mathbf{E}_0 \left[ \sum_{t=1}^T (R(A^*) - R(A_t)) \right]$  denote expected regret for an arbitrary distribution  $\mathbf{P}_0$  over  $\Theta$ , and an arbitrary policy  $\pi$  for choosing matchings that is adapted to the filtration  $\mathcal{F}_t$ . Worst-case regret for any given policy  $\pi$  is equal to  $\sup_{\Theta} \mathcal{R}(\delta_{\Theta}, \pi)$ , and minimax regret is equal to  $\inf_{\pi} \sup_{\Theta} \mathcal{R}(\delta_{\Theta}, \pi)$ .

Sion’s minimax theorem (a generalisation of von Neumann’s minimax theorem) implies that

$$\sup_{\mathbf{P}_0} \inf_{\pi} \mathcal{R}(\mathbf{P}_0, \pi) = \inf_{\pi} \sup_{\mathbf{P}_0} \mathcal{R}(\mathbf{P}_0, \pi).$$

<sup>4</sup> The following discussion draws on the blog-post <https://banditalgs.com/2019/03/17/bayesianminimax-duality-for-adversarial-bandits/>, accessed 8 December 2021.



**Table 1.** Models and priors for matching.

**Continuous outcomes**

$$Y_{jt} \sim N(\Theta_j, \sigma^2)$$

$$\Theta_j = \Gamma_{u_j}^u + \Gamma_{v_j}^v + \Gamma_{u_j, v_j}^{uv}$$

$$\Gamma_{u_j}^u \sim N(0, \tau_{\Gamma^u}^2), \quad \Gamma_{v_j}^v \sim N(0, \tau_{\Gamma^v}^2), \quad \Gamma_{u_j, v_j}^{uv} \sim N(\mu, \tau_{\Gamma^{uv}}^2),$$

**Binary outcomes**

$$Y_{jt} \sim \text{Bernoulli}(\Theta_j)$$

$$\Theta_j = \frac{1}{1 + \exp(-(\Gamma_{u_j}^u + \Gamma_{v_j}^v + \Gamma_{u_j, v_j}^{uv}))}$$

$$\Gamma_{u_j}^u \sim N(0, \tau_{\Gamma^u}^2), \quad \Gamma_{v_j}^v \sim N(0, \tau_{\Gamma^v}^2), \quad \Gamma_{u_j, v_j}^{uv} \sim N(\mu, \tau_{\Gamma^{uv}}^2),$$

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

$$Y_{jt} \sim \text{Beta-Binomial}(\alpha_j, \beta_j, \bar{y})$$

$$A_j = m \cdot \Theta_j, \quad B_j = m \cdot (1 - \Theta_j)$$

$$\Theta_j = \frac{1}{1 + \exp(-(\Gamma_{u_j}^u + \Gamma_{v_j}^v + \Gamma_{u_j, v_j}^{uv}))}$$

$$\Gamma_{u_j}^u \sim N(0, \tau_{\Gamma^u}^2), \quad \Gamma_{v_j}^v \sim N(0, \tau_{\Gamma^v}^2), \quad \Gamma_{u_j, v_j}^{uv} \sim N(\mu, \tau_{\Gamma^{uv}}^2)$$

*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  $\mu$  for continuous outcomes, by  $\tau_{\Gamma^u}^2$ ,  $\tau_{\Gamma^v}^2$ ,  $\tau_{\Gamma^{uv}}^2$ , and  $\mu$  for binary outcomes, and by  $\tau_{\Gamma^u}^2$ ,  $\tau_{\Gamma^v}^2$ ,  $\tau_{\Gamma^{uv}}^2$ , and  $\mu$  for discrete outcomes with bounded support. We propose to use some diffuse prior for these hyper-parameters.

Our prior-independent bound on the regret of Thompson sampling immediately provides an upper bound on  $\sup_{\mathbf{P}_0} \inf_{\pi} \mathcal{R}(\mathbf{P}_0, \pi)$ . Furthermore

$$\inf_{\pi} \sup_{\mathbf{P}_0} \mathcal{R}(\mathbf{P}_0, \pi) \geq \inf_{\pi} \sup_{\Theta} \mathcal{R}(\delta_{\Theta}, \pi),$$

that is, restricting attention to priors which concentrate at a point immediately yields worst-case stochastic regret. We thus get an upper bound on minimax regret of  $\sqrt{\frac{1}{2} J T M \cdot [\log(\frac{J}{M}) + 1]}$ .

## 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  $\Theta_j$ . We need to specify this distribution of  $Y_{jt}$ , as well as a joint prior distribution of the parameters  $\Theta_j$  across  $j$ .

Each of these models assumes that the match-effect  $\Theta_j$  is determined by the sum of type-effects  $\Gamma_{u_j}^u$  and  $\Gamma_{v_j}^v$ , plus an interaction effect  $\Gamma_{u_j, v_j}^{uv}$ . For continuous outcomes, we assume that  $\Theta_j$  is directly given by this sum. For binary or discrete outcomes, we assume that  $\Theta_j$  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  $\Theta_j$  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  $\Theta$ . 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 maximising 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  $\Theta$  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 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

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

Then  $A_t$  follows the distribution required for Thompson sampling; that is, it satisfies equation (2.1).

In order to form  $1 - \alpha$  credible sets for the parameters  $\Theta_j$  given the history  $\mathcal{F}_t$ , one can 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  $\hat{\Theta}_{jt}$  across these draws.

## 5. APPLICATION: REFUGEE RESETTLEMENT

The United States has historically been the world's largest destination of resettled refugees.<sup>5</sup> President Joe Biden pledged to resettle 125,000 refugees in the fiscal year starting from October

<sup>5</sup> The resettlement process, which benefits only a small fraction of the world's 25 million refugees, is highly regulated and well organised. Many people, of course, also seek asylum by arriving irregularly.

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 US government. As a result, HIAS, one of nine US resettlement agencies, optimises the placement of the resettled refugees using its recommendation system called *Annie*<sup>TM</sup><sub>MOORE</sub>. However, *Annie*<sup>TM</sup>'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*<sup>TM</sup> 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*<sup>TM</sup> 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 US ties. Applicants with US ties (e.g., US resident friends or family) are automatically resettled to the affiliate where their US ties reside. Applicants with no US 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 US 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 US ties.<sup>6</sup> All affiliates are anonymised. We treat 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. Affiliates have a limited capacity in hosting refugees. The annual capacity is suggested by the resettlement agencies and approved by the US 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 US ties actually resettled to each affiliate in a given year.<sup>7</sup>

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 US 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.

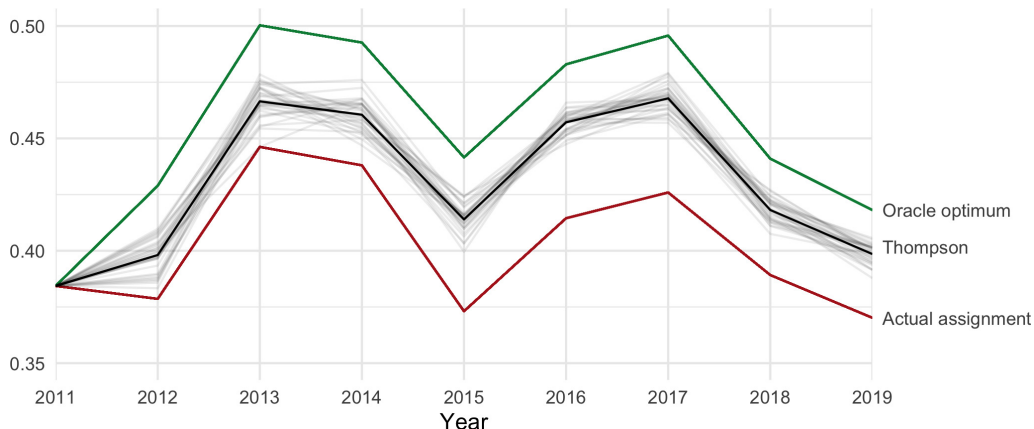
**Model** We fit the hierarchical Bayesian model for binary outcomes described in Table 1 to these data, and set  $\Theta_0$  to the posterior mean for this model, as described in Section 4.1. In the simulations described next, we sample counterfactual outcomes for refugees allocated using the Thompson algorithm based on the estimated parameter values  $\Theta_0$ .

**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

<sup>6</sup> In their analysis, Ahani et al. (2021) also pool some affiliates because of small numbers of observations.

<sup>7</sup> In practice, the quotas apply to the total number of *refugees* rather than *families* resettled in each affiliate (Ahani et al., 2021) and resettlement agencies are allowed to exceed their official capacity by 10 per cent without further approval. Moreover, there are feasibility constraints on which refugees can be placed in which affiliate (e.g., not all affiliates 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).



**Figure 1.** Simulated expected employment rates by year.

*Notes:* Simulations based on refugee resettlement data described in Section 5. Employment rates for refugees with no US ties only. 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  $\Theta_0$ , subject to actual capacity constraints.

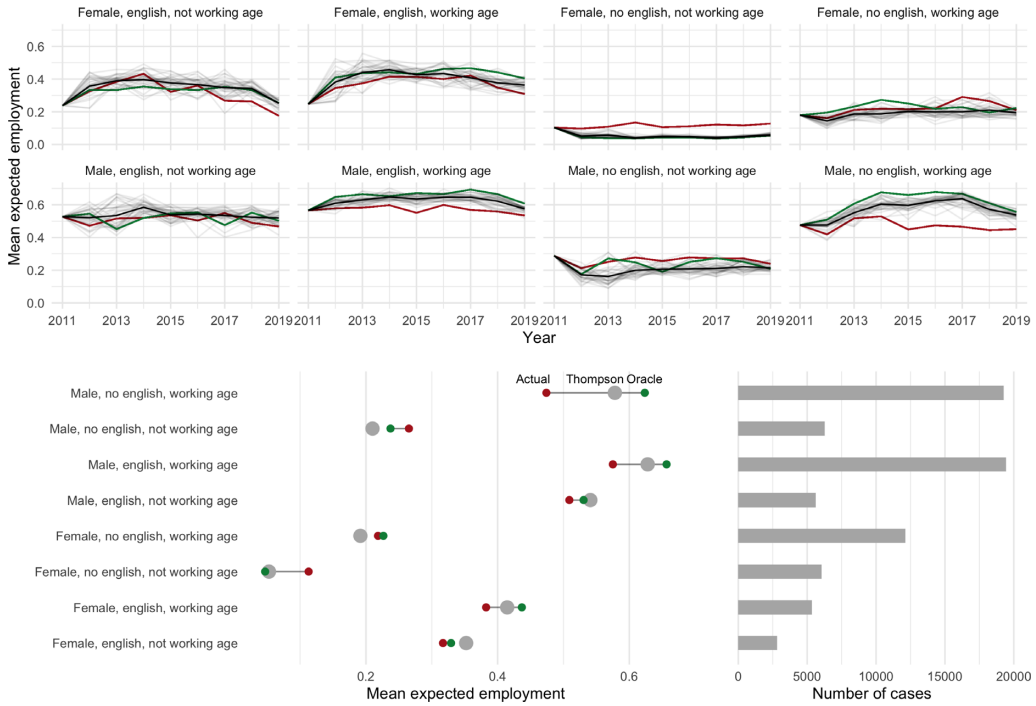
January to their affiliates on 1 October because of the lags between matching and arrival. Because employment is measured after 90 days, when we match refugees in period  $t$ , we only have the employment information for refugee families who arrived up to and including quarter  $t - 2$ .<sup>8</sup>

We match refugees with US ties to their actual affiliates. For all the refugees without US ties, we match them to affiliates using the Thompson algorithm. This matching has to satisfy the capacity constraints of affiliates described above. We solve for the optimal matching for a given draw  $\hat{\Theta}$  from the posterior using linear programming (Bansak et al., 2018). Since we set the total quarterly quota of the affiliates equal to the number of refugees arriving that quarter, and since 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 summarises 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  $\Theta_0$  for each refugee type. Figure 1 shows that the oracle-optimal matching boosts employment by around five percentage points compared to the actual assignment, from around 40 per cent to around 45 per cent. 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 nonworking-age,

<sup>8</sup> 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 2.** Simulated expected employment rates by year and by type.

*Notes:* Simulations based on refugee resettlement data described in Section 5. Employment rates for refugees with no US ties only. 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  $\Theta_0$ , subject to actual capacity constraints.

nonEnglish-speaking women or men as primary applicants lose out. This illustrates that maximisation of overall employment rates might not lead to an increase in employment rates for each subgroup.

Figure D1 in Appendix D 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.

Our model assumes that potential outcome distributions are stationary. Figure D.2 in Appendix D shows that the actual employment rates of different refugee types are not stationary. This is not surprising given that refugee flows can be highly volatile and, moreover, from 2017 HIAS has been using *Annie*<sup>TM</sup>MOORE in their assignment process. Further research might study a nonstationary version of our model.

**The importance of prior dependence** Table D1 in Appendix D, finally, demonstrates the importance of using a hierarchical prior, rather than an independent prior as considered in Russo and Van Roy (2016). Table D1 repeats our main simulations, but uses a prior where the components

$\Theta_j$  follow an independent normal distribution, conditional on the hyper-parameters  $\mu$  and  $\tau^2$ . As Table D1 shows, the hierarchical prior leads to considerably higher employment rates, especially in the initial periods of Thompson sampling. This is as expected: the hierarchical prior captures important structure of the underlying data. Using this prior information leads to better allocations. Over time, the prior gets dominated by the data either way, and thus prior information becomes less relevant for algorithm performance.

## 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 US. 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, and Rediet Abebe for her advice.

## FUNDING

Teytelboym was supported by Economic and Social Research Council Grant ES/R007470/1.

## REFERENCES

- Agrawal, S. and N. Goyal (2012). Analysis of Thompson sampling for the multi-armed bandit problem. *Proceedings of the 25th Annual Conference on Learning Theory* vol. 23, pp. 39.1–39.26.
- Ahani, N., T. Andersson, A. Martinello, A. Teytelboym and A. C. Trapp (2021). Placement optimisation in refugee resettlement. *Operations Research* 69(5), 1349–650.
- Ahani, N., P. Gözl, A. D. Procaccia, A. Teytelboym and A. C. Trapp (2021). Dynamic placement in refugee resettlement. *Proceedings of the 22nd ACM Conference on Economics and Computation*, p. 5, New York, NY, USA: Association for Computing Machinery.
- Arnosti, N. and P. Shi (2020). Design of lotteries and wait-lists for affordable housing allocation. *Management Science* 66(6), 2291–307.

- Audibert, J.-Y., S. Bubeck and G. Lugosi (2014). Regret in online combinatorial optimisation. *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), 1876–902.
- Azevedo, E. M. and E. G. Weyl (2016). Matching markets in the digital age. *Science* 352(6289), 1056–57.
- 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), 325–29.
- Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery* 5(1), 27–36.
- Biró, P., J. van de Klundert, D. Manlove, W. Pettersson, T. Andersson, L. Burnapp, P. Chromy, P. Delgado, P. Dworzak, B. Haase, et al. (2021). Modelling and optimisation in European kidney exchange programmes. *European Journal of Operational Research* 291(2), 447–56.
- Bubeck, S. and M. Sellke (2020). First-order Bayesian regret analysis of Thompson sampling. *Proceedings of the 31st International Conference on 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). <https://www.osti.gov/biblio/1430202>.
- Chen, W., Y. Wang and Y. Yuan (2013). Combinatorial Multi-Armed Bandit: General Framework and Applications. In D. Sanjoy and M. David (eds.), *Proceedings of the 30th International Conference on Machine Learning*, vol. 28(1), pp. 151–59, PMLR: Atlanta, Georgia, USA.
- Gai, Y., B. Krishnamachari and R. Jain (2012). Combinatorial network optimisation with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking* 20(5), 1466–78.
- Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)* 41(2), 148–64.
- 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.
- Jagadeesan, M., A. Wei, Y. Wang, M. Jordan and J. Steinhardt (2021). Learning equilibria in matching markets from bandit feedback. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang and J.W. Vaughan (eds.), *Advances in Neural Information Processing Systems*, vol. 34, pp. 3323–35.
- Johari, R., V. Kamble and Y. Kanoria (2021). Matching while learning. *Operations Research* 69(2), 655–81.
- Kasy, M. and A. Sautmann (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica* 89(1), 113–32.
- Kaufmann, E., N. Korda and R. Munos (2012). T sampling: An asymptotically optimal finite-time analysis. *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 C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, 28, pp. 1297–305.
- Kveton, B., Z. Wen, A. Ashkan and C. Szepesvári (2015). Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pp. 535–43.
- 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. *Proceedings of Thirty Fourth Conference on Learning Theory*, pp. 2965–92.
- Lattimore, T. and C. Szepesvári (2020). *Bandit Algorithms*. United Kingdom: Cambridge University Press.

- MacDonald, D. E. (2019). Foster care: A dynamic matching approach. Working Paper.
- 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. Preprint 2006.06613, arXiv.
- Robinson-Cortés, A. (2019). Who gets placed where and why? An empirical framework for foster care placement. Working Paper.
- Russo, D. and B. Van Roy (2016). An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research* 17(1), 2442–71.
- 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), 500–22.
- Thakral, N. (2016). The public-housing allocation problem: Theory and evidence from Pittsburgh. Working Paper.
- 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), 285–94.
- van Dijk, W. (2019). The socio-economic consequences of housing assistance. Working Paper.
- Wager, S. and K. Xu (2021). Diffusion asymptotics for sequential experiments. Preprint 2101.09855, arXiv.
- Waldinger, D. (2021). Targeting in-kind transfers through market design: A revealed preference analysis of public housing allocation. *III*(8), 2660–96.
- Wang, S. and W. Chen (2018). Thompson sampling for combinatorial semi-bandits. *Proceedings of the 35th International Conference on Machine Learning*, pp. 5114–22.
- Zimmert, J. and T. Lattimore (2019). Connections between mirror descent, Thompson sampling and the information ratio. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, 32. <https://doi.org/10.48550/arXiv.1905.11817>.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publihe/sher's website:

Online Appendix  
Replication Package

*Co-editor Victor Chernozhukov handled this manuscript.*



## APPENDIX 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:**

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

- **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:**

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

- **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:

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

## APPENDIX 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 lemmata, before combining them in the proof of Theorem 3.1 itself.

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

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

**Proof of Lemma B.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  $\Theta_t$  and the definition of Thompson sampling.

(B3) By the Cauchy-Schwarz inequality (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 B.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 B.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  $\mathbf{E}_t[\tilde{Y}_{jt}] = \bar{\Theta}_{jt} \cdot D_{KL}(\bar{\Theta}_{jt}^*, \bar{\Theta}_{jt})$  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  $\tilde{Y}_{jt}$ ,  $I_t(A_j^*; \tilde{Y}_{jt})$ :

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

and thus

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

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

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

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

These steps hold for the following reasons.

(B6) Because the second term in equation (B5) is nonnegative.

(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 B.3 (BOUNDING THE SUM OF COMPONENT-WISE INFORMATION).

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

**Proof of Lemma B.3:**

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

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

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

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

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

These steps hold for the following reasons.

(B10) The chain rule of mutual information.

- (B11) The entropy reduction form of mutual information and the nonnegativity of (conditional) entropy.  
 (B12) The definition of entropy for  $A_j^*$ .  
 (B13) Jensen's inequality.  
 (B14) The inequality  $\log(1+x) \leq x$  for  $x = \frac{M}{J-M}$ .

□

**Proof of Theorem 3.1:**

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

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

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

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

These steps hold for the following reasons.

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

□

## APPENDIX C. RANDOMISATION INFERENCE

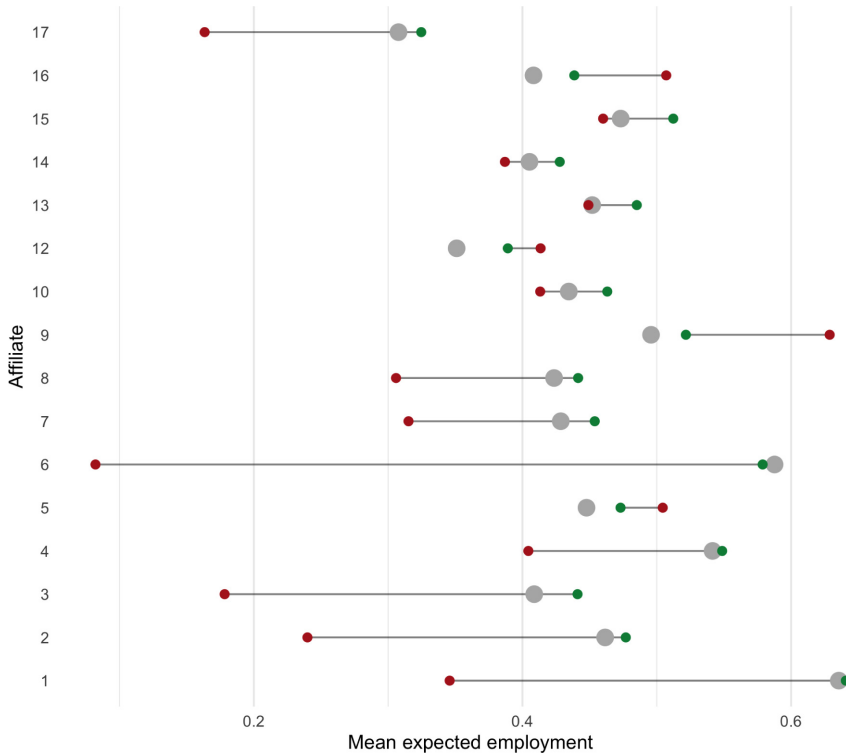
An alternative to Bayesian inference discussed in Section 4.2 is randomisation (permutation) inference. In the context of treatment effect estimation, randomisation inference can be used to test the null hypothesis that treatment does not affect any outcome, so that for instance  $Y_i^1 = Y_i^0$  for 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  $Y_i^0(a)$  for any counterfactual action  $a \in \mathcal{A}$  under the null hypothesis  $H^0$ , given knowledge of  $Y_t(A_t)$  for the realised 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 formalise 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.

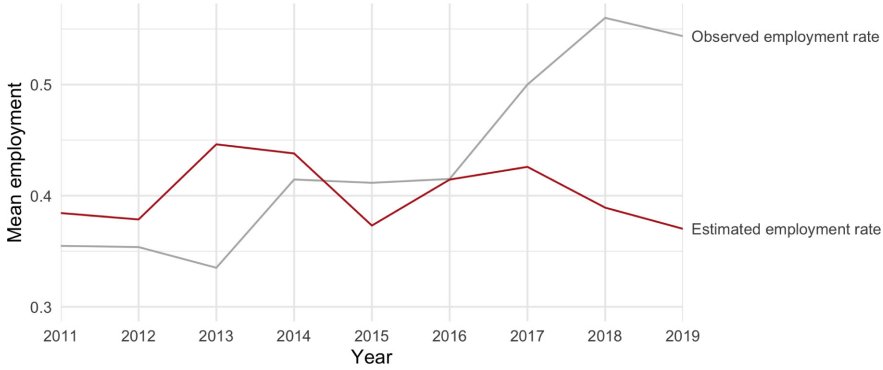
Given some specification of counterfactual outcomes, we can sample counterfactual histories  $\tilde{\mathcal{F}}_t$  by re-running the Thompson sampling algorithm iteratively. In each period  $s$ , draw  $\tilde{\Theta}_{t'}$  and the corresponding  $\tilde{A}_{t'}$  from the posterior given  $\tilde{\mathcal{F}}_{t'}$ . Impute a counterfactual outcome vector  $Y_{t'}^0(\tilde{A}_{t'})$ , based on the null hypothesis to be tested. Update the history  $\mathcal{F}_{t'}$  by adding  $\tilde{A}_{t'}$ ,  $Y_{t'}^0(\tilde{A}_{t'})$ , and iterate for the next period. Once  $t' = t$ , calculate a realisation 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.

APPENDIX D. ADDITIONAL EMPIRICAL RESULTS



**Figure D1.** Actual and counterfactual expected employment rates by affiliate.

*Notes:* Simulations based on refugee resettlement data described in Section 5. Employment rates for refugees with no US ties only. Grey: average of 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  $\Theta_0$ , subject to actual capacity constraints.



**Figure D2.** Observed versus estimated employment rates.

*Notes:* Based on refugee resettlement data described in Section 5. Employment rates for refugees with no US ties only. This figure compares observed employment rates in each year (grey) to estimated employment rates based on the estimated parameters  $\Theta_0$  and the actual demographics of arrivals and their allocation to affiliates.

**Table D1.** Independent versus hierarchical prior for Thompson sampling.

Year	Actual	Hierarchical prior	Independent prior	Difference
2011	0.384	0.384	0.384	0.000
2012	0.379	0.398	0.382	0.016
2013	0.446	0.466	0.460	0.007
2014	0.438	0.461	0.460	0.001
2015	0.373	0.414	0.410	0.004
2016	0.414	0.457	0.454	0.003
2017	0.426	0.468	0.468	0.000
2018	0.389	0.418	0.418	0.000
2019	0.370	0.399	0.400	-0.002

*Notes:* This table compares the performance of Thompson sampling in simulations for two different priors: The hierarchical prior (our preferred approach, as discussed in the main text, and shown in Figure 1), and the prior with independence across matches (the approach in the existing literature). The variable ‘Difference’ is the increase of employment due to using the hierarchical prior, relative to the independent one. Employment rates for refugees with no US ties only.