Adaptive combinatorial allocation:
How to use limited resources while learning what works

Maximilian Kasy Alexander Teytelboym

May 2020

Introduction

Many policy problems have the following form:

- Resources, agents, or locations need to be allocated to each other.
- There are various feasibility constraints.
- The returns of different options (combinations) are unknown.
- The decision has to be made repeatedly.

Sketch of setup

- There are J **options** (e.g., matches) available to the policymaker.
- Every period, the policymaker's **action** is choose at most M options.
- Before the next period, the policymaker observes the **outcomes** of every chosen option (combinatorial semi-bandit setting).
- Policymaker's **reward** is the sum of the outcomes of chosen options.
- Policymaker's **objective** is to maximize the cumulative expected rewards.
- Alternatively, policymaker's objective to minimize **expected regret**—the cumulative difference between the (same) optimal action taken every period and the actual actions.

Example 1: Gender composition in classrooms

- **Outcome**: classroom average test score.
- **Batch size** (M): # classrooms every term/year.
- **# Options** (J): classroom size × # classrooms.
- **Constraints**:
  - respecting classroom capacities;
  - total # girls (# boys) across all classrooms is fixed.

Example 2: Foster family placement

- **Outcome**: a measure of well-being of foster children.
- **Batch size** (M): # children.
- **# Options** (J): # children × # foster families.
- **Constraints**:
  - respecting capacity of each foster family;
  - siblings must be placed together;
  - child-family match feasibility.

Example 3: Combinations of therapies

- **Outcome**: health outcomes of patients.
- **Batch size** (M): # patients.
- **# Options** (J): # patients $\times 2^{\#\text{therapies}}$.
- **Constraints**:
  - respecting constraints on medical resources (e.g., doctor time, testing kits etc.);
  - avoiding deadly therapy combinations for certain patients (e.g., allergies).

Overview of the results

- In each example, the number of actions available to the policymaker is huge, e.g., there are $\binom{J}{M}$ ways to choose M out of J possible options/matches.
- The policymaker's decision problem is a computationally intractable dynamic stochastic optimization problem.
- Our heuristic solution is **Thompson sampling**—in every period the policymaker chooses an action with the posterior probability that this action is optimal.
- We derive a **finite-sample, prior-independent bound on expected regret**: surprisingly, per-unit regret only grows in $\sqrt{J}$ and does *not* grow in M.
- We illustrate the performance of our bound with **simulations**.
- We would love feedback on two proposed **applications**—experimental (MTurk) and empirical (refugee resettlement).

Literature

- Optimal solutions to bandit problems:
  Gittins (1979); Keller and Rady (1999).
- Thompson sampling:
  Thompson (1933); Russo et al. (2018).
- Performance guarantees:
  Agrawal and Goyal (2012, 2013); Audibert et al. (2014).

  We draw in particular on:
  Russo and Van Roy (2016) and Bubeck and Sellke (2020).
- Adaptive policies in economics:
  Kasy and Sautmann (2019); Kasy and Teytelboym (2020); Caria et al. (2020).

Introduction

Setup

Performance guarantee

Simulations

Applications

Setup

- Options $j \in \{1, \dots, J\}$.
- Sufficient resources to select only $M \leq J$ options.
- Feasible combinations of options:

$$a \in \mathcal{A} \subseteq \{a \in \{0,1\}^J : \|a\|_1 = M\}.$$

- Periods: $t = 1, \dots, T$.
- Vector of potential outcomes (i.i.d. across periods):

$$Y_t \in [0,1]^J.$$

- Average potential outcomes:

$$\Theta_j = \mathbf{E}[Y_{jt}|\Theta].$$

- Prior belief over the vector $\Theta \in [0,1]^J$ with arbitrary dependence across j.

Observability

- After period t, we observe outcomes for all chosen options:

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

- Thus actions in period t can condition on information

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

- These assumptions makes our setting a "semi-bandit" problem.
(Because we observe more than just $\sum_j a_j \cdot Y_{jt}$ as in a bandit problem!)

Objective and regret

- Reward for action a:
$$\langle a, Y_t \rangle = \sum_j a_j \cdot Y_{jt}.$$
- Expected reward:
$$R(a) = \mathbf{E}_t[\langle a, Y_t \rangle | \Theta] = \langle a, \Theta \rangle.$$
- Optimal action:
$$A^* \in \underset{a \in \mathcal{A}}{\operatorname{argmax}} \; R(a) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \; \langle a, \Theta \rangle.$$
- Expected *regret* at T:
$$\mathbf{E}_1 \left[\sum_{t=1}^{T} (R(A^*) - R(A_t)) \right].$$

Thompson sampling

- Take a random action $a \in \mathcal{A}$, sampled according to the distribution

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

- This assumption implies in particular that

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

- Introduced by Thompson (1933) for treatment assignment in adaptive experiments.

Introduction

Setup

Performance guarantee

Simulations

Applications

Regret bound

Theorem

Under the assumptions just stated,

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

Features of this bound:

- It holds in finite samples, there is no remainder.
- It does not depend on the prior distribution for Θ.
- It allows for prior distributions with arbitrary statistical dependence across the components of Θ.
- It implies that Thompson sampling achieves the efficient rate of convergence.

Regret bound

Theorem

Under the assumptions just stated,

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

Verbal description of this bound:

- The worst case expected regret (per unit) across all possible priors goes to 0 at a rate of 1 over the square root of the sample size, $T \cdot M$.
- The bound grows, as a function of the number of possible options J, like $\sqrt{J}$ (ignoring the logarithmic term).
- Worst case regret per unit does not grow in the batch size M, despite the fact that action sets can be of size $\binom{J}{M}$!

Key steps of the proof

1. Use Pinsker's inequality to **relate expected regret** to the information about the optimal action A^*. Information is measured by the **KL-distance** of posteriors and priors.

   (This step draws on Russo and Van Roy (2016).)

2. Relate the **KL-distance** to the **entropy reduction** of the events $A_j^* = 1$.

   The combination of these two arguments allows to bound the expected regret for option j in terms of the entropy reduction for the posterior of A_j^*.

   (This step draws on Bubeck and Sellke (2020).)

3. The total **reduction of entropy** across the options j, and across the time periods t, can be no more than the **sum of the prior entropy** for each of the events $A_j^* = 1$, which is bounded by $M \cdot \left[\log\left(\frac{J}{M}\right) + 1\right]$.

Introduction

Setup

Performance guarantee

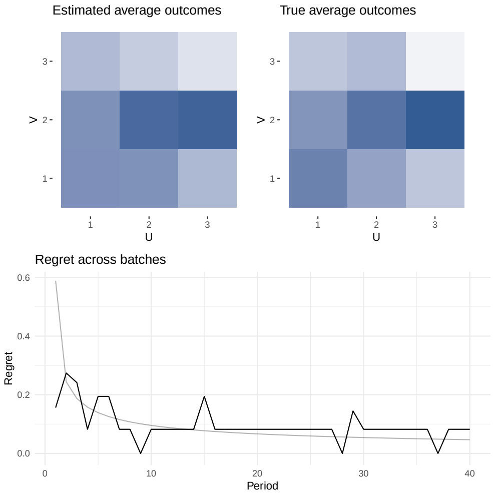
Simulations

Applications

Simulations

- Two-sided, one-to-one matching.
  - Equal number of types on either side: $U = V$, ranging from 3 to 8.
  - Batch size (number of units on either side) equal to number of options $M = J = U \cdot V$, ranging from 9 to 64.
  - $T = 40$ periods (batches).
- Θ is a draw from the model
  - $\Theta_{u,v} = g\left(\alpha_u + \beta_v + \gamma_{u,v}\right)$,
  - where α_u, β_v and $\gamma_{u,v}$ are i.i.d. $N(0,1)$,
  - and $g(x) = \exp(x)/(1 + \exp(x))$ is the logit link function.
- We use the same model as prior for Thompson sampling.
- The plots show:
  1. True Θ and estimated Θ at the end of the simulation.
  2. Actual per-unit regret in each period.
  3. Our theoretical bound on expected regret.

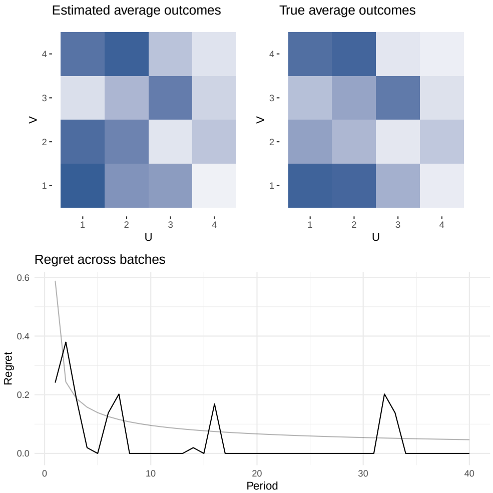
Simulations



Estimated average outcomes

True average outcomes

Regret across batches

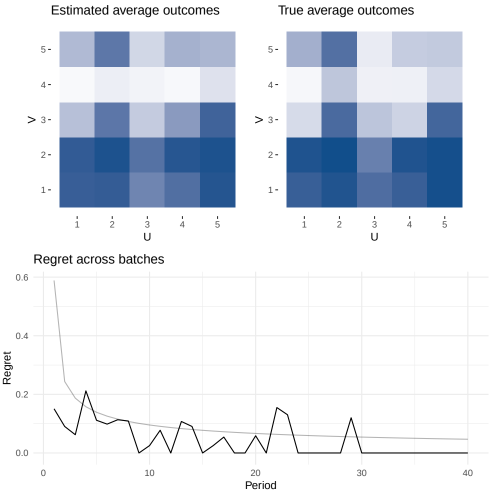


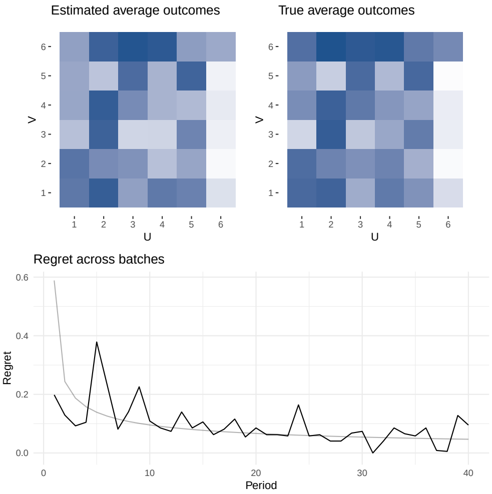
Estimated average outcomes

True average outcomes

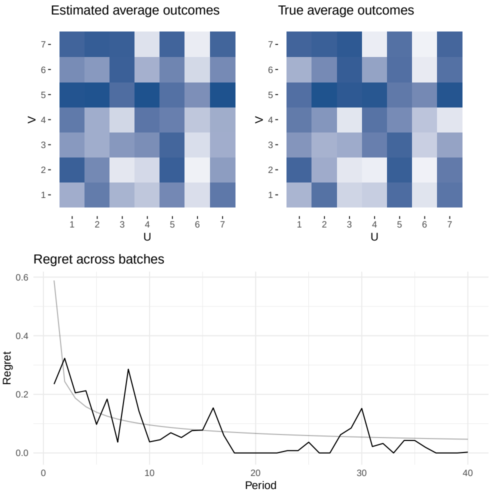
Regret across batches

Simulations



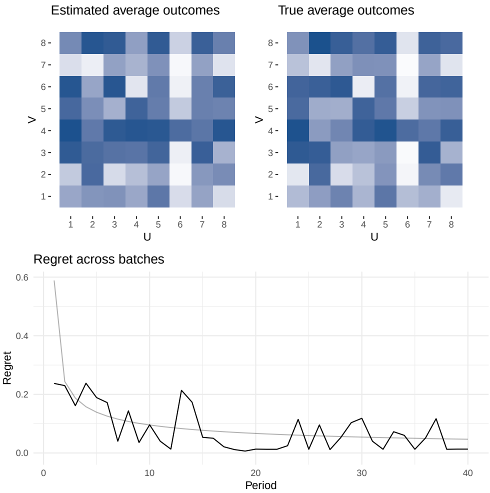


Simulations


Estimated average outcomes

True average outcomes

Regret across batches


Estimated average outcomes

True average outcomes

Regret across batches

MTurk Matching Experiment

- **Outcome**: self-assessed usefulness to personalized advice (score 1–5).
- **Batch size** (M): $\#$ senders $= \#$ receivers.
- **# Options** (J): $\#$ sender types $\times$ $\#$ receiver types.
- **Constraints**: one-to-one matching.

MTurk Matching Experiment: Proposed Design

- 4 types = {Indian, American} × {College, No College}
- 16 agents per batch, 4 of each type.
- Instruction to sender:

  Reflect on your past experience as a worker. [...] Write a message to another person, who is from India and college-educated, that includes useful advice, tips, and stories from your own experiences.

- Instruction to receiver: Read the message and score (1–5), e.g.,:

  I would benefit from receiving another message from this person.

  This message contained advice that is useful to me.

  The person who wrote this message is good at sensing what others are feeling.

Adaptive Refugee Resettlement

- **Outcome**: 90-day employment of refugees resettled in the US.
- **Batch size** (M): # refugee families (weekly)
- **# Options** (J): # refugees families × # localities.
- **Constraints**:
  - locality capacity constraints;
  - service provision feasibility (multidimensional knapsack constraints).

Adaptive Refugee Resettlement: Proposed Simulation

- Detailed data on refugee resettlement from HIAS for 2011-2020.
- Around 20 localities and thousands of refugees.
- Refugees arrive every week; observe their employment after 90 days.
- Observe many refugee covariates; treat localities as fixed.
- Use employment probabilities estimated from past data as “true” parameters and simulate the performance of Thompson sampling on a recent year of data.
- Should give us an idea of how quickly the agency can adapt its matching system to sudden changes in refugee flows.

Thank you!