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

Maximilian Kasy   Alexander Teytelboym

August 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.

Examples

1. **Demographic composition of classrooms**
   - Distribute students across classrooms,
   - to maximize test scores in the presence of (nonlinear) peer effects,
   - subject to overall demographic composition, classroom capacity.
2. **Foster family placement**
   - Allocate foster children to foster parents,
   - to maximize child outcomes,
   - subject to parent capacity, keeping siblings together, match feasibility.
3. **Combinations of therapies**
   - Allocate (multiple) therapies to patients,
   - respecting resource constraint, medical compatibility.

Sketch of setup

- There are J **options** (e.g., matches) available to the policymaker.
- Every period, the policymaker's **action** is to choose at most M options.
- Before the next period, the policymaker observes the **outcomes** of every chosen option (combinatorial semi-bandit setting).
- The policymaker's **reward** is the sum of the outcomes of the chosen options.
- The policymaker's **objective** is to maximize the cumulative expected rewards.
- Equivalently, the policymaker's objective is to minimize **expected regret**—the shortfall of cumulative expected rewards relative to the oracle optimum.

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**.
- Work in progress: **Applications**—experimental (MTurk) and observational (refugee resettlement).

Setup

- Options $j \in \{1, \ldots, J\}$.
- Only sufficient resources to select $M \leq J$ options.
- Feasible combinations of options:

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

- Periods: $t = 1, \ldots, T$.
- 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 the information

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

- These assumptions make our setting a "semi-bandit" problem:
  We observe more than just $\sum_j a_j \cdot Y_{jt}$,
  as we would in a bandit problem with actions a!

Objective and regret

- Reward for action a:

$$\langle a, Y_t \rangle = \sum_j a_j \cdot Y_{jt}.$$

- Expected reward:

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

- 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.

Regret bound

Theorem

Under the assumptions just stated,

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

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

MTurk Matching Experiment: Proposed Design

- Matching message senders to receivers based on types.
- 4 types $= \{\text{Indian, American}\} \times \{\text{Female, Male}\}$
- 16 agents per batch, 4 of each type, for both senders and recipients.
- Instruction to sender:

  In your message, please share advice on how to best reconcile online work with family obligations. In doing so, please reflect on your own past experiences. [. . .] The person who will read your message is an Indian woman.

- Instruction to receiver: Read the message and score on 13 dimensions (1–5), e.g.,:

  The experiences described in this message are different from what I usually experience.

  This message contained advice that is useful to me.

  The person who wrote this understands the difficulties I experience at work.

Introduction

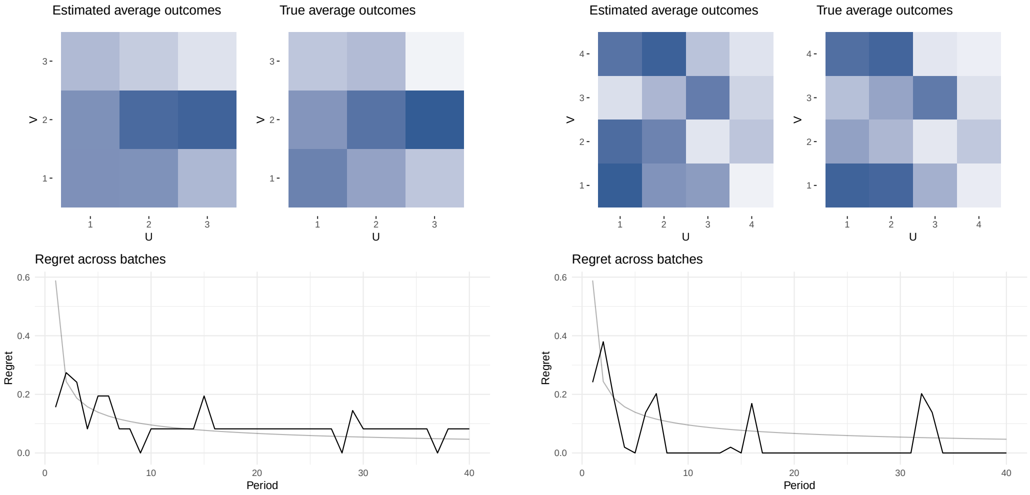
Setup

Performance guarantee

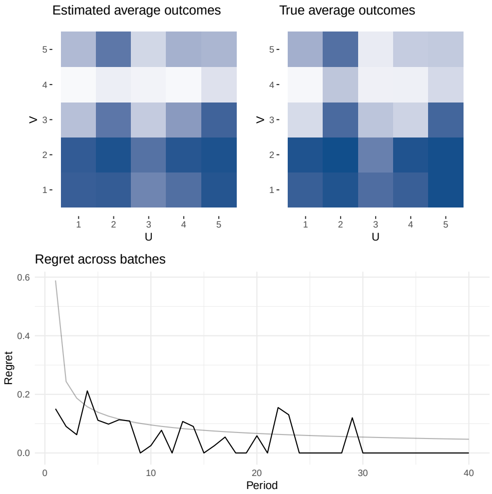
Applications

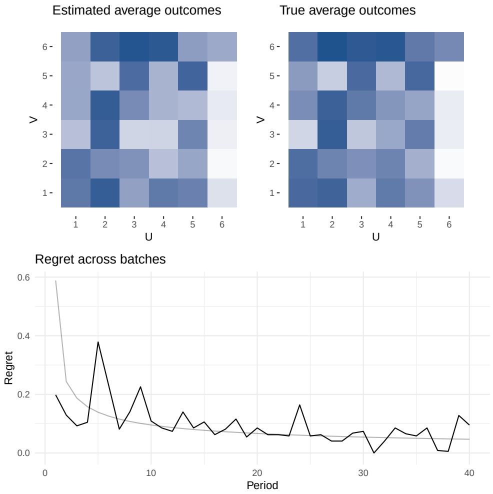
Simulations

Simulations

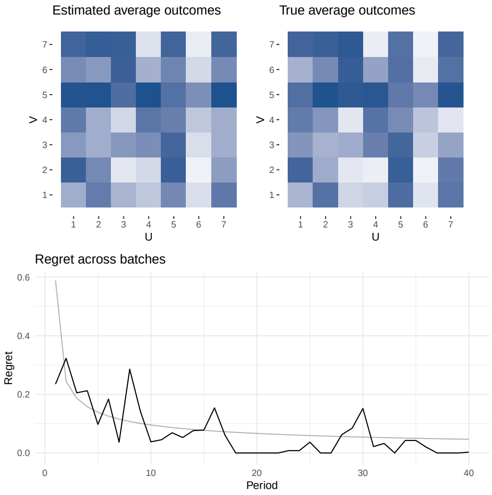


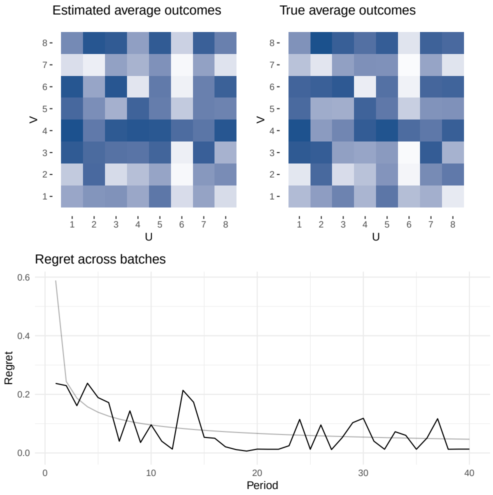
Simulations





Simulations





Thank you!