Contextual Bandits in a Survey Experiment on Charitable Giving: Within-Experiment Outcomes versus Policy Learning

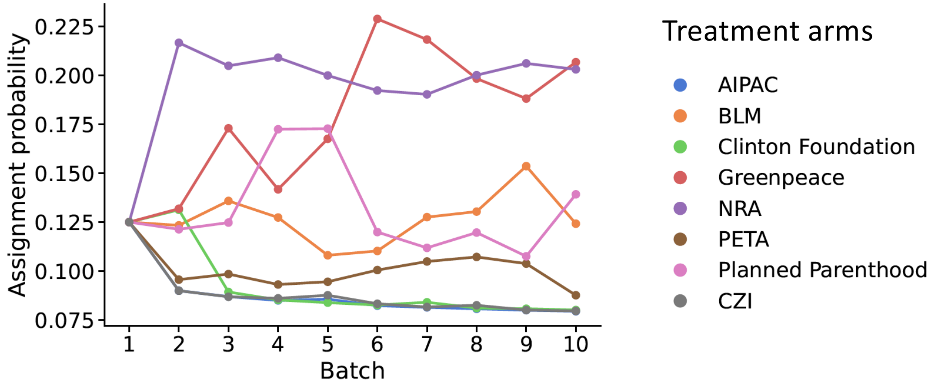
Susan Athey, Undral Byambadalai, Vitor Hadad,

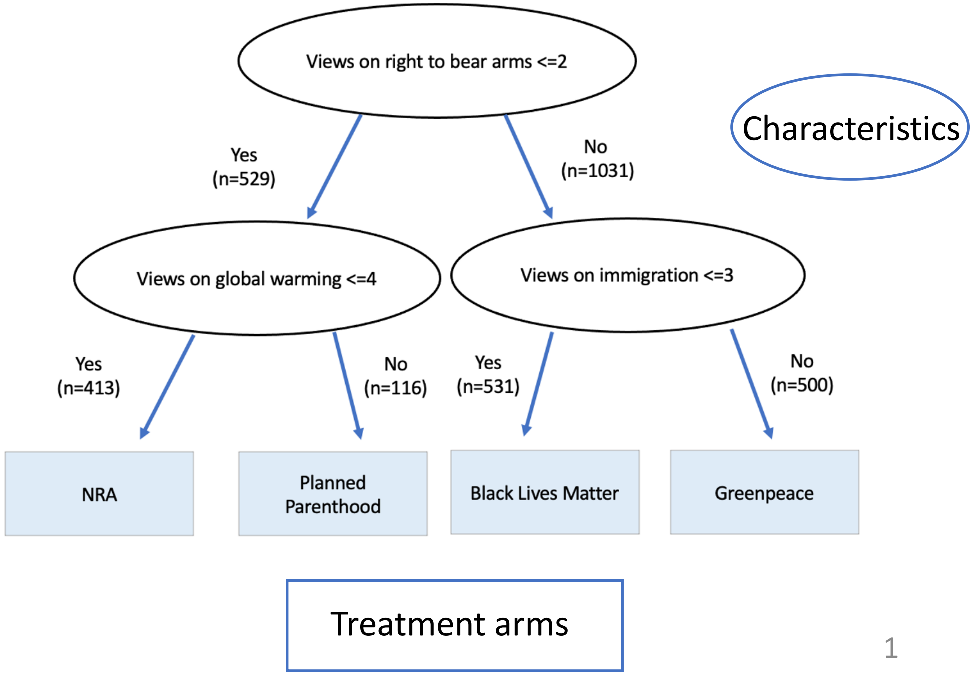
Sanath Kumar Krishnamurthy, Weiwen Leung, Joseph Jay Williams

Microeconometrics Class of 2020 and 2021 Conference, Duke University

Introduction

- Our goal: designing a “contextual bandit” - an **adaptive experiment** with multiple arms - where goal is to learn a **targeted treatment assignment rule**
- Tension arises between **within-experiment** outcome maximization and finding best policy to use AFTER experiment (**“policy learning”**)
- Propose a heuristic algorithm that **balances the two goals**.
- **Implement** our method in charitable giving experiment.
- **Compare** with other existing contextual bandit algorithms using semi-synthetic data based on our experimental data.





Setup and notation

We consider the stochastic **contextual bandit** setting with K treatment arms.

Treatment arms:

- $w_t \in [K] \equiv \{1, \dots, K\}$

User arrives at time t:

- $x_t \in \mathbb{R}^p$ covariates (context)
- $Y_t(1), Y_t(2), \dots, Y_t(K)$ potential outcome vector

Algorithm at time t:

- Observes covariates x_t
- Uses past observations to construct assignment probabilities p_t
- Selects a treatment arm $w_t \sim p_t(\cdot \,|x_t)$
- Observes outcome $Y_t(w_t) \in \mathbb{R}$

Unknown to the algorithm:

- Conditional mean outcome function:

$$f(x, w) := \mathbb{E}[Y_t(w)|x]$$

- Optimal policy:

$$\pi_f(x) := \arg\max_w f(x, w)$$

Goal 1: Cumulative regret

Most common objective for contextual bandit algorithms: cumulative regret minimization (maximize expected outcome DURING the experiment)

Cumulative regret: $\sum_{t=1}^{T}\left(f\left(x_t, \underbrace{\pi_f(x_t)}_{\text{Optimal policy}}\right) - f(x_t, \underbrace{w_t}_{\text{Selected treatment arm}})\right)$

Conditional mean outcome function:

$$f(x,w) := \mathbb{E}[Y_t(w)|x].$$

Optimal policy:

$$\pi_f(x) := \underset{w}{\operatorname{argmax}} f(x,w).$$

Goal 2: Policy learning (aka "Simple regret")

Policy value is given by

$$R_f(\pi) = \mathbb{E}\left[f\big(x, \pi(x)\big)\right].$$

Given a policy class Π, the optimal in-class policy is given by

$$\pi^* \in \mathrm{argmax}_{\pi \in \Pi} R_f(\pi).$$

At the end of an adaptive experiment, we also want to learn a policy $\hat{\pi} \in \Pi$ with low simple regret.

Simple regret: $R_f(\pi^*) - R_f(\hat{\pi})$

Tension between simple regret (policy learning) and cumulative regret (within-experiment outcomes)

Consider the task of constructing the assignment rule p_t.

- We want the following to be small for simple regret:

$$V(p_t, \pi^*) := E\left[\frac{1}{p_t(\pi^*(x)|x)}\right]$$

- However, for cumulative regret we want the following to be large:

$$E\left[\sum_{w \in [K]} f(x, w) p(w|x)\right]$$

- If we know π^*, we can set $p_t(\pi^*(x)|x) = 1$ for all x and do well on both objectives.

Uncertainty in estimating π^* introduces tensions between the two quantities.

- Uniformly sampling arms ensures $V(p_t, \pi^*) = K$.
- In attempting to place a higher probability on the estimated optimal arm at any context, "aggressive algorithms" may make $V(p_t, \pi^*) > K$.

See Krishnamurthy et al 2023 (arxiv.org/pdf/2307.02108.pdf) for recent developments

Survey experiment

Contexts:

age, gender, race, religious or not, urban/rural, political affiliation, last donation

views on immigration, views on global warming, views on right to bear arms, views on abortion

how often watch/read Fox News, CNN, WSJ

Treatment:

Please take a few seconds to review the information below. **In the next page, we'll ask you a question about this organization.**

AIPAC

The American Israel Public Affairs Committee (AIPAC) is a lobbying group that advocates pro-Israel policies to the Congress and Executive Branch of the United States. The current president of AIPAC is Betsy Berns Korn.

Outcome:

How would people like you feel if we donated 1,000 USD to the organization shown in the previous page?

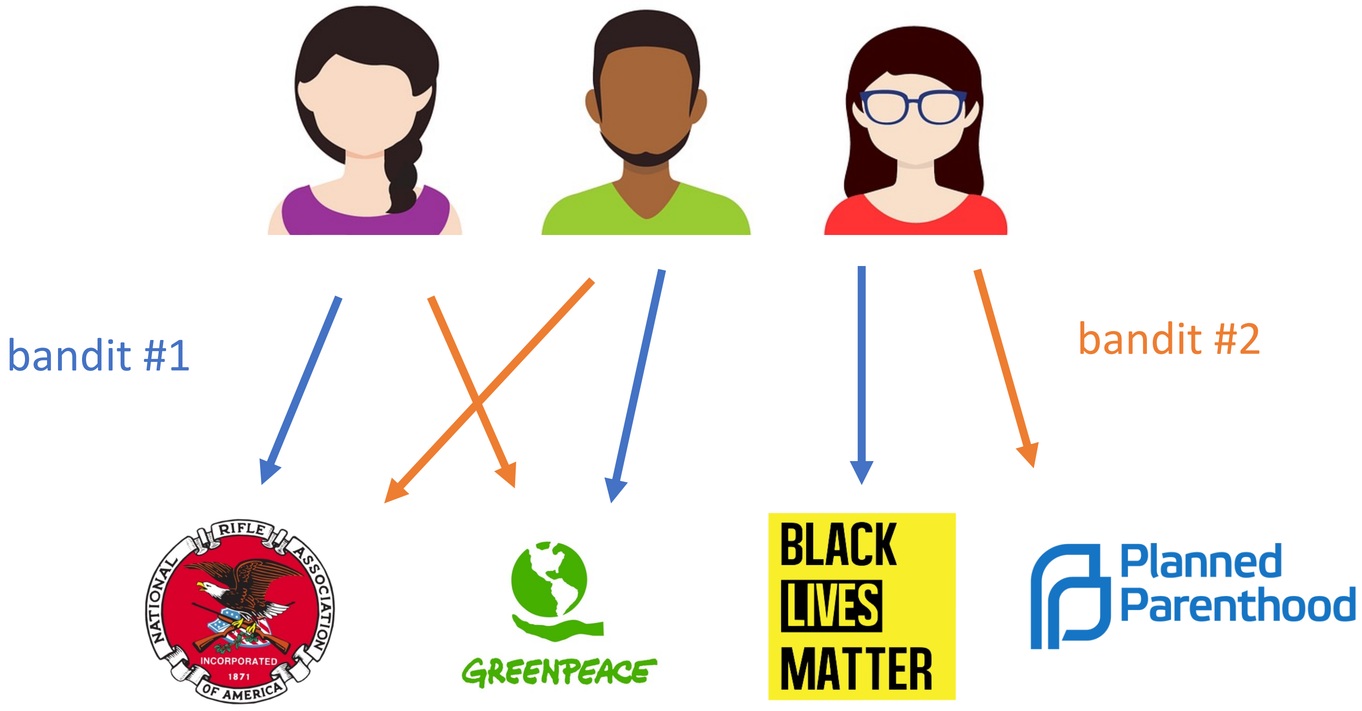
Please drag the slider to indicate your estimate, with ***-10 being extremely dissatisfied***, and ***10 being extremely satisfied***.

-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10

Feelings Thermometer

Simulations based on semi-synthetic data

- Contextual bandits algorithms guide data collection ⇒ not straightforward to reanalyze historically collected data to compare algorithms
    - For a given x, a different algorithm would assign a different treatment than what was observed
- Running many parallel experiments to compare algorithms can be costly ⇒ rely on simulations based on semi-synthetic data



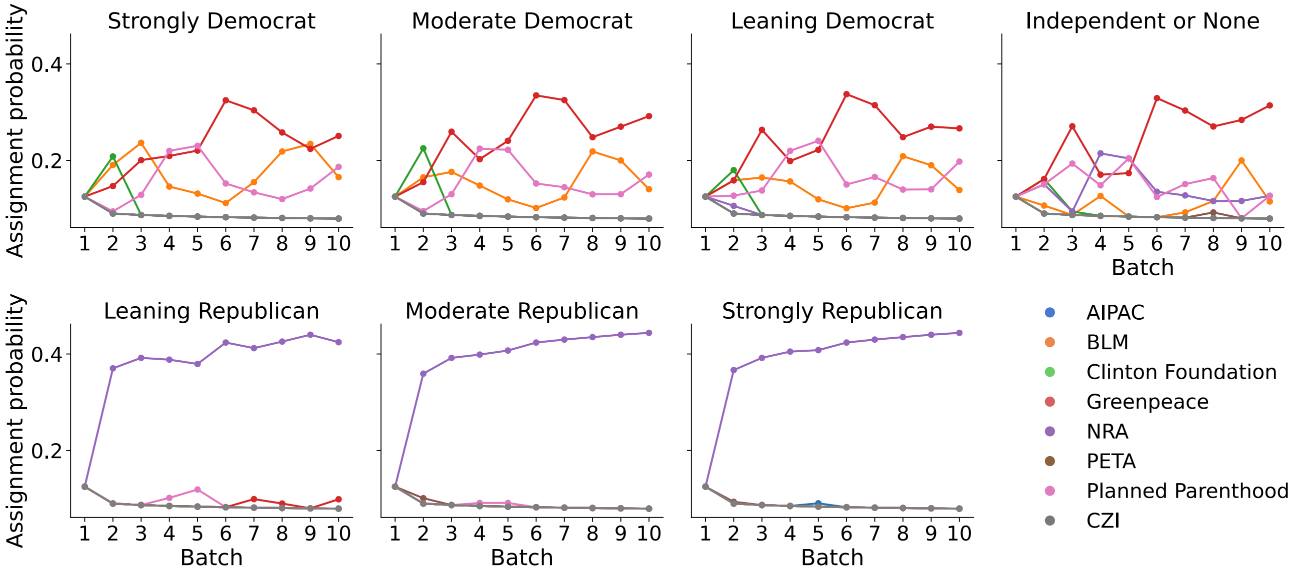
TreeBagging algorithm

Obtain assignment probabilities using tree-policy based bagging algorithm

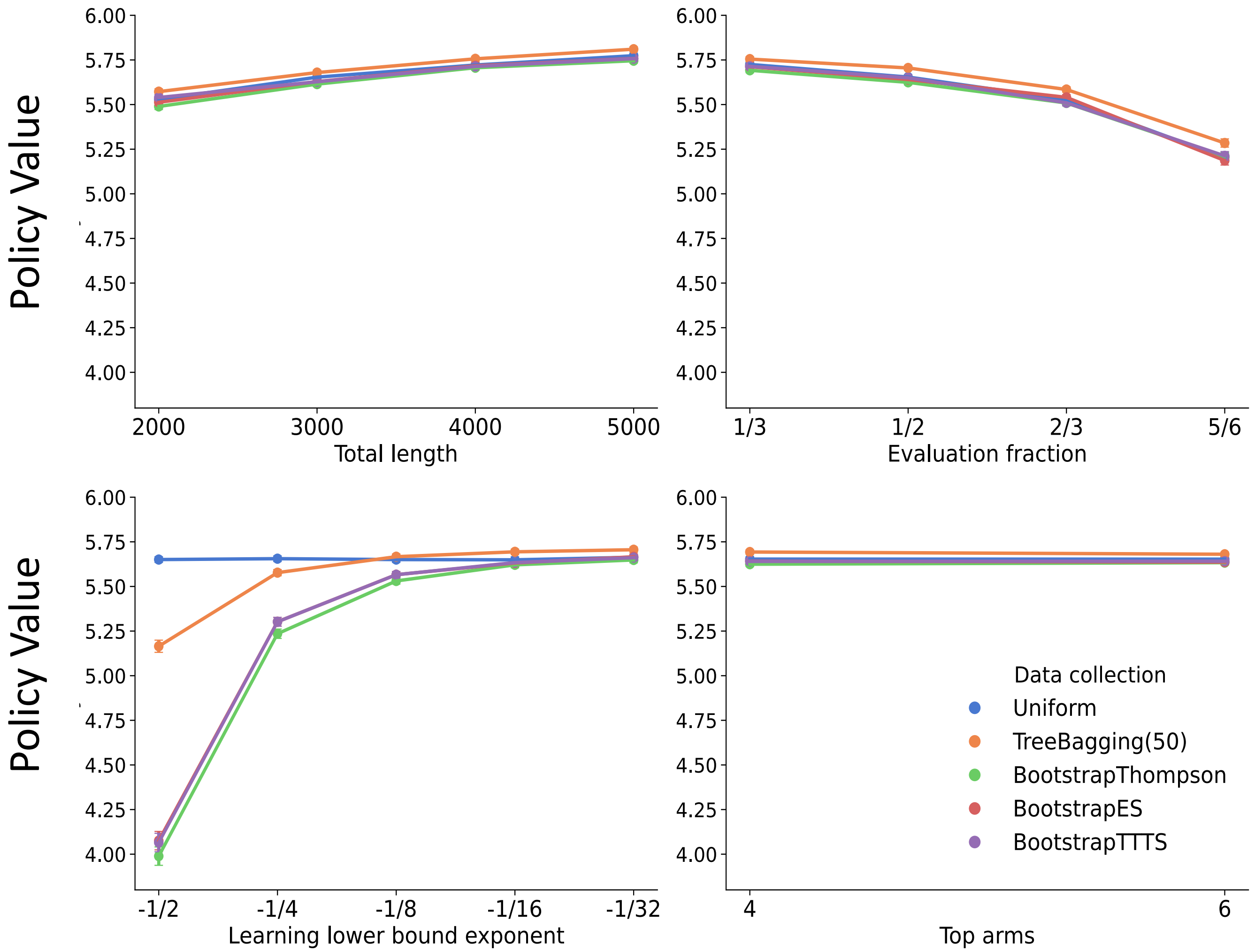
(1) Impose **a decaying lower bound** on the assignment probabilities p_t (bounds variance of policy estimate $V(p_t, \pi^*)$)

(2) **Tree-bagging** with **shallow trees** avoids extrapolation from limited data (robust to mis-specification), avoids using arms that show benefits for very small set of covariates

(3) At the end of the experiment, **drop least-favored arms** and learn a policy using only the top arms. (known to work well in non-contextual bandits)



Uniform treatment assignment (pure RCT) beats adaptive assignment for policy learning



Conduct semi-synthetic simulations based on pilot data.

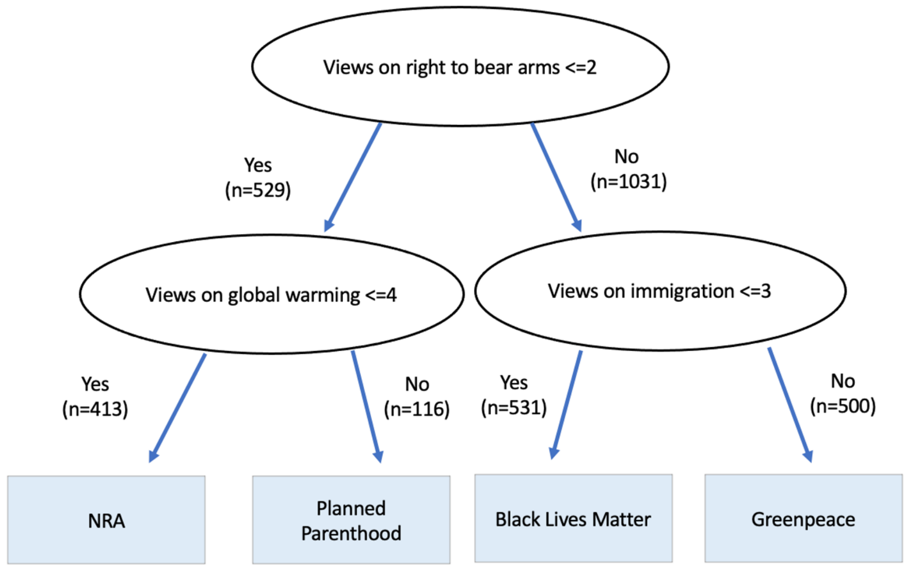
Each subplot shows the average value of learned policies across simulations as we vary one tuning parameter (keeping rest at "optimal" values).

Contextual bandit algorithms have been optimized for cumulative regret.

Uniform randomization beats most contextual bandits.

Traditional (-1/2) decay rates for lower bound on assignment bad for policy learning.

Targeted policy vs. best non-targeted policy



	Value	Std.err	Diff	Std.err	p-value
Best non-targeted policy (Greenpeace)	4.687	0.208			
Targeted policy	5.653	0.216	0.966	0.300	0.001

Views on immigration: The US government needs to get tougher on immigration
Views on global warming: The US government should do more to prevent global warming
Views on right to bear arms: The right to bear arms should be limited

1- Strongly disagree, 2 - Somewhat disagree, 3 - Neither agree nor disagree, 4 - Somewhat agree, 5 - Strongly agree

Conclusion

- We consider the problem of designing an adaptive experiment when the goal is to learn a personalized treatment assignment rule.
- Existing contextual bandit algorithms are too "aggressive" in discarding arms and don't do well in policy learning compared to uniform randomization.
- We propose a heuristic algorithm called TreeBagging and apply it in a real-world experiment, learning a targeted treatment assignment policy that significantly outperforms the best non-targeted policy.
- Semi-synthetic simulations show that TreeBagging outperforms uniform randomization for policy learning while yielding a substantial reduction in cumulative regret; not true for standard contextual bandit algorithms.

Thank you!