Treatment Choice with Bandit

János K. Divényi

March 14, 2018

Central European University

How to assign individuals to treatments?

Traditional literature of Manski (2004), Kitagawa and Tetenov (2017) derives the best treatment rules based on a given experimental sample.

How to assign individuals to treatments?

Traditional literature of Manski (2004), Kitagawa and Tetenov (2017) derives the best treatment rules based on a given experimental sample.

My approach

Experiment itself is a decision.

Consider sequential choice: observing outcomes before deciding about new participants. How to balance between exploring and exploiting?

Contribution

1. Borrow the multi-armed bandit methodology from ML literature and apply it to the Rubin Causal Model
2. Show on JTPA that it results in considerable welfare gains
3. Ask open questions (and hope to answer them later)

Motivation

Motivation

We want to spend tax payers' money effectively on large social programs.

Current approach: run RCT to learn about the effect.

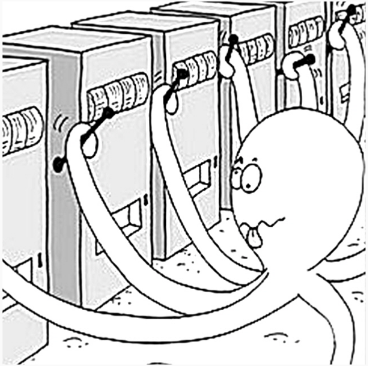
Motivation

We want to spend tax payers' money effectively on large social programs.

Current approach: run RCT to learn about the effect.

Digitalization makes experimenting cheaper:
one-shot experimentation $\rightarrow$ constant exploration-exploitation

Multi-armed bandit



How to optimize which arm to pull?

Learn the rewards $\leftrightarrow$ Pull the highest

Exploration $\leftrightarrow$ Exploitation

Relation to literature

- **Treatment choice** Manski (2004), Dehejia (2005), Kitagawa and Tetenov (2017)
- **Multi-armed bandit** Robbins (1952), Bubeck and Cesa-Bianchi (2012), Szepesvári and Lattimore (2018)
- **Dynamic treatment allocation** Perchet et al. (2016), Kock and Thyrsgaard (2018)

Formal setup

Problem

Rubin Causal model, binary treatment, potential outcomes: $Y(1), Y(0)$

Problem

Rubin Causal model, binary treatment, potential outcomes: $Y(1), Y(0)$

Goal: find treatment rule $\pi : i \to \{0, 1\}$ that maximizes utilitarian welfare

Problem

Rubin Causal model, binary treatment, potential outcomes: $Y(1), Y(0)$

Goal: find treatment rule $\pi : i \rightarrow \{0, 1\}$ that maximizes utilitarian welfare

Sequential arrival: j is assigned $\rightarrow$ $Y(\pi_j)$ is observed $\rightarrow$ $j + 1$ is assigned...

How to maximize welfare?

Common approach: derive *regret*, ie. expected welfare loss relative to the maximum feasible welfare:

$$R(n) = \sum_{i=1}^{n} \mathbb{E}\left[Y_i(\pi^*(i)) - Y_i(\pi(i))\right].$$

Good rules achieve low regret across all states of nature - *minimax optimality*

Traditional approach: RCT

1. Choose a sample size of m.
2. Assign the first m individuals with 50% probability to the treatment.
3. After m individuals, compare the average outcomes (treatment effect).
4. If the effect is positive, apply the treatment to everyone onwards.

Bandit approach: UCB

1. Assign the first two individuals to treatment and control.
2. Calculate averages and get an upper bound for the estimate.
3. Assign the individual to the treatment if the upper bound of its mean estimate is higher.

Bandit approach: UCB

1. Assign the first two individuals to treatment and control.
2. Calculate averages and get an upper bound for the estimate.
3. Assign the individual to the treatment if the upper bound of its mean estimate is higher.

Intuitively: choose the treatment if (1) we are uncertain about its expected outcome (exploration), or (2) we are certain that its expected outcome is high (exploitation).

Bandit approach: UCB

1. Assign the first two individuals to treatment and control.
2. Calculate averages and get an upper bound for the estimate.
3. Assign the individual to the treatment if the upper bound of its mean estimate is higher.

Intuitively: choose the treatment if (1) we are uncertain about its expected outcome (exploration), or (2) we are certain that its expected outcome is high (exploitation).

It is shown to be minimax optimal (Lai and Robbins, 1985)

Bandit approach: UCB (Szepesvári and Lattimore, 2018)

Upper bound for each outcome $k \in \{0, 1\}$ calculated for each individual i

$$\bar{Y}_{i-1}(k) + \sqrt{\frac{2 \cdot \log\left(1 + i \cdot \log^2(i)\right)}{N_k(i-1)}}$$

where $N_k(i-1)$: $\#$ individuals assigned to treatment k before i arrives

The bound can be derived from the *Hoeffding's bound*

$$\mathbb{P}\left(\left|\bar{Y}_n - \mathbb{E}[Y]\right| > \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$$

using

$$\exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) = \frac{1}{1 + n\log^2(n)}$$

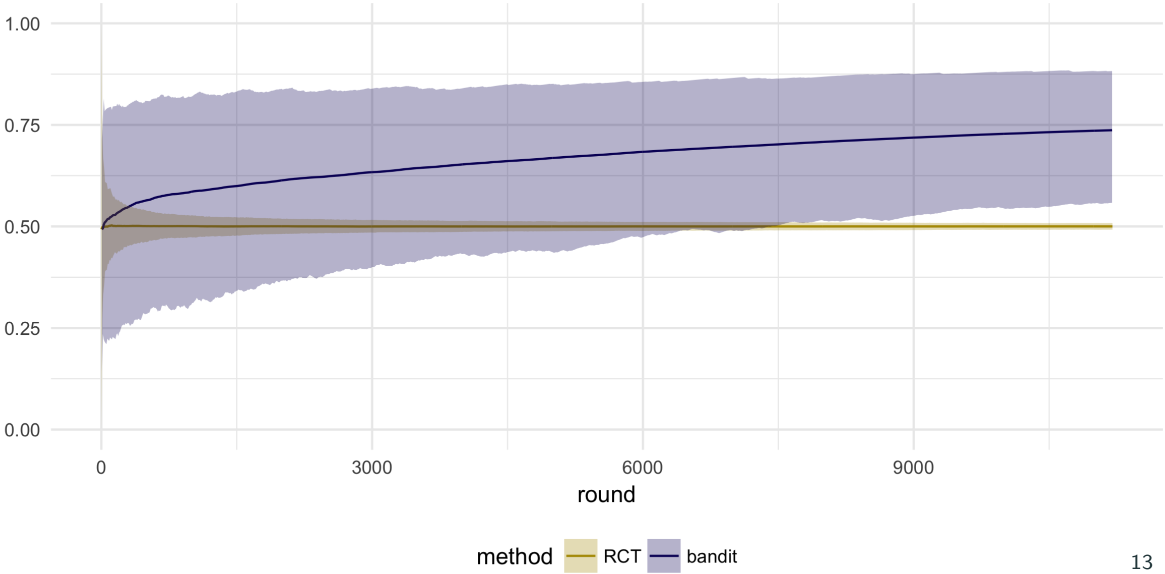
Illustration

JTPA study

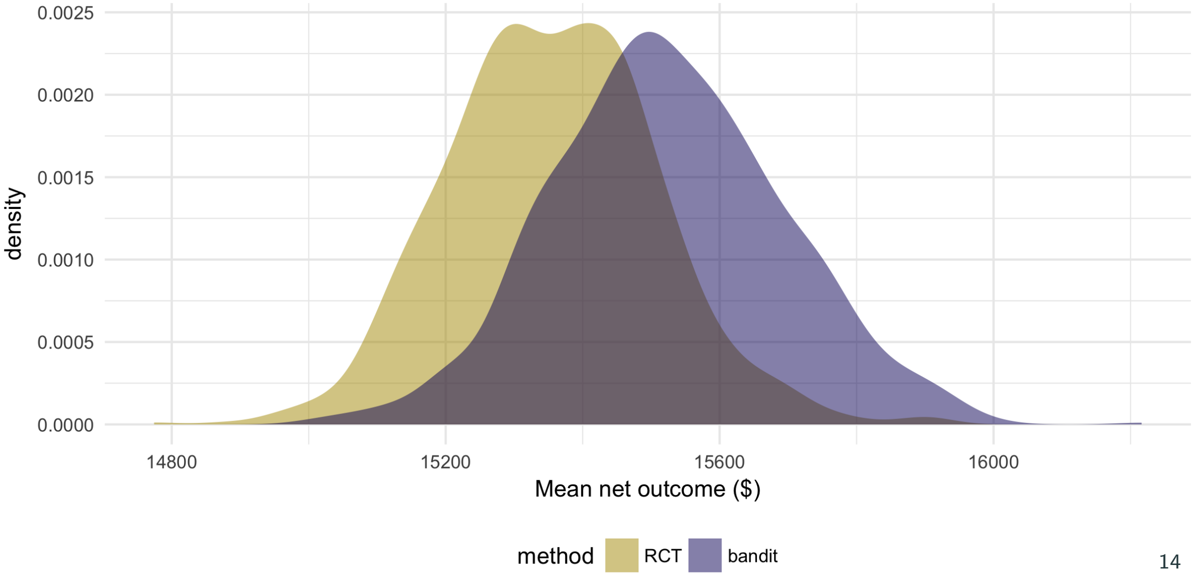
JTPA study

	Assignment		
	Treatment	Control	All
Number of participants	7,487	3,717	11,204
Share of trainees	64.2%	1.5%	
Mean outcome	\$16,200	\$15,041	\$15,815
ITT			**\$1,159**
Mean net outcome	\$15,703	\$15,029	\$15,480
net ITT			**\$674**

Share of participants assigned to treatment - The bandit is adaptive...



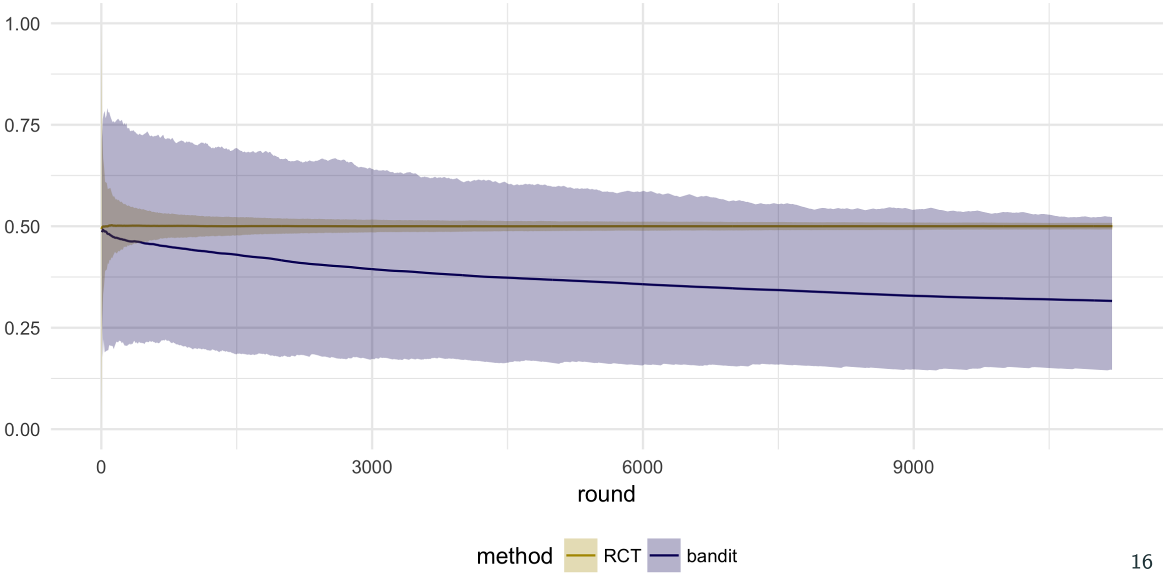
...that results in better expected outcome



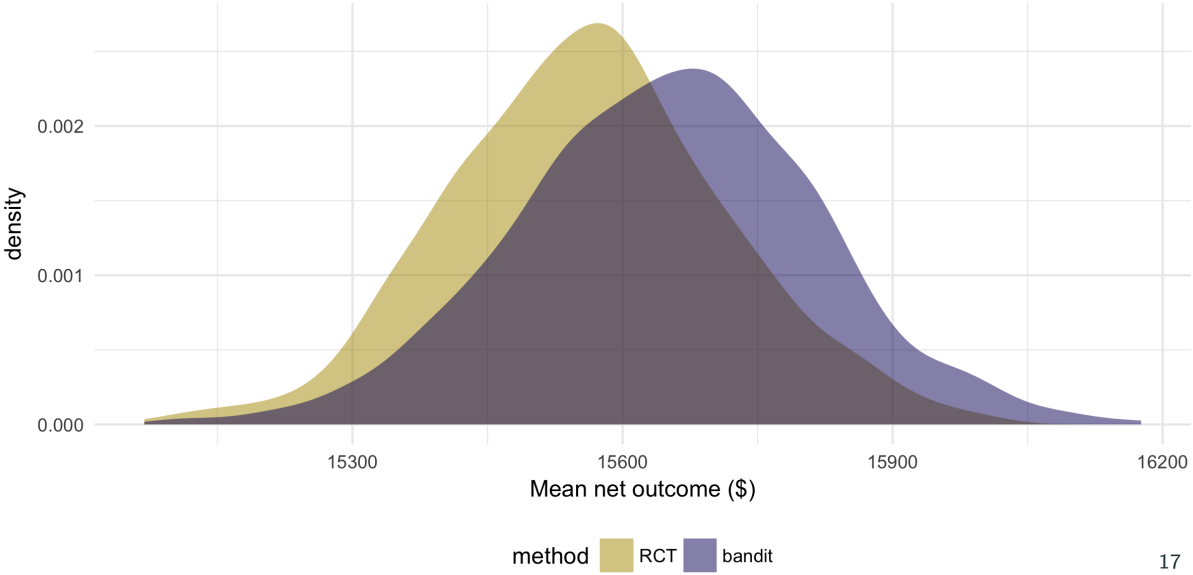
JTPA study - simulated no effect case

	Assignment		
	Treatment	Control	All
Number of participants	7,487	3,717	11,204
Share of trainees	64.2%	1.5%	
Mean outcome	$15,812	$15,821	$15,815
ITT			**-$9**
Mean net outcome	$15,316	$15,810	$15,480
net ITT			**-$495**

Share of participants assigned to treatment - no effect case



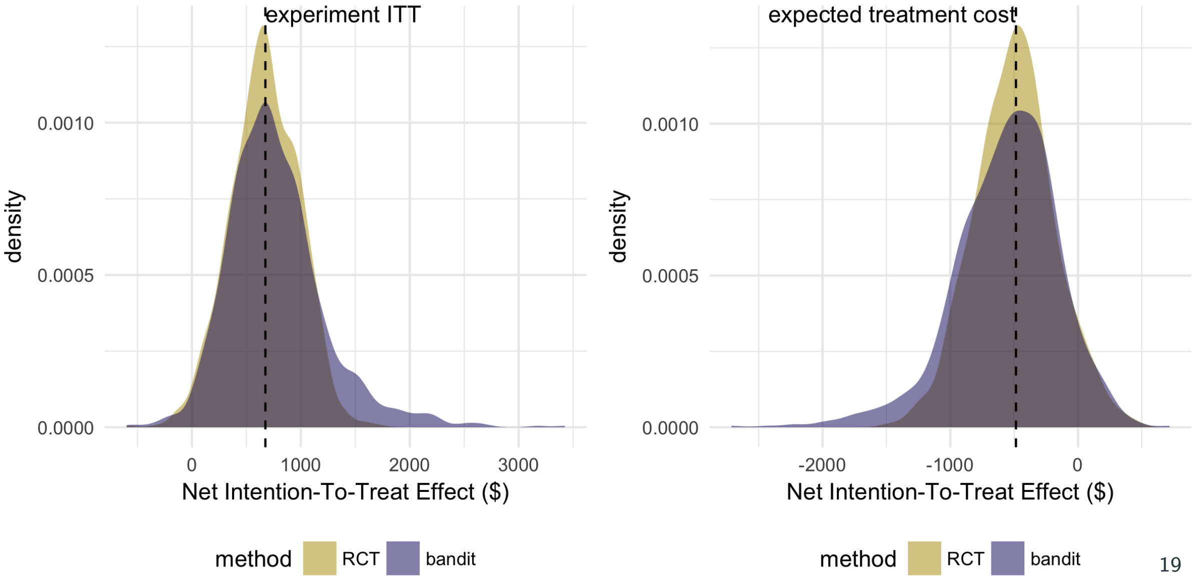
no effect case



Bandit leads to considerable gains in aggregate welfare

	Positive effect	No effect
bandit	\$15,525	\$15,649
RCT	\$15,356	\$15,560
Difference	\$169	\$89
Gain in welfare	**\$1,893,476**	**\$997,156**

The price to pay: the implicit treatment effect estimator is biased



Way ahead

- Understand the behavior of the treatment effect estimator
- Use other prominent experiments for illustration (e.g. from development economics)
- Consider covariates
- Justify bandit algorithm choice: UCB, SE, adaptive SE, etc.

Thank you for your attention