Treatment Choice with Bandit *

Preliminary Draft

János K. Divényi †
Central European University

March 13, 2018

Abstract

In this paper I extend upon the traditional treatment choice literature by introducing sequential choice. There arises a natural trade-off between exploration and exploitation. I borrow the multi-armed bandit methodology from the machine learning literature, and apply it within the Rubin Causal Model. I rephrase the theoretical results of the machine learning literature to show that the traditional approach of randomized control trial (RCT) is not optimal. For illustration, I take a well-known and economically relevant example (JTPA study), and show that taking into account the exploration-exploitation trade-off, one can win almost $2 million relative to RCT. The price to pay is the loss of an unbiased estimator for the treatment effect.

*I thank Gábor Kézdi and Róbert Lieli for advice. Jenő Pál and Sándor Sóvágó provided helpful comments. The data and replication files are available upon request.

†divenyi_janos@phd.ceu.edu

1 Problem

My focus is on the problem of treatment choice. There is a treatment with possibly many variations whose effects on the individuals are unknown. Individuals arrive gradually, and a policy-maker has to decide who should get which treatment based on some observables and an objective function.

I would like to extend upon the growing literature of treatment choice (e.g. Manski, 2004; Kitagawa and Tetenov, 2017) by considering sequential choice: instead of assuming a given (randomized) sample, the assignment itself is also a control variable. The policy-maker has to decide about a treatment rule by considering the trade-off between learning about the response function (exploration) and applying the best possible rule (exploitation). The exploration-exploitation trade-off is widely researched in computer science (see the multi-armed bandit literature, e.g. in Bubeck and Cesa-Bianchi, 2012), but has not yet been applied to the potential outcome framework of Rubin Causal Model (see e.g. in Imbens and Wooldridge, 2009).

The traditional approach and a widespread practice in economics is to run a randomized controlled trial until a sufficient sample size is collected, and use its result to create a welfare-maximizing treatment rule to apply afterwards. I argue that one should take the decision about experimentation into account as well. Assigning units to likely inferior treatments during the experiment harms the outcome. Experimentation has cost (direct and opportunity cost). The sooner one gets the right treatment, the higher the aggregate welfare will be. Considering these aspects, and balancing between exploration and exploitation, one can get higher expected welfare.

Relate to literature:

- personalized medicine: e.g. (Qian and Murphy, 2011) - sequential observations and assignments for the same individual
- treatment choice: (Manski, 2004; Dehejia, 2005; Kitagawa and Tetenov, 2017) - there is an experimental sample given a priori
- dynamic treatment rules: (Perchet et al., 2016; Kock and Thyrsgaard, 2018) - most adjacent, but without the focus on Rubin's model of causality, no thoughts about the estimation of treatment effect

2 General setup

There is a multi-valued treatment, and $D_i \in \mathcal{D} = \{0, 1, ..., S\}$ denotes the treatment assignment of unit i ($D_i = 0$ stands for the no-treatment case, i.e. the status qou). $X_i \in \mathcal{X} \subset \mathbb{R}^{d_x}$ are observable pre-treatment characteristics. $\{Y_i(s)\}$ are potential outcomes that would have been observed if i's treatment status were $D_i = s$ (potential outcomes might include the cost of

the corresponding treatment). The population is thus characterized by P, a joint distribution of $(X_i, Y_i(0), Y_i(1), ..., Y_i(S))$.

Individuals arrive independently in a gradual way: (1) observation i is drawn from P, (2) the policy-maker chooses $D_i = s$, and (3) the outcome $Y_i(s)$ is realized. Then, the next observation is drawn, and so on. The total number of arriving individuals is denoted by n.

The objective of the policy-maker is to maximize the total welfare on the arriving individuals. For simplicity, I assume a utilitarian welfare function, so the problem collapses to maximizing the sum of the individual outcomes. The policy-maker wants to choose a conditional treatment rule $\pi(X) : \mathcal{X} \to \mathcal{D}$ to decide which unit should get which treatment. Formally, the policy-maker's problem is to find $\pi(X)$ that maximizes the expected welfare:

$$\max \mathbb{E} \sum_i Y_i(\pi(X_i))$$

This objective is equivalent to minimizing the expected welfare loss (regret) relative to the maximum feasible welfare. Let's denote the best possible treatment rule by $\pi^*(X_i)$, such that $Y_i(\pi^*(X_i)) = \max Y_i(d)$ for each i. The expected regret after treating n individuals is

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

The policy-maker's goal is to find a rule that minimizes the expected regret. It is common in the literature to use minimax optimization: find a rule that minimizes the worst-case regret. Intuitively, the policy-maker would like to choose a rule that behaves uniformly well across all states of nature (Manski, 2004).

The main difference of my setup to the recent literature on treatment choice is that there is no experimental sample a priori, that is excluded from the optimization problem and can be used to learn about the treatment response function. We start with no information about the treatment response function, and aim for a rule that maximizes the expected welfare.

3 Theoretical results for a simple case

For simplicity, assume a binary treatment $D_i \in \{0, 1\}$ and omit pre-treatment variables ($d_x = 0$). In this case, the optimal treatment rule is constant: either treat everyone ($\pi = 1$) or treat no one ($\pi = 0$). The "treat everyone" rule is optimal if and only if $\mathbb{E}[Y_i(1)] > \mathbb{E}[Y_i(0)]$, that is the average treatment effect is positive: $ATE = \mathbb{E}[Y_i(1) - Y_i(0)] = \tau > 0$. However, without any information, we do not know which rule to choose. We should explore first the outcomes with and without treatment to be able to decide.

Let's assume – without loss of generality – that the treatment effect is positive, so the best choice would be to treat everyone ($Y_i(\pi^*) = Y_i(1)$ for each i). Then the expected regret after treating n individuals is

$$R_n = n\mathbb{E}\left[Y_i(1)\right] - \sum_{i=1}^{n} \mathbb{E}\left[Y_i(\pi_i)\right]$$

Denoting the number of individuals assigned to the (inferior) no-treatment case by n_0 (that could be a random variable depending on the policy π), the regret can be also written as follows:

$$\begin{aligned} R_n &= n\mathbb{E}\left[Y_i(1)\right] - \mathbb{E}\left[(n - n_0)\right]\mathbb{E}\left[Y_i(1)\right] - \mathbb{E}\left[n_0\right]\mathbb{E}\left[Y_i(0)\right] \\ &= \mathbb{E}\left[n_0\right]\mathbb{E}\left[(Y_i(1) - Y_i(0))\right] \\ &= \mathbb{E}\left[n_0\right]\tau \end{aligned}$$

Traditional rule (Randomized Controlled Trial).

1. Choose a sample size of m.
2. Assign the first m individuals with p probability to the treatment (typically, $p = 0.5$).
3. After m individuals, test[1] whether the average treatment effect is positive.
4. If the effect is positive, apply the treatment to each individual onwards.

The rule can be written formally as (with $k \in 0, 1$)

$$\pi_i = \begin{cases} 1 & \text{if } i \leq m \text{ with probability } p \\ 0 & \text{if } i \leq m \text{ with probability } 1 - p \\ \arg\max_k \bar{Y}(k) & \text{if } i > m \end{cases}$$

To calculate the expected regret of this rule, we have to have a sense about how well the averages calculated after m observations could reveal the better outcome. In other words, we have to know how far the average could be from the expected value. Let's recall some important results from the statistics literature of concentration inequalities (see e.g. Rigollet and Hütter, 2017).

Definition. *(sub-Guassianity) A random variable X is said to be sub-Gaussian[2] with variance proxy σ^2 if $\mathbb{E}[X] = 0$ and its moment generating function satisfies*

[1] Typically, it means a standard hypothesis test. For simplicity, I consider just the simple comparison of means. This is in line with the *Conditional Empirical Success rule of Manski (2004).*

[2] Intuitively, sub-Gaussianity means that the corresponding distribution has at most as thick tails as the Gaussian distribution with variance σ^2. It can be shown that any bounded zero mean random variable is sub-Gaussian with a variance proxy that depends on the size of its support.

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right), \textit{ for all } \lambda \in \mathbb{R}$$

Theorem 1. *If $X_1, X_2, ..., X_n$ are independent sub-Gaussian random variables with variance proxy σ^2, their sum is also sub-Gaussian with variance proxy σ^2/n.*

Lemma. *If X is sub-Gaussian with variance proxy σ^2, then for any $\varepsilon > 0$*

$$\mathbb{P}(X > \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right) \textit{ and } \mathbb{P}(X < \varepsilon) \geq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

Proof. Follows by applying the *Chernoff* bound. □

Let us assume that $Y_i(k) - \mathbb{E}[Y(k)]$ are independent, sub-Gaussian random variables with variance proxy σ^2, and denote their average after n observation as $\bar{Y}_n(k) = \frac{1}{n}\sum Y_i(k)$. Using the previous results, we can derive bounds for the concentration of the averages as

$$\mathbb{P}\left(\left|\bar{Y}_n(k) - \mathbb{E}[Y(k)]\right| > \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) \tag{1}$$

for any $\varepsilon > 0$.

The traditional rule suffers $(1-p)\cdot m \cdot \tau$ regret in expected value during the experimentation. For individuals $j > m$ the regret is positive only if the average of the better outcome (treatment) was lower than the average of the worst outcome (no treatment) from the first m individuals. So the expected regret can be written as

$$R_n = \tau\left((1-p)\cdot m + \mathbb{P}\left(\bar{Y}_{(1-p)\cdot m}(0) > \bar{Y}_{p\cdot m}(1)\right)(n-m)\right)$$

After a bit restructuring, substituting $p = 0.5$, and applying Theorem 1 and Equation 1, we can write

$$\begin{aligned}
\mathbb{P}\left(\bar{Y}_{m/2}(0) > \bar{Y}_{m/2}(1)\right) &= \mathbb{P}\left(\bar{Y}_{m/2}(0) - \bar{Y}_{m/2}(1) > 0\right) \\
&= \mathbb{P}\left(\bar{Y}_{m/2}(0) - \mathbb{E}[Y_0] - (\bar{Y}_{m/2}(1) - \mathbb{E}[Y_1]) > \tau\right) \\
&\leq \exp\left(-\frac{m\tau^2}{8\sigma^2}\right)
\end{aligned}$$

So we can bound the expected regret of the traditional RCT rule by

$$R_n \leq \tau\left(\frac{m}{2} + \exp\left(-\frac{m\tau^2}{8\sigma^2}\right)(n-m)\right) \tag{2}$$

This is a worst-case bound, showing an upper limit of the expected regret. It grows linearly with n (if the experimental sample size, m, is fixed)[3].

Bandit rule. (Upper Confidence Bound)

1. Estimate $Y(1)$ and $Y(0)$ by the corresponding average outcomes up until the actual round.
2. Form suitable confidence bounds around the averages that can be derived from Equation 1.
3. Assign to program for which the upper confidence bound is the higher.

Intuitively, we will 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).

The multi-armed bandit literature (started with Robbins, 1952) provides a wide variety of different rules that each solves the exploration-exploitation trade-off differently. I chose the Upper Confidence Bound (UCB) as it is one of the most popular algorithms. Recent papers related to economics use another method, called successive elimination[4] (see Perchet and Rigollet, 2013; Kock and Thyrsgaard, 2018). For now, the exact method does not really matter until it features the required properties.

The UCB algorithm was introduced by Lai and Robbins (1985) who showed that it achieves a regret bound of order $log(n)$, given that the potential outcomes ("rewards") are i.i.d. They also proved that the policy is minimax optimal, that is no other policy can be better asymptotically.

4 Illustration

4.1 Setup

For illustration, I use the well known National Job Training Partnership Act (JTPA) study (Bloom et al., 1997). I take the experimental sample that was used by the influential paper of Abadie et al. (2002). This sample has been used many times for illustrative purposes in the treatment choice literature (see among others Kitagawa and Tetenov, 2017). Participants of the JTPA study assigned to the treatment group were offered job training.

Table 1 shows the main numbers of the experiment. The program seems to be effective. The average earnings of the treatment group is $ higher, even though only 64% of them actually got the training. This shows a positive intention-to-treat effect (ITT), that is my main interest here focusing on treatment assignment rules. The positive ITT more than compensates for the actual cost of the treatment.

[3] If m can be chosen depending on n, one can reach a regret that satisfies $R_n = O\left(n^{2/3}\right)$ but not better, see Szepesvári and Lattimore (2018).

[4] Successive elimination makes it easier to account for further observable information, in the form of covariates.

Table 1: Descriptive statistics of JTPA experiment

	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**

Mean outcome is calculated as 30 month earnings. Mean net outcome accounts for the occasional cost of training ($774, borrowed from Bloom et al., 1997).

This particular study shows a positive effect, however, it could have been the case that the job training just did not help. This is the point of the experiment: to measure whether it makes sense to apply a treatment. To illustrate the other typical state of the world, I simulate a "no effect" scenario as well. Here, I just randomly shuffle the earnings across the individuals. Table 2 summarizes the key numbers for this case. Obviously, the number of participants and share of trainees are the same. However, the mean outcomes are close to each other, giving a close-to-zero ITT. As the treatment helps nothing, the net ITT is negative, as we have to pay the cost for nothing. The net ITT is the expected cost of the assignment: the cost times the increased probability of having to pay it.[5]

Table 2: Descriptive statistics of JTPA experiment - 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**

Mean outcome is calculated as 30 month earnings. Mean net outcome accounts for the occasional cost of training ($774, borrowed from Bloom et al., 1997).

[5] $774 * (0.642 - 0.015) \sim 495$)

For my case, I assume sequential arrival: participants are assigned to the treatment/control group, and their main outcome (earnings in next 30 months) can be observed before the next participant arrives. This assumption is clearly unrealistic, but serves for illustrative purposes. Many times we run experiments that show (obviously less extreme) sequential feature: we can already observe the outcome of some participants when still assigning new ones. We can also imagine a scenario in which such programs (e.g. how to help unemployed) run very long.

For the illustration, I bootstrap the original sample (e.g. sample randomly with replacement). I take the group of participants assigned to the treatment as representative for the distribution of $Y(1)$ and the control group participants as representative for the distribution of $Y(0)$. Random assignment should be enough to provide justification for this method. I can simulate an assignment decision by drawing from the corresponding sample (with replacement). With this procedure, I do not need to assume anything about the nature of the treatment effect (e.g. whether it is homogeneous), but can exploit the data at hand.

I apply the different treatment rules detailed in the previous section on the described scenarios. For RCT, I will use the apriori best fifty-fifty assignment rule. As Tables 1 and 2 show, in the real experiment, more than half of the participants were assigned to the treatment group. I run 1,000 simulation runs for each rule and scenario.

4.2 Results

The main results are depicted in Figures 1 and 2 for the positive effect case, and in Figures 3 and 4 for the no effect case. As the bandit algorithm continuously adjusts the treatment rule based on the previous observations, it learns faster than the RCT. For the positive effect case, it is assigning more and more participants to the treatment, whereas for the no effect case (negative net effect case) it is assigning less and less. RCT does not change its assignment policy throughout the experiment. Learning faster results in higher welfare at the end. The gain in welfare is considerable, amounting to million dollars in expected value (see Table 3).

Table 3: Mean net outcomes in different scenarios

	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**

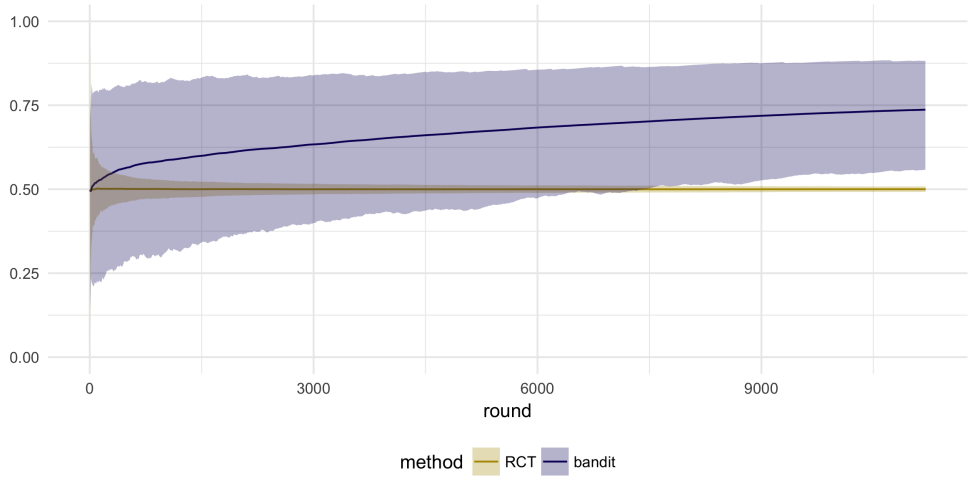

Figure 1: Share of individuals assigned to treatment outcomes.
Every 10th round is plotted, the shaded area shows the corresponding range between the 5% and 95% quantiles. The bandit algorithm gradually learns that the net effect is positive, and assigns more and more participants to the treatment. This results in higher welfare.
Number of simulations = 1,000.

What is the price to pay? Figures 5 and 6 show the estimated intention-to-treat effect along with the observed effect in the original sample. As we can see, the implicit treatment effect estimator of the bandit method is biased agains zero.[6]

The classical treatment choice literature considers the case when someone wants to create a treatment rule based on an existing sample, and apply it for the future. For better comparison, let's simulate into the future. I consider 10,000 additional participants. For RCT, I assign everyone to the group whose average outcome were higher after the experiment (CES rule of Manski, 2004; Kitagawa and Tetenov, 2017)). For the bandit, I proceed with the algorithm as the experiment did not end (that is the point of the bandit). Table 4 summarizes the result.

5 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.

[6]This could be an interesting point. What is the exact reason of losing the unbiasedness? It should be related to the "peeking" error of A/B tests. Is there a way to adjust the result somehow to get an unbiased estimate? What is the asymptotic distribution of the bandit treatment effect estimator?

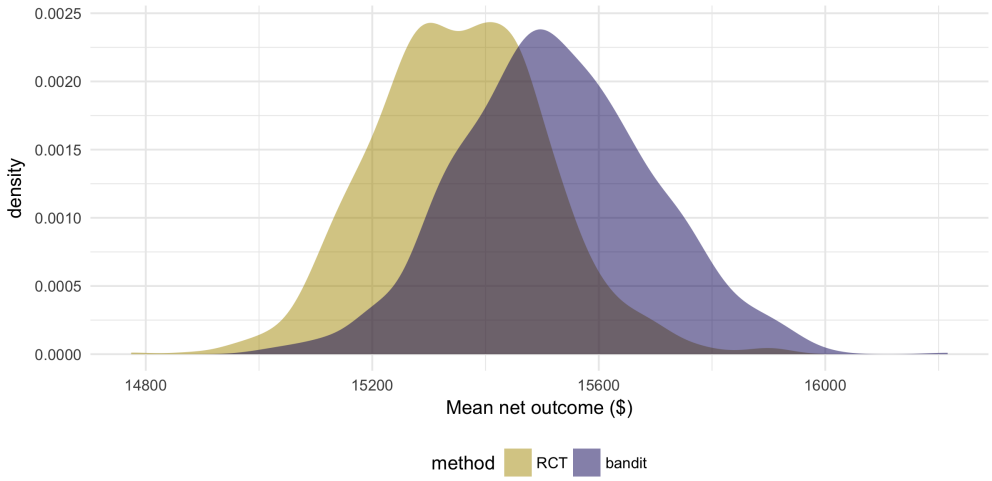

Figure 2: Mean outcome by individual.
With the bandit algorithm, we win about $169 on each individual, following from the higher probability of assigning them to the treatment. This is about the quarter of the net intention-to-treat effect.
Number of simulations = 1,000.

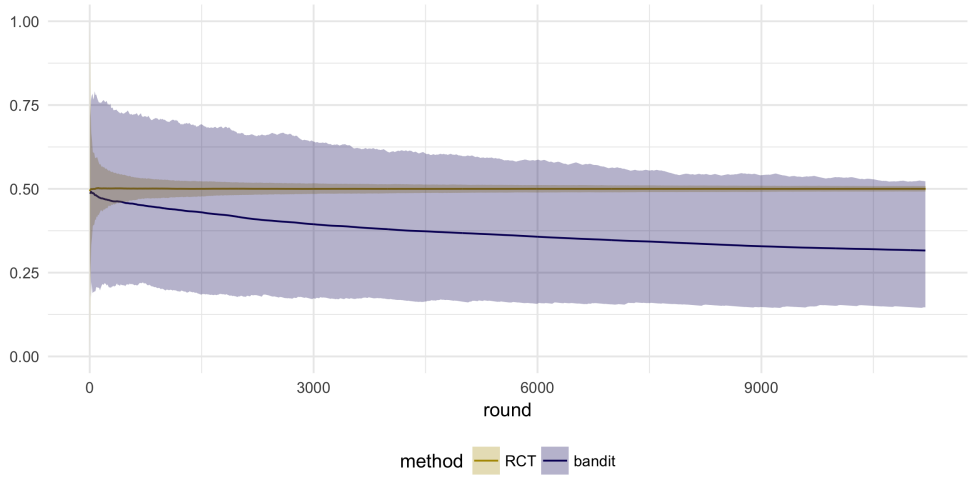

Figure 3: Share of individuals assigned to treatment outcomes - No effect case.
Every 10th round is plotted, the shaded area shows the corresponding range between the 5% and 95% quantiles. The bandit algorithm gradually learns that the net effect is negative, and assigns less and less participants to the treatment. This results in higher welfare.
Number of simulations = 1,000.

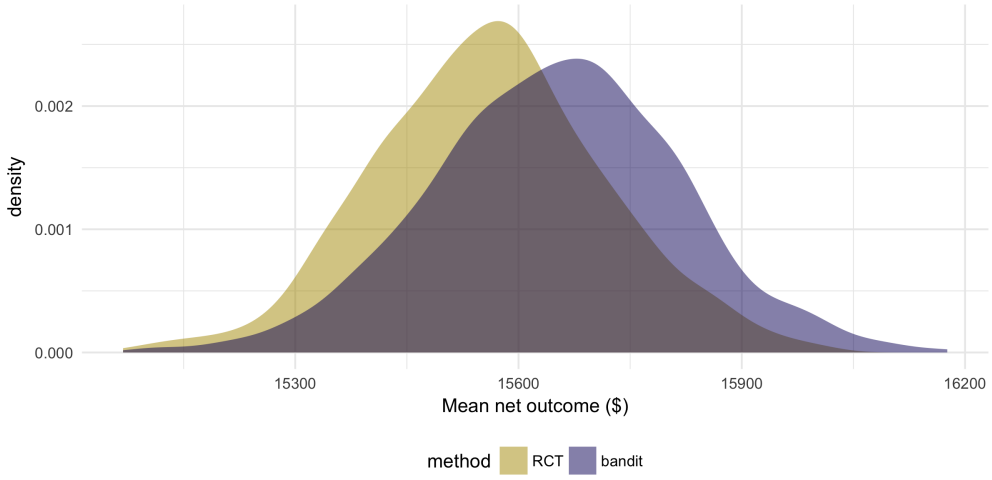

Figure 4: Mean outcome by individual - No effect case.
With the bandit algorithm, we win about $89 on each individual, following from the lower probability of assigning them to the treatment. This is about 18% of the absolute value of the net intention-to-treat effect.
Number of simulations = 1,000.

Table 4: Mean net outcomes in different scenarios, accounting for the future

	Positive effect		No effect	
	bandit	RCT	bandit	RCT
Mean outcome	**$15,560**	$15,511	$15,667	**$15,682**
Minimum outcome	**$15,065**	$14,943	**$15,190**	$15,075

Larger values are in bold. When considering a future run, RCT beats the bandit in expected value for the no effect case. For most of the simulation runs, it realizes that the treatment is not worth it, and stops assigning anyone to the treatment. In contrast, the bandit adopts only gradually. However, for some runs, the decision made by RCT is wrong: it assigns everyone to the wrong treatment. That has a large cost. The bandit cannot err that much because of the gradual adoption, so it beats RCT in the minimax sense for both scenarios.

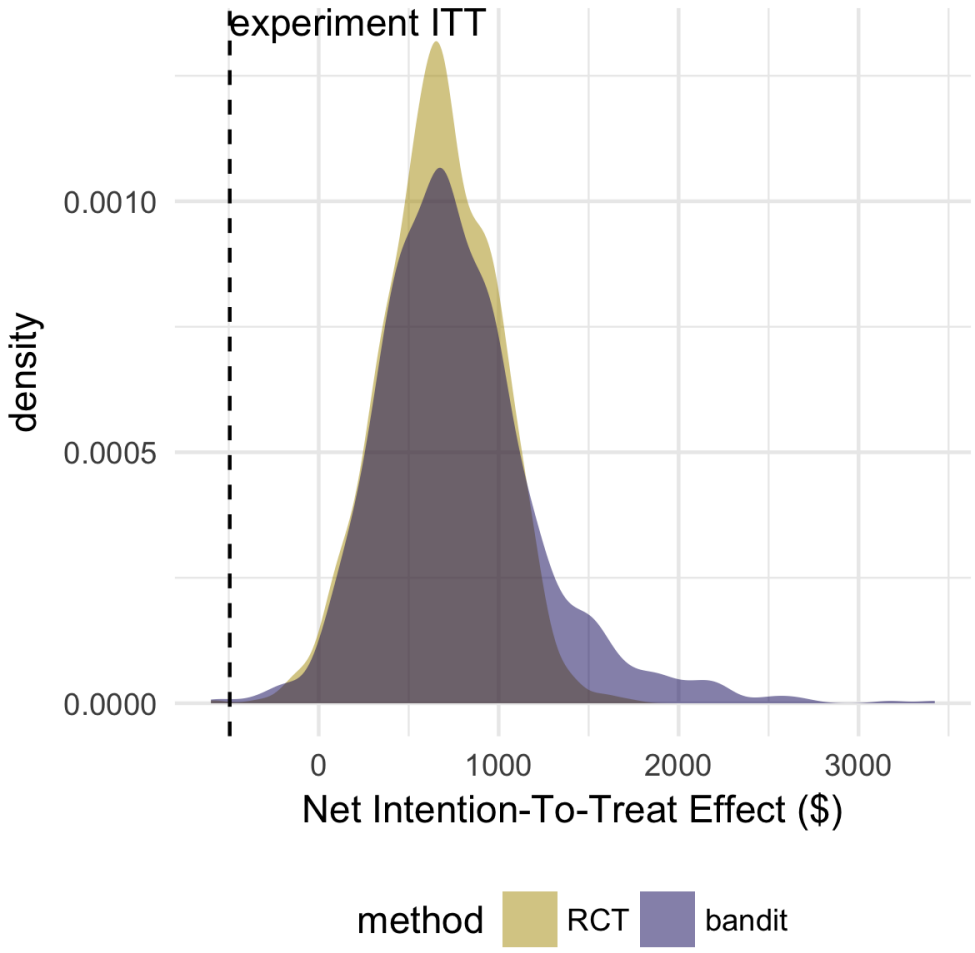

Figure 5: Estimated treatment effect.
The dashed line shows the average net intention-to-treat effect of the actual experiment ($674).
The bandit algorithm overestimates the effect by about 14% (mean of bandit net ITT: $771).
Number of simulations = 1,000.

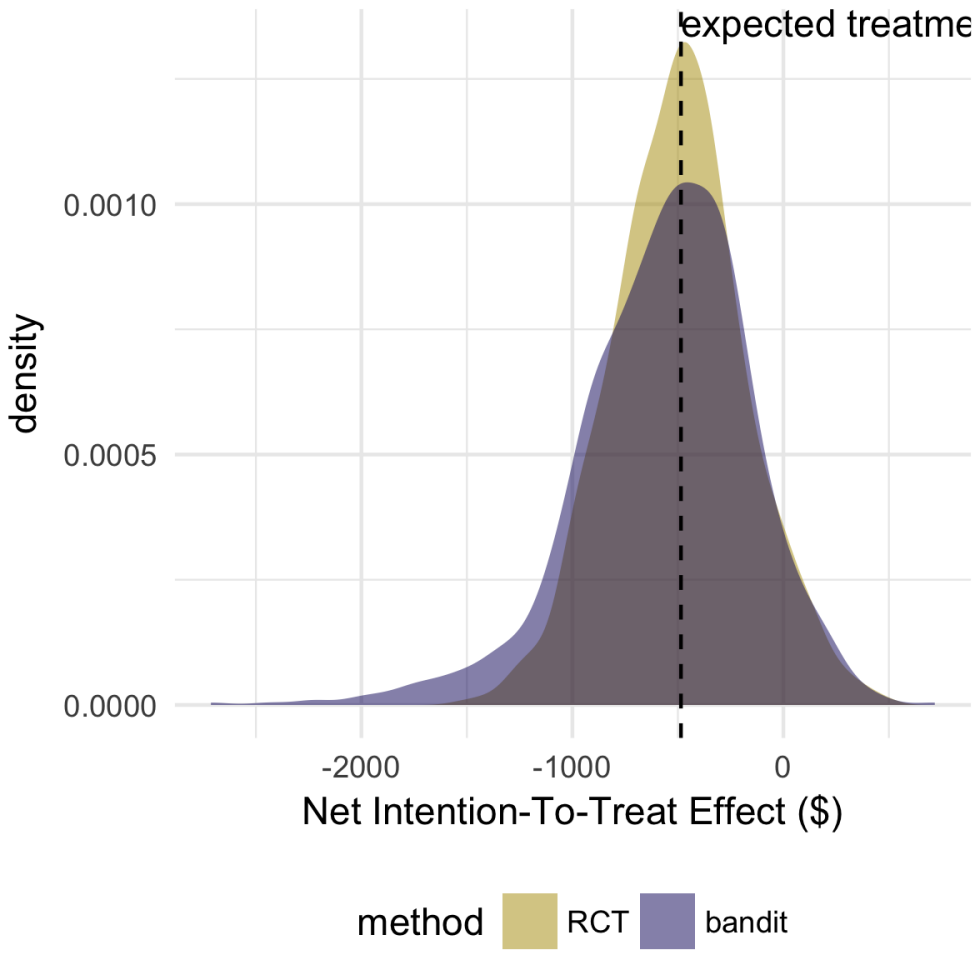

Figure 6: Estimated treatment effect - No effect case.
The dashed line shows the average net intent-to-treat effect of the actual experiment, that is the cost of the treatment times the increased probability of having to pay it (-$774 * 0.625 = -$484).
The bandit algorithm overestimates the effect by about 13% (mean of bandit net ITT: -$561)
Number of simulations = 1,000.

References

Abadie, Alberto, Joshua Angrist, and Guido Imbens, "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 2002, *70* (1), 91–117.

Bloom, Howard S., Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos, "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study," *The Journal of Human Resources*, 1997, *32* (3), 549–576.

Bubeck, Sébastien and Nicolò Cesa-Bianchi, "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems," *Foundations and Trends in Machine Learning*, 2012, *5* (1), 1–122.

Dehejia, Rajeev H., "Program evaluation as a decision problem," *Journal of Econometrics*, 2005, *125* (1-2 SPEC. ISS.), 141–173.

Imbens, Guido W and Jeffrey M Wooldridge, "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 2009, *47* (1), 5–86.

Kitagawa, Toru and Aleksey Tetenov, "Who should be treated? Empirical welfare maximization methods for treatment choice," 2017.

Kock, Anders Bredahl and Martin Thyrsgaard, "Optimal dynamic treatment allocation," 2018.

Lai, T. L. and Herbert Robbins, "Asymptotically Efficient Adaptive Allocation Rules," *Advances in Applied Mathematics*, 1985, *6* (1), 4–22.

Manski, Charles F., "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 2004, *72* (4), 1221–1246.

Perchet, Vianney and Philippe Rigollet, "The multi-armed bandit problem with covariates," *Annals of Statistics*, 2013, *41* (2), 693–721.

_ , _ , Sylvain Chassang, and Erik Snowberg, "Batched bandit problems," *Annals of Statistics*, 2016, *44* (2), 660–681.

Qian, Min and Susan A Murphy, "Performance guarantees for individualized treatment rules," *The Annals of Statistics*, 2011, *39* (2), 1180–1210.

Rigollet, Philippe and Jan-Christian Hütter, "High dimensional statistics," *MIT course notes*, 2017.

Robbins, Herbert, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, 1952, *58* (5), 527–536.

Szepesvári, Csaba and Tor Lattimore, "Bandit Algorithms," *banditalgs.com*, 2018.