Eliminating Bias in Treatment Effect Estimation Arising from Adaptively Collected Data *

Preliminary Draft

János K. Divényi †
Central European University

December 20, 2019

Abstract

It has been demonstrated that bandit algorithms that collect data adaptively - balancing between exploration and exploitation - can achieve higher average outcomes than the "experiment first, exploit later" approach of the traditional treatment choice literature. However, there is little work on how data arising from such algorithms can be used to estimate treatment effects. This paper contributes to this growing literature in three ways. First, a systematic simulation exercise characterizes the behavior of the standard average treatment effect estimator on adaptively collected data: I show that the treatment effect estimation - that results from two negatively biased means - suffers from amplification bias, and illustrate that this bias increases in noise and adaptivity. I also show that the traditional correction method of inverse propensity score weighting can even exacerbate this bias. Second, I suggest an easy-to-implement bias correction method: limiting the adaptivity of the data collection by requiring sampling from all arms results in an unbiased IPW estimate. Lastly, I demonstrate a trade-off between two natural goals: maximizing the expected welfare and having a good estimate for the treatment effect. I show that my correction method extends the set of choices regarding this trade-off, yielding higher expected welfare while allowing for an unbiased and relatively precise estimate.

*I thank Róbert Lieli for advice. Gábor Békés, Marc Kaufmann, Dániel Kehl, Gábor Kézdi, Sergey Lychagin, Miklós Koren, Arieda Muço, Jenő Pál, Sándor Sóvágó and participants of the Brown Bag Seminar at Central European University, Big Data Econometrics Seminar at the University of Aix-Marseilles, and the summer workshop of the Hungarian Society of Economics for doctoral students provided helpful comments. The simulation code is available on GitHub.

†divenyi_janos@phd.ceu.edu

1 Introduction

We are often interested in whether an innovative treatment should be introduced and applied for individuals arriving in succession. Suppose an online shop wants to change its pricing scheme. They can experiment with a new scheme introducing it to part of their daily visitors, with the ultimate goal of applying the better scheme as soon as possible to maximize their profit. Once they change to the new scheme, they also want to know how much value they can hope from it for their next year's budget, i.e. they also want to measure the treatment effect.

This problem is ubiquitous today. Innovation is crucial to survival. We want to apply the procedure that yields the best expected outcome according to our current knowledge (status quo) but we also want to experiment with new ideas that might yield even higher outcome (exploitation versus exploration, earning versus learning). We are also interested in learning what to expect from introducing an innovation.

My paper is at the intersection of the traditional treatment choice literature of econometrics and the growing literature on multi-armed bandits of machine learning. I show that in situations similar to the example above, the profit maximizing or welfare goal and the estimation goal go against each other: strategies that aim for the former lose on the latter. I illustrate this trade-off by a systematic Monte Carlo study using established solutions (such as randomized controlled trial) for these kinds of problems and provide intuitive explanations for the observed patterns. I suggest a new method that extends the policy-maker's choice in the welfare-estimation space.

The standard procedure in economics to decide on the introduction of a new pricing scheme is to first learn its effect, and then to introduce it if the effect is positive. The traditional treatment choice literature (e.g. Manski 2004, Dehejia 2005, Hirano and Porter 2009, Kitagawa and Tetenov 2018, Athey and Wager 2019) assumes that an experimental sample with randomized assignment exists and derives the welfare-maximizing policy rule given the information that can be learnt on the previously collected data. The welfare of the experimental subjects is disregarded. However, in practice, exploration and exploitation do not naturally separate. The decision-maker always decides (sometimes unconsciously) whether it is worth experimenting or simply applying the best practice.

Multi-armed bandit algorithms (for comprehensive reviews see , e.g., Lattimore and Szepesvári 2019, Slivkins 2019) seek to optimize the exploration-exploitation trade-off suggesting heuristic rules that "learn and earn" in parallel. Instead of aiming for a one-off decision, they involve a sequence of decisions where each decision balances between experimenting and exploiting. As such, it is suitable for situations where the feedback is quick (as in our pricing scheme example). The goal is to maximize the expected welfare during the whole process, including the experimentation phase. Bandit algorithms continuously balance between choosing

the treatment arm with the highest expected payoff (exploitation) and choosing treatment arms that are not yet known well (exploration) – the result of each decision contributes to later decisions. There is a quickly evolving literature (in the field of computer science) that investigates different algorithms in different setups and prove their optimality by various criteria. As algorithms aim to find the arm with the highest expected reward (or finding the better pricing scheme), measuring the exact effect of the various arms relative to a baseline is not part of the problem considered.

In this paper, I study a well-known bandit heuristic, *Thompson sampling*, suggested by Thompson (1933). I chose this method because it is one of the most well-known algorithms, it is widely used in the industry (see e.g. Graepel et al. 2010, Scott 2010) and it is a probabilistic rule that has some appealing features I am going to rely on later. However, the focus is not on the specific heuristic, but on the basic features of adaptively collected data when used for statistical inference. All of my results should extend to other popular heuristics that are deterministic, such as the Upper Confidence Bound algorithm (see e.g. Lai and Robbins 1985).

I run Monte Carlo simulations to characterize the behavior of established strategies for the following treatment choice problem: There are two treatments (status quo and innovation, control and treatment) and individuals arriving in groups (batches) should be assigned to one of them. The policy-maker has two goals: to maximize welfare (profit) on arriving individuals and to measure the treatment effect. The potential outcomes are Gaussian, and the individual-level treatment effect is fixed but its magnitude (relative to the variation in the potential outcomes) is *ex ante* unknown. The length of the process (total number of arriving individuals, also called as "horizon") is finite but also unknown.

My simulation results highlight an important trade-off: strategies that result in good estimators (such as the randomized controlled trial during the whole process) suffer from a huge opportunity cost (resulting from assigning too many individuals to the inferior treatment) – whereas the bandit algorithm that optimizes for welfare leads to a biased treatment effect estimate. The bias is positive in absolute value (amplification bias) and increases in the relative size of the treatment effect and in the speed of adaptivity of the algorithm (smaller batches). The traditional bias correction method of inverse propensity weighting (IPW) does not work – in fact, it can even exacerbate the bias. Some straight-forward solutions form transitions between the two extremes (e.g. explore-then-commit strategy), so they provide good choices for policy-makers who have both welfare and estimation goals in their minds. Finally, I suggest an easy-to-implement bias correction method: limiting the adaptivity of the data collection by requiring sampling from all arms. Using inverse propensity weighting on data that arise from limited adaptivity results in an unbiased treatment effect estimate, whereas it preserves almost all of the welfare gain stemming from adaptivity. I show that limiting extends the set of choices regarding

the welfare-estimation trade-off relative to the established strategies.

The performance of bandit algorithms in a stochastic context are measured by their expected reward (total welfare) relative to the reward gained by the best possible policy (which is usually infeasible). The difference between these two measures is the expected (cumulative) regret. We can easily calculate it if the distributions of the potential outcomes and the length of the horizon are known but typically this is not the case. Therefore, each bandit is characterized by their worst-case regret (within a given set of environments). In my setup, I assume the potential outcomes are Gaussian with unit variance but we do not know the expected values, nor the total number of individuals. The worst-case regret of any algorithm is the expected regret achieved by the bandit given the worst possible parameters for the environment.

The seminal paper of Lai and Robbins (1985) proved that in most environments any bandit algorithm should pull suboptimal arms at least $o(\log(n))$ times in the long run. The resulting regret features as an asymptotic lower bound on regret. A bandit algorithm is called optimal, if it reaches this asymptotic (minimax) lower bound. Note that minimax optimality is not a property of the policy alone, but a property of the policy together with set of environments and horizon.

Korda et al. (2013) prove that Thompson sampling on Gaussian rewards with known variance is asymptotically optimal. However, it does not mean that we cannot improve on it in a given setting (in my case, for a given treatment effect, and a given horizon). I extend the existing theoretical results on the performance of Thompson sampling by simulating its expected behavior in various setups.

The standard optimality results refer to setups where individuals arrive one-by-one. I assume that individuals arrive in groups (or batches) to better simulate real-world problems. Perchet et al. (2016) show that unless the number of batches is really small (much smaller than the logarithm of the number of individuals), batches do not form a constraint regarding the existing optimality results for the bandits.

A new line of research focuses on optimal experimentation design where the goal is to learn the treatment effect (see Kasy (2016) for one-off experiments, and Hahn et al. (2011) for adaptive experiments). Another deals with adaptive treatment assignment where the goal is to choose among a set of policies for large-scale implementation (Kasy and Sautmann 2019). The latter is especially close to my work, especially regarding the setup (treatment assignment in an experiment with several batches) but there is a major difference: these works assume away the welfare of the experimental subjects and only focus on learning. I consider both welfare and estimation under adaptive treatment assignment.

Some recent papers deal with estimation concerns arising from adaptively collected data. Nie et al. (2018) prove in theory that the estimated means of the treatment arms suffer from negative bias. They suggest a complex modification of the data collection process to eliminate

bias. However, they deal with arm's means, not with treatment effect, and they do not investigate the behavior of the bias in various setups.

Villar et al. (2015) compare various bandit algorithms in terms of outcome and also estimation performance in a simulated clinical trial. They focus on different algorithms, not on the general mechanism of the adaptive data collection. They do not deal with any bias-correction.

Dimakopoulou et al. (2018) look at so called contextual bandits that include observable variables in the algorithms to capture heterogeneity in the treatment effect. They focus on bias in treatment effect originating from imbalances in the observables, not the general characteristic of the standard treatment effect estimator even if the effect itself is constant.

To my knowledge, this is the first paper that compares different strategies in the welfare-estimation space. I also contribute to the field in two other ways: with the systematic simulation exercise that characterizes the behavior of adaptive data collected in various setups, and with my suggested strategy of "limiting" which extends the choice set of the policy-maker in the welfare-estimation space.

The paper is structured as follows. Section 2 gives a formal setup for the problem. Section 3 illustrates the basic mechanisms on simulating a simple setup. Section 4 introduces my easy-to-implement bias correction method, the limiting. Section 5 and 6 detail the results of the systematic Monte Carlo study, altering the parameters for uncertainty and length of horizon, respectively. Section 7 concludes.

2 General setup

There is a set of n individuals indexed by $i \in \{1, ..., n\}$ whose outcome Y is of interest. There is a binary treatment $W_i \in \{0, 1\}$ where $W_i = 0$ stands for the no-treatment case, i.e. the status quo. $\{Y_i(1), Y_i(0)\}$ are potential outcomes that would have been observed for individual i with or without the treatment (potential outcomes might include the cost of the corresponding treatment). The actual (observed) outcome is $Y_i = Y_i(1)W_i + Y_i(0)(1 - W_i)$. Let us denote the expected value of the potential outcomes by $\mu_w = \mathbb{E}[Y_i(w)]$, for $w \in \{0, 1\}$. The individual-level treatment effect is fixed, i.e. $Y_i(1) = Y_i(0) + \tau$ for each i where τ denotes the average treatment effect: $\tau = \mu_1 - \mu_0$.

Individuals arrive randomly in m equal-sized batches $(B_1, B_2, ..., B_m)$. The batch size is denoted by n_B so $mn_B = n$. Arrival is sequential and the outcome is observed right after the assignment. The process can be described as follows:

1. A group of individuals $i \in B_j$ arrive, and are assigned to either treatment or control.
2. Outcomes $\{Y_i\}_{i \in B_j}$ are observed.
3. A next group of individuals $i \in B_{j+1}$ arrive and the first two steps are repeated.

The policy-maker chooses an appropriate assignment rule π that maps each individual to a treatment arm depending on its batch (given the observations from all the previous batches): $\pi : \{1, ..., n\} \to \{0, 1\}$. The policy-maker has two goals: she cares about the total welfare of individuals, and she also wants to estimate the treatment effect τ with an unbiased, precise estimator. I assume a utilitarian welfare function, so the total welfare is measured simply by the sum of outcomes: $\sum_i Y_i$. The estimator is measured by its expected bias and its mean squared error (MSE).

The traditional way to deal with these problems consists of two steps: first, concentrate on the estimation goal and run a Randomized Controlled Trial (RCT) on an experimental sample, and then, focus on the outcome and form a deterministic rule based on the experiment's result that can be applied from then on (subject of the classic treatment choice literature). This process can be translated to my case as the strategy of Explore-then-commit (ETC):

Explore-then-commit (ETC)

1. Choose a sample size n_E for an experiment. Typically, this is done by assuming a minimum size for the treatment effect and calculating a required sample size that yields enough power given a predetermined false positive rate (or significance level).

2. Split the first k batches equally between treatment and control until there are at least n_E individuals assigned, i.e. $(k-1)n_B < n_E$ and $kn_B \geq n_E$.

3. After k batches, estimate the average treatment effect by comparing the treatment and control averages calculated on the collected data[a]:

$$\hat{\tau}^{(k)} = \hat{\mu}_1^{(k)} - \hat{\mu}_0^{(k)} = \frac{\sum_{i \in \bigcup_{j=1}^{k} B_j} Y_i W_i}{\sum_{i \in \bigcup_{j=1}^{k} B_j} W_i} - \frac{\sum_{i \in \bigcup_{j=1}^{k} B_j} Y_i (1 - W_i)}{\sum_{i \in \bigcup_{j=1}^{k} B_j} (1 - W_i)}$$

4. Apply the assignment with the higher mean to everyone onwards:

$$W_i = \arg\max_w \left\{ \hat{\mu}_w^{(k)} \right\} \text{ for } i \in \bigcup_{j=k+1}^{m} B_j$$

[a] Comparing the averages corresponds to the Conditional Empirical Success Rule of Manski (2004).

Bandit algorithms allow for more dynamic adaptivity, blending experimentation with exploitation. The strategy of Thompson Sampling can be described as follows[1]:

[1] For more detail, see Russo et al. (2017)

Thompson Sampling (TS)

1. Split the first batch equally between treatment and control.
2. Form beliefs about the treatment and control means by deriving posterior distributions using normal density with calculated averages (assuming that standard deviation is known)[a]:

$$\mathcal{N}\left(\hat{\mu}_1^{(k)}, \frac{\sigma^2}{n_1^{(k)}}\right) \text{ for treatment, and } \mathcal{N}\left(\hat{\mu}_0^{(k)}, \frac{\sigma^2}{n_0^{(k)}}\right) \text{ for control,}$$

   where

$$n_1^{(k)} = \sum_{i \in \bigcup_{j=1}^{k} B_j} W_i, \; n_0^{(k)} = \sum_{i \in \bigcup_{j=1}^{k} B_j} (1 - W_i).$$

3. Assign individuals to the treatment in the next batch by the probability that the treatment mean is higher than the control mean. Technically, for each individual draw $\tilde{\mu}_{i,w} \sim \mathcal{N}\left(\hat{\mu}_w^{(k)}, \frac{\sigma_w^2}{n_w^{(k)}}\right)$ for $i \in B_{k+1}$ and $w \in \{0, 1\}$, and assign the individual to the treatment of the higher draw

$$W_i = \arg\max_d \{\tilde{\mu}_{w,i}\}$$

4. Repeat from step (2) until assigning the last batch.

[a] This is equivalent to the posterior of mean of a normal variable with known variance using non-informative Jeffreys prior

Intuitively, we will choose the treatment more likely (for a larger fraction of individuals in the batch) if (1) we are uncertain about its expected outcome (exploration), or (2) we are certain that its expected outcome is high (exploitation).

3 The statistical and welfare properties of bandit

3.1 Parametrization

I show the basic mechanisms of the bandit algorithm using a simple setup. I assume – without loss of generality – a positive average treatment effect with unit value ($\tau = 1$). The population consists of $n = 10,000$ individuals with normally distributed potential outcomes ($\sigma = 10$). The noise-to-signal ratio is high to make the treatment effect hard to measure, and thus, the problem interesting. The potential outcomes are constructed such that $\mu_1 = 1$ and $\mu_0 = 0$ within the

population. Individuals arrive randomly in ten batches, each of size 1000.

In this case, the (infeasible) optimal treatment rule is to treat everyone ($\pi = 1$) that would achieve a total welfare of $10,000$. Due to the fact that the treatment effect is normalized and is fixed for everyone, the sum of outcomes equals to the sum of individuals assigned to the treated, so both measures express the total welfare. Without information about the sign of the treatment effect, we do not know a priori whether the "treat everyone" or the "treat no-one" strategy is better. We should explore first the outcomes with and without treatment to be able to decide.

To simulate how the bandit algorithm works I run $20,000$ simulations. The runs differ only in the sequence how the individuals arrive; they all use the same population of 10,000 with the average of potential outcomes equaling to 0 and 1, respectively.

3.2 Adaptivity, welfare, bias

Data collection with the bandit algorithm is adaptive: after every batch we decide how to split the next one between treatment and control. As the treatment effect is positive, TS assigns an increasing number of people to the treatment on average as it is learning the effect on the incoming batches (see Figure 1). Because of the highly volatile outcome, individual Monte Carlo runs do not necessarily follow this trend, some runs event start to the wrong direction. However, by the last batch, every run succeeds in learning the positive effect, and most runs achieve the optimal allocation ("treat everyone").

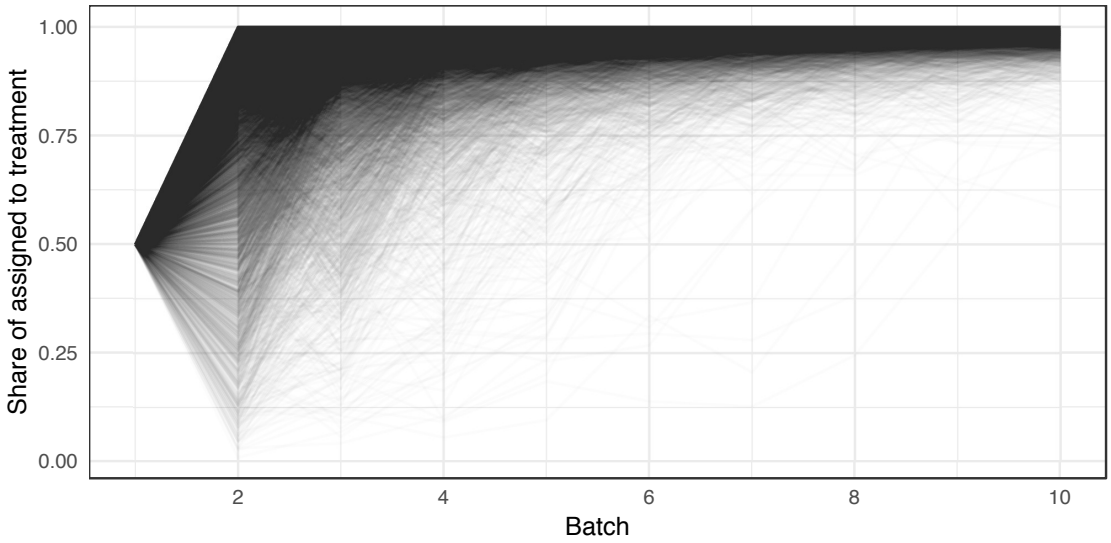


Figure 1: Share of batches assigned to treatment across simulation runs. The bandit is adaptive: the share of treated is increasing on average and tend to the optimal allocation ("treat everyone"). Number of simulations = 20,000.

Thanks to the adaptivity that assigns more and more people to the treatment with better

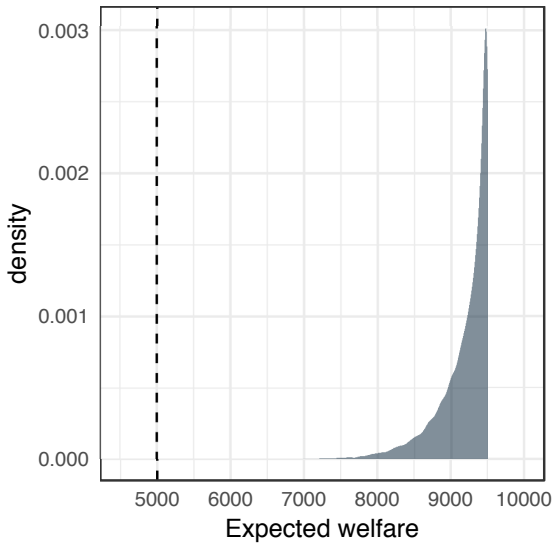
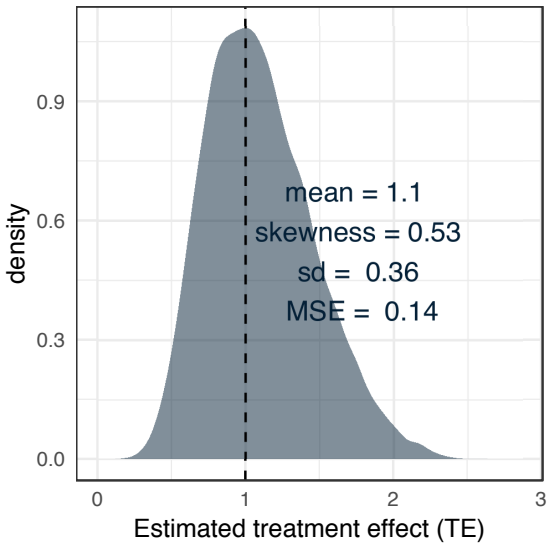


Figure 2: Distribution of welfare across simulation runs. The dashed line shows the expected welfare of a random allocation. The bandit achieves close to optimal welfare with a high probability but even the fat tail towards zero clearly outperforms the random split. Number of simulations = 20,000.

outcome (relative to the random $50-50\%$ split in the first batch), bandit results in a higher total welfare than what is expected from a random split (RCT). Figure 2 shows the distribution of welfare across simulation runs. Due to the first batch, there is a natural upper limit on welfare that the algorithm could achieve: if the algorithm correctly learns the positive treatment effect in the first batch with full certainty, everyone onwards will be assigned to the treatment, and only the random half of the first batch is assigned to the inferior, zero-outcome control. In this setup, we have 10 batches, so the natural upper limit is $9.5n = 9500$. The expected welfare distribution has a long tail to the left showing the runs that get lost at start. Each run beats the random allocation confidently.

The bandit algorithm performs well regarding the outcome goal. However, there is a price to pay. The treatment effect itself is harder to measure. The standard estimator that uses the difference in the overall averages is biased on adaptively collected data (see the left panel of Figure 3): we overestimate the real effect by 10%. This bias is due to magnitude errors, as none of the runs results in a negative estimated treatment effect (which is in line with Figure 1 that shows that every run ends up assigning most people in batch to the treatment). In addition to the bias, the distribution of the estimator has other unfavorable properties: it is highly volatile (the standard deviation is almost 20 times larger than what we would expect from a totally random allocation), and it is positively skewed so the standard hypothesis tests also fail.

The treatment effect estimator is a result of two estimated means. Looking at the treatment



	Control	Treatment
Mean	-0.10	1.00
Skewness	-0.63	-1.49
Std. dev.	0.33	0.03
MSE	0.12	0.00

Figure 3: Distribution of estimated treatment effect. The estimator is biased, mainly because of the biased estimate of the control mean. Number of simulations = 20,000.

and control averages (see the right panel of Figure 3), one can see that the bias originates from a negatively biased control mean estimate. The treatment mean estimate is approximately unbiased but it is also negatively skewed similarly to the control estimate. This result is in line with Nie et al. (2018) who prove that adaptive data collection – under natural conditions – leads to negatively biased sample means. For our setup, the bias in treatment mean is negligibly small (around 0.2%). The positive bias and the positive skewness of the treatment effect estimator is a result of the negatively biased and negatively skewed estimators of the group means where the bias in the control mean is dominating.

The bias in group means results from an asymmetry in sampling that is an inherent feature of the adaptive data collection. For the sake of an intuitive understanding of this process, let us focus only on the control estimate where the bias is larger. As the first batch is a simple random split, the first batch average is an unbiased estimate for the control mean ($\mathbb{E}[\hat{\mu}_1^{(1)}] = \mu_1$). However, the actual estimate contains some estimation error ($\hat{\mu}_1^{(1)} = \mu_1 + \varepsilon_1^{(1)}$). If this error is negative ($\varepsilon_1^{(1)} < 0$), there will be a positive error in the treatment effect estimate. As a result, the bandit's belief will be distorted towards the treatment being effective, so more individuals will be assigned to the treatment and only a few to the control. Few new observations in the control group cannot compensate for the original error in the control estimate. However, if the error in the first batch is positive ($\varepsilon_1^{(1)} > 0$), the belief will be distorted towards the treatment being ineffective, so more individuals will be assigned to control, and these new observations can outweigh the original error in the control estimate.

Figure 4 provides a visual illustration for this mechanism. If the first batch results in a negative

control estimate, this error is more likely to remain there also in the overall estimate of the experiment, than in the case when the first batch results in a positive control estimate.

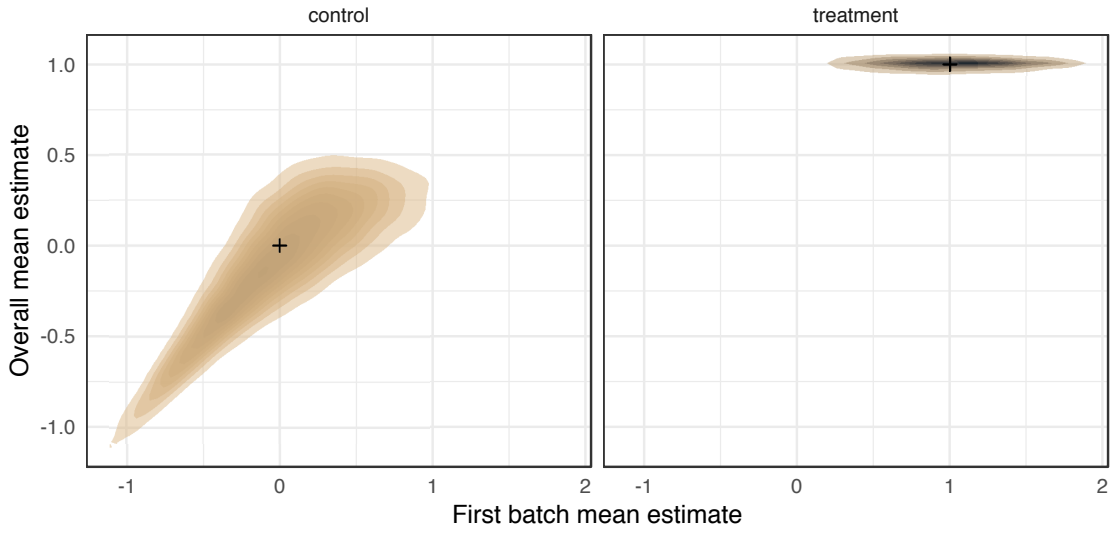


Figure 4: Estimated means using only the first batch versus the estimated means on the whole sample (density: darker regions mean higher density). The importance of the first batch estimate is clear, especially for the control outcome: an underestimated group mean from the first batch remains uncompensated in the overall estimate. Number of simulations = 20,000.

Note that this asymmetry by the estimation error is not restricted to the first versus later batches but is present throughout the whole process. It is only most visible after the first batch as the first round of assignment does not depend on previous observations.

The asymmetry can be highlighted using a simple decomposition: The treatment effect estimate being the difference of two mean estimates can be expressed as the difference of linear combinations of corresponding group batch averages. The treatment and control averages can be calculated as weighted averages of the batch group averages where the weights are the shares of the given batch within the total size of the given group (see Equation 1). The batch group estimates are unbiased as they arise from simple random splits of batches (only the way how the split is done changes but it does not matter for the sake of unbiasedness). The bias in the overall averages results only from compositional effect: as a negative error in the estimate of a given batch leads to under-sampling in the following batches, it means lower weights for these batches, thus, a relatively higher weight to the given erroneous batch. In contrast, a positive error leads to over-sampling in the following batches, which gives a relatively lower weight for the erroneous batch. Also, over-sampling in the next batch quickly leads to the correction of the error, thus the over-sampling itself remains only a temporary issue.

$$
\begin{aligned}
\hat{\tau} &= \frac{\sum_{i=1}^{n} Y_i W_i}{\sum_{i=1}^{n} W_i} - \frac{\sum_{i=1}^{n} Y_i (1 - W_i)}{\sum_{i=1}^{n} (1 - W_i)} \\
&= \sum_{j=1}^{m} \underbrace{\frac{\sum_{i \in B_j} Y_i W_i}{\sum_{i \in B_j} W_i}}_{\substack{\text{batch treated} \\ \text{average}}} \underbrace{\frac{\sum_{i \in B_j} W_i}{\sum_{i=1}^{n} W_i}}_{\substack{\text{share of batch} \\ \text{within all treated}}} - \sum_{j=1}^{m} \underbrace{\frac{\sum_{i \in B_j} Y_i (1 - W_i)}{\sum_{i \in B_j} (1 - W_i)}}_{\substack{\text{batch control} \\ \text{average}}} \underbrace{\frac{\sum_{i \in B_j} (1 - W_i)}{\sum_{i=1}^{n} (1 - W_i)}}_{\substack{\text{share of batch} \\ \text{within all control}}}
\end{aligned} \tag{1}
$$

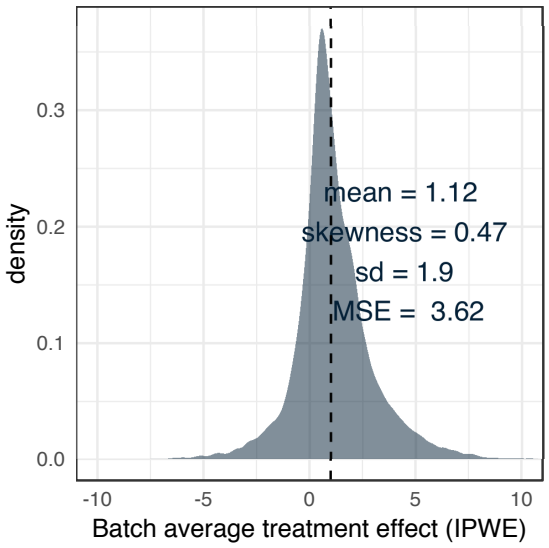
3.3 Inverse Propensity Weighting

Following from the fact that each group average is an unbiased estimate for the corresponding group mean, it seems to be a good idea to take the simple average of the batch averages to estimate the means. This method gets rid of the compositional effect and takes the averages of multiple unbiased estimates. As I prove in Equation 2, this is the same as calculating the Inverse Propensity Weighted Estimator (IPWE). As individuals arrive in batches, propensity scores ($p_i = \mathbb{P}(W_i = 1)$) within batches are constant. This method is mentioned as a possible way for correcting the bias by multiple sources (e.g. Nie et al. 2018, Dimakopoulou et al. 2018).

$$
\begin{aligned}
\hat{\tau}_{IPWE} &= \frac{1}{n} \left(\sum_{i=1}^{n} \frac{Y_i W_i}{p_i} - \sum_{i=1}^{n} \frac{Y_i (1 - W_i)}{1 - p_i} \right) \\
&= \frac{1}{n} \sum_{j=1}^{m} \left(\sum_{i \in B_j} \frac{Y_i W_i n_B}{\sum_{i \in B_j} W_i} - \sum_{i \in B_j} \frac{Y_i (1 - W_i) n_B}{\sum_{i \in B_j} (1 - W_i)} \right) \\
&= \frac{1}{m} \sum_{j=1}^{m} \underbrace{\frac{\sum_{i \in B_j} Y_i W_i}{\sum_{i \in B_j} W_i}}_{\substack{\text{batch treated} \\ \text{average}}} - \frac{1}{m} \sum_{j=1}^{m} \underbrace{\frac{\sum_{i \in B_j} Y_i (1 - W_i)}{\sum_{i \in B_j} (1 - W_i)}}_{\substack{\text{batch control} \\ \text{average}}}
\end{aligned} \tag{2}
$$

However, for my case, it does not seem to be effective (see Figure 5). Instead of eliminating the bias, it even exacerbated the problem. The volatility of the estimator got larger by at least an order of magnitude, resulting in a negative estimate (sign error) in 20% of the runs.

The reason why this method leads to unsatisfactory results lies again in the asymmetry of sampling. Taking the average of averages as explained above should work but only if there are averages available to average on. However, in some cases the bandit might assign everyone to the treatment (recall Figure 1) leaving no control assignees to use for calculating the control batch average. These cases are exactly the ones where the treatment effect is estimated with the highest positive error (hence the extreme assignment share of the treated). See Table 1 for details: runs



	Control	Treatment
Mean	-0.12	1.00
Skewness	-0.44	-1.82
Std. dev.	1.89	0.04
MSE	3.59	0.00

Figure 5: Distribution of Inverse Propensity Weighted Treatment Effect (IPWE). The situation is even worse: larger bias with much higher variance. Number of simulations = 20,000.

with smaller number of batches with any control assignees overestimate the treatment effect more and more. The majority of runs end up with assigning to control in every batch, however, they result in an underestimated treatment effect, a natural consequence of selection.

Table 1: Comparison of IPWE by number of batches with control assignment

# of batches with controls	1	2	3	4	5	6	7	8	9	10
IPWE	1.99	1.85	1.76	1.79	1.71	1.46	1.64	1.32	1.22	0.80
Probability	2.0%	3.2%	3.5%	3.5%	3.8%	4.4%	5.2%	6.8%	11.4%	56.3%

Selection bias: Runs with controls in every batch underestimate the treatment effect while runs with batches without controls overestimate the treatment effect, using the average of averages (IPWE) for estimator. Number of simulations = 20,000.

Figure 6 provides a visual illustration for this phenomenon on the control group. The left panel shows that each batch average in itself is an unbiased estimate for the corresponding control mean. As we tend to sample less and less control in later batches, the estimate is more and more volatile. The right panel shows how the average of averages evolve through batches. If the average of averages after a given batch is small, we tend to sample either less control in the following batch so we update the average with a more volatile average, or no control at all so we do not update the average. This process results in the negatively biased, negatively skewed distribution plotted with the darkest color in the chart.

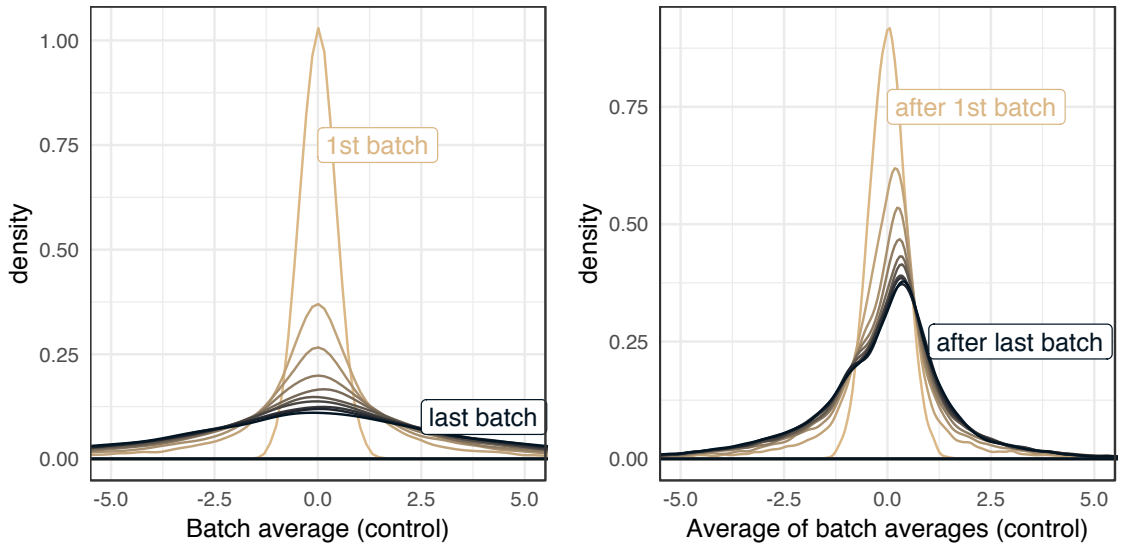


Figure 6: Batch average for the control mean across batches. Each batch in itself is unbiased. Average of batch averages is getting biased due to selection. Number of simulations = 20,000.

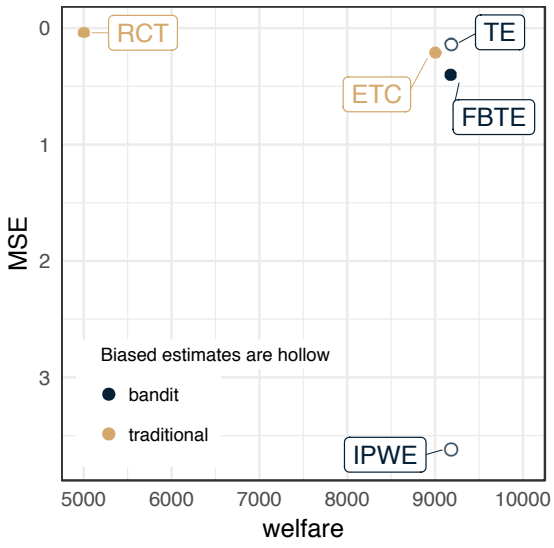
3.4 Welfare-Estimation Trade-off

To have an unbiased estimate, there is a straight-forward shortcut: use only the data collected in the first batch for estimation. As the first batch is a small-scale RCT, the First Batch Treatment Effect (FBTE) estimator is unbiased. However, this strategy loses on efficiency as it drops 90% of the observations.

This trade-off is inherent in this problem. Running an RCT on the whole sample results in the best estimator (that makes RCT as the gold standard for measuring an effect), however, it suffers a high opportunity cost by assigning too many people to inferior treatment.

Figure 7 summarizes this trade-off by comparing the performance of different strategies in this space: the welfare achieved on the individuals (x axis) and the precision of the estimator (measured by mean squared error, y axis). I assume that biased estimates are inferior irrespectively of their precision as the policy-maker can only use an unbiased estimate (in the figure, they are shown as hollow circles for comparison). The best strategy would be a shaded point at the top right corner: with a total welfare of $10,000$ and an unbiased treatment effect estimate with zero MSE. Obviously, such a strategy does not exist.

The above discussed estimators on adaptively collected data (TE, IPWE) gain a lot on welfare relative to RCT, but they give biased estimates making them misleading in some situations. The FBTE loses more on MSE than TE, but it yields an unbiased estimate with the same gain on welfare. ETC is the explore-then-commit strategy using the first batch for exploration (i.e. $n_E = n_B$) and then making a decision for the following batches; its estimator is more precise than the FBTE – in exchange for giving up some of the welfare gain. A policy-maker without an extreme



RCT: random split of the whole sample

ETC: random split of the first batch then treat everyone onwards based on the measured result

TE: Thompson sampling with the standard TE estimator

FBTE: Thompson sampling using only the first batch to calculate the TE estimate

Figure 7: Performance of different strategies in the welfare-estimation space (n_B = 1000). Welfare and estimation goals are working against each other. Number of simulations = 20,000.

marginal rate of substitution between welfare and estimation precision would choose between the ETC or FBTE strategies.

3.5 Choosing batch size

In many cases, the policy-maker has some control over the batch size (n_B). Even if the arrival of individuals is dictated by a natural process, one can still increase the natural batch size by collapsing batches. Therefore, the previously discussed strategies might have multiple variants that use different batch sizes for a given setup. This subsection investigates how batch size affects the welfare and the estimation aspects of the different strategies.

One would expect that smaller batch size (more batches, quicker adaptivity) leads to higher welfare, as it extends the possibilities of the policy maker. To validate this expectation, I simulate the same setup of $10,000$ individuals with 10 different batch sizes[2]. Smaller batch size mean more batches, thus, more frequent allocation decisions. Figure 8 showing the expected welfare by batch size only partially justifies our expectation: generally, smaller batch size leads to higher expected welfare, but focusing on the small batch size region (left panel) reveals that being too "quick" can also do harm; the optimum is at $n_B = 100$. Smaller batch sizes give the chance of reacting more quickly to a positive treatment effect, hence, suffering less opportunity cost. However, it also means deciding based on more volatile estimates, increasing the probability of "getting lost", and

[2]The simulated values are the followings: 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, and 10,000. The maximum value corresponds to a simple random split on the whole sample.

adapting to the wrong pattern.

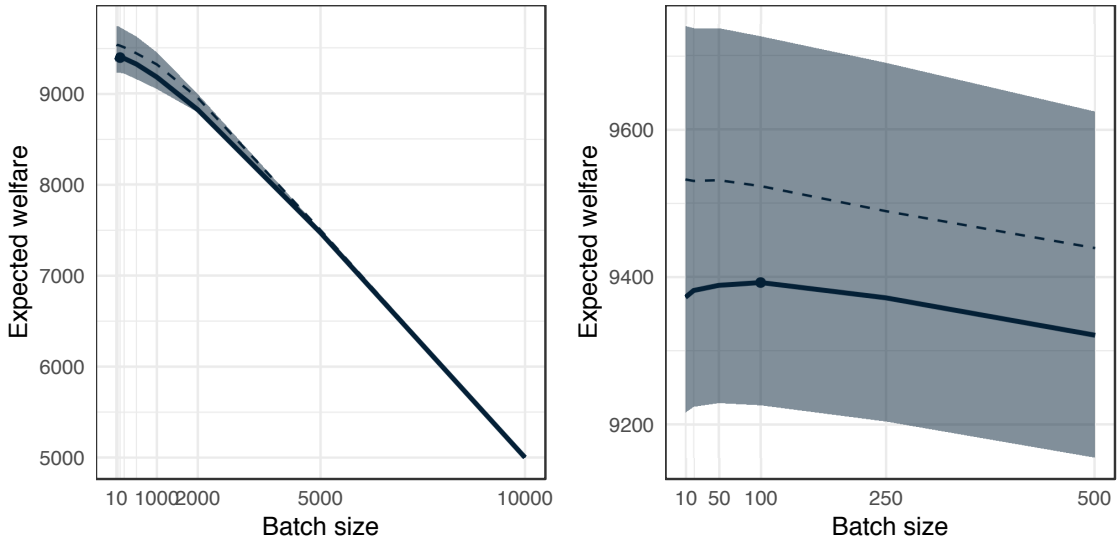


Figure 8: Expected welfare by batch size. The shaded area show the interquartile range, the dashed line is the median. The point depicts the batch size with the maximum welfare. The right panel focuses on small batch sizes. Smaller batches (quicker adaptivity) generally leads to higher welfare, but only after a certain point: really small batch size can harm. Number of simulations = 20,000.

Figure 9 illustrates this phenomenon by showing the probability of under-performing a simple random split in terms of welfare at each point of the process, for different batch sizes. At the beginning, quicker adaptivity allows for smaller opportunity cost as with smaller batch sizes the algorithm can allocate less people to the inferior treatment (recall that the first batch size is a random split). However, quicker adaptivity also means making decisions based on more volatile measures due to smaller sample sizes. These decisions turn out more likely to be false, therefore, the probability of under-performing remains relatively high at the later stages of the process. The welfare result of Figure 8 originates from these two contradicting processes.

The behavior of the batch size parameter let us raise an interesting analogy from the machine learning literature: regularization (see e.g. Hastie et al. 2001) is the process of adding information to prevent over-fitting (e.g. by shrinking coefficients). Regularization leads to higher bias to gain on variance, increasing predictive accuracy. In our case, larger batch size means more regularization: it loses on opportunity cost at the beginning, but wins on generalization in the longer term – especially if the noise is high.

The fact, that for this given setup a constrained algorithm works better than a less constrained one does not contradict to the literature. The Thompson Sampling algorithms is a general solution, working well in different setups whose parameters (mainly τ and n) are ex-ante

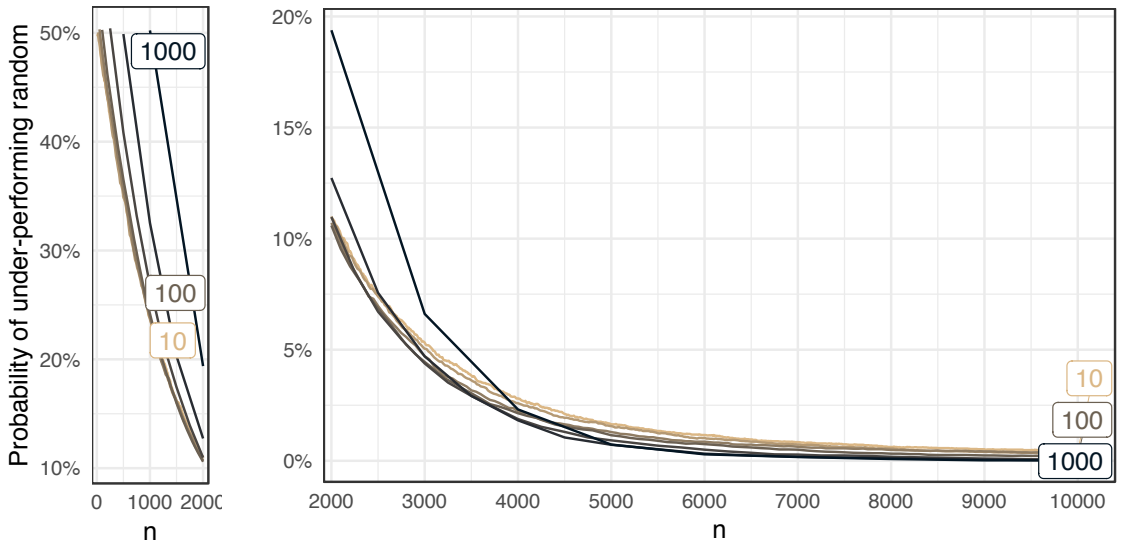


Figure 9: Evolution of bandit algorithms by batch sizes (lighter lines depict smaller batch sizes) across arriving individuals, measured by the probability of under-performing a simple random split at each point of the process. Quicker adaptivity results in smaller opportunity cost at the beginning (left panel), but leads to higher probability of getting wrong at later stages (right panel). Number of simulations = 20,000.

unknown. As we are going to see later, regularization by avoiding too small batches helps only if the noise is high, or equivalently, if the treatment effect is small.

Batch size also affects the performance on estimation. Quicker adaptivity (smaller batch size) leads to higher welfare only at the cost of worse (more biased and more volatile) estimators. This is in accordance with the welfare-estimation trade-off discussed in the previous subsection. IPWE can mitigate the bias for small batches, but generally, the improvement is small and it also means a much higher mean squared error (see Figure A18 in Appendix for details).

We can extend Figure 7 about welfare-estimation trade-off with varying batch size (see Figure 10). Adaptive data collection and using the First Batch Treatment Effect (FBTE) clearly dominates the Explore-then-Commit (ETC) strategy for small batch sizes (batch size $<$ 1000), but it loses somewhat on large batch sizes. IPWE and the standard TE estimator on adaptively collected data are out of play because of the highly biased estimates[3]. I show them on the chart only for comparison. A policy-maker who values the estimation relatively more should go with ETC, but if welfare is also important, FBTE seems to be a better option. Our initial choice of bath size $=$ 1000 seems to be the closest to the optimal for both strategies.

[3]TE can work with really large batches, but that means essentially no adaptivity – as in the case of ETC

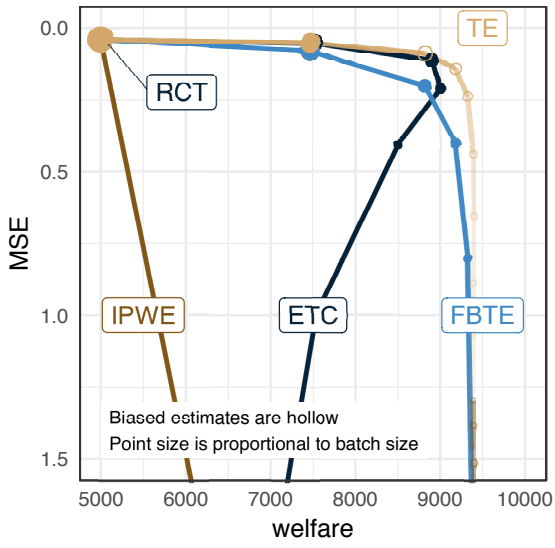


Figure 10: Performance of different strategies in the welfare-estimation space (varying batch size). Generally, quicker adaptivity leads to higher welfare but also larger MSE. ETC with moderate batch size works well, but smaller batch size harms not MSE but also welfare. FBTE approximates the standard TE with higher MSE but ensuring an unbiased estimate. Number of simulations = 20,000.

4 Bias Correction with Limitation

Additionally to the previously discussed strategies I suggest a new one that improves on the IPWE by a slight modification of the data collection process that results in an unbiased estimate. As I showed in section 3.3, the reason why the standard IPWE is biased is that the algorithm does not assign to both groups in each batch, and this unanimous assignment asymmetrically depends on previous observations: If there is a negative error in the average of the control averages after several batches, we are more likely to stop assigning anyone to the control afterwards, thus, end up with an erroneous overall IPW estimator. In contrast, if the error is positive, we are more likely to assign new arrivals to the control, and the later averages can compensate for the early error.

A simple solution for this issue is to ensure that in each batch, people are assigned to both groups, that is to limit the propensity score away from the extremes of zero and one. Although this method needs the modification of the data collection process, in the digital world this is typically not very costly. Also, this solution is easy-to-implement, does not require any complex calculations as the suggestion of Nie et al. (2018).

Limited Thompson Sampling

The difference to the native Thompson Sampling is highlighted in bold.

1. Split the first batch equally between treatment and control.
2. Form beliefs about the treatment and control means by deriving posterior distributions using normal density with calculated averages (assuming that standard deviation is known).
3. Assign individuals to the treatment in the next batch by the probability that the treatment mean is higher than the control mean. **If this probability is too extreme, use a limited probability instead. Denoting the amount of limitation by L, and the probability after the kth batch by $p^{(k)}$, the assigning probability is $\tilde{p}^{(k)} = \max\left(\min\left(p^{(k)}, 1-L\right), L\right)$.**
4. Repeat from step (2) until assigning the last batch.

Following from the discussion of section 3.3, the smallest possible limitation (e.g. 1% for the batch size of 100) would yield an unbiased estimate. The amount of limitation incorporates the welfare-estimation trade-off. Limiting to higher extent requires higher opportunity cost, but also allows for more robust estimates. It forms a smooth transition between two endpoints: the unlimited bandit (0% limit, previously used in TE, IPWE and FBTE strategies) and a random split of the full sample (50% limit, RCT).

Figure 11 shows the effect of limitation on welfare and estimation goals simulating 8 different limit levels[4]. As expected, higher limit means lower welfare and more precise inverse-propensity-weighted estimate[5].

The cost on welfare is linear and is the same for each batch size (nominally, this means a higher cost for lower batch sizes that yield higher welfare with the unlimited strategy). However, the loss in welfare and the gain in precision is disproportionate: while the loss is linear in the amount of limitation, the gain is not: using a 1% limit, MSE drops dramatically for each batch size (by as much as 80% for batch size = 2000 - see right panel) while it costs no more than 1% of welfare (left panel).

It is interesting to note that limitation affects differently the different batch sizes. Small and large batch sizes induce lower cost than the middle range for a given limit. This is the result of two factors: First, limitation acts as a regularization tool, similarly to what we have seen with larger

[4] 0%, 0.5%, 1%, 2%, 5%, 10%, 15% and 20%.

[5] Limitation also decreases the bias of the standard TE estimator, but due to the inherent weighting in Equation 1, some bias remains until the limit reaches the level of the simple random split.

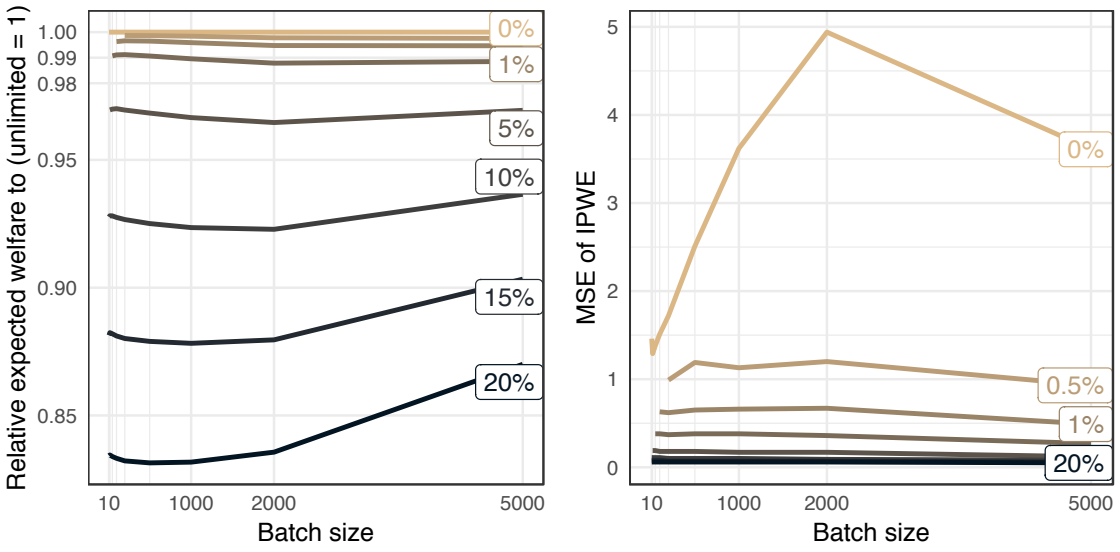


Figure 11: Welfare and estimation performance of limiting by batch size. Higher limits incur higher welfare cost (left panel) but bring more precision (right panel). The loss and gain by the amount of limit are disproportionate. Number of simulations = 20,000.

batch sizes. Limitation decreases the probability of over-fitting, and can thus improve welfare for some runs. Second, limitation obviously does not affect the simple random split of the first batch. For larger batch sizes, the share of the first batch is higher, thus, the limitation cost is relatively lower.

On the other hand, the improvement on the estimation precision is about stable by batch size. This result follows from the fact that limitation is defined as share of the batch, so it means closely the same for each batch size. Higher limitation - in line with approaching the simple random split strategy - also improves the skewness of the estimator and the variance of the reached welfare.

As the estimation improvement does not depend on the batch size, strategies with quicker adaptivity should fare better in the welfare-estimation space. The left panel of Figure 12 shows the performance of different limitation strategies. Lower limitation can achieve higher welfare, but only for an growing cost on MSE. The lines are close to horizontal, showing that lower batch sizes can achieve higher expected welfare for practically no estimation cost. Different points of this chart depict different parametrizations of the limited IPWE strategy; some of them dominate each other (e.g. large batch sizes with low limitation are clearly worse than smaller batch sizes with higher limitation). Connecting the best parametrizations give us the Performance Frontier of this strategy in the welfare-estimation space. Any of these point could be achieved by choosing an appropriate batch size (n_B) and amount of limitation (L).

The right panel of figure shows only the frontier for the limited IPWE strategy, along with our previous strategies. Limitation with inverse-propensity-weighting extends the possibilities of the

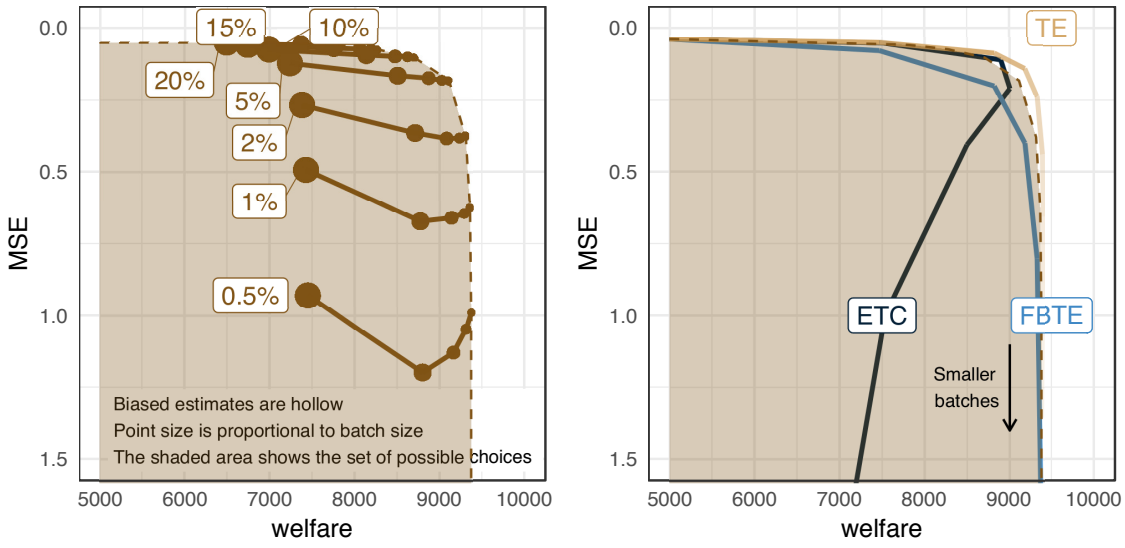


Figure 12: Performance of different strategies in the welfare-estimation space. The right panel shows the outcomes for IPWE with different limits along with the Performance Frontier. The left panel shows only this frontier along with the previous strategies: limited IPWE extends the possibilities by approximating the TE strategy while also ensuring an unbiased estimate. Number of simulations = 20,000.

policy-maker: It gets the closest to the TE strategy with also allowing for an unbiased estimate, and dominates both of our previous best strategies (ETC and FTBE with batch size of 1000). If the policy-maker cares about welfare as well, collecting data adaptively with some limitation and estimating the treatment effect with inverse-propensity-weighting is the best strategy.

5 Monte Carlo Simulation for Uncertainty

5.1 Simulation setup

The previous sections illustrated the mechanisms of adaptive data collection and the welfare-estimation trade-off by using a simple setup of a given treatment effect (equaling to a 10th of the standard deviation of the potential outcomes) and a given horizon (10,000 individuals). Here, I extend these results by conducting a systematic Monte Carlo simulation with different levels uncertainty. I investigate the effect of the size of treatment effect by holding τ fixed at unit value and vary σ. As the important measure regarding learning is the relative effect size τ/σ, it does not matter which one is fixed. Fixing τ allows me to directly compare the welfare and estimation performance of the strategies.

I investigate 7 different values for σ with $n = 10,000$[6]. Each setup is simulated with 10 values

[6] $\sigma \in \{1, 2, 5, 10, 15, 20, 30\}$

of batch size and 8 values of limit[7], 10 − 50 thousand runs for each[8].

5.2 Results on welfare and bias

Figure 13 summarizes the results of the expected total welfare and the bias in the traditional treatment effect estimate by batch size for each σ. Less uncertainty (smaller variation in the potential outcomes) increases the expected gain and decreases the bias. Both of these results are intuitive.

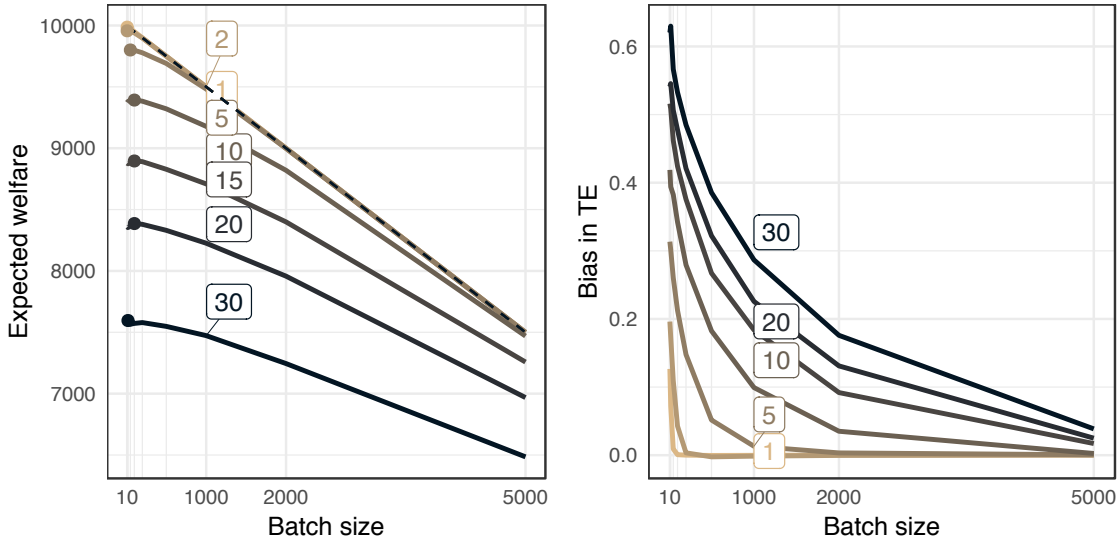


Figure 13: Expected total welfare (left panel) and bias in the traditional treatment effect estimator (right panel) with different batch sizes (on x axis) for different levels of noise (labelled). The dashed line on left shows the maximum welfare that each strategy could achieve, the points show the batch sizes with the maximum welfare for a given σ.
Number of simulations = 10-50,000.

Unlike in the illustrative setup, for low levels of noise the quickest adaptivity results in the highest expected welfare. For these setups, the danger of over-fitting is low, so regularizing by increasing the batch size does not help, only brings a higher opportunity cost.

There is another interesting pattern to note: For welfare, each line approaches the one with the smallest σ, some also reach it. This means that less uncertainty does not lead to higher outcome under a certain value of σ where this value depends on the batch size. The reason for this is that for each batch size there is a maximum of outcome that cannot be exceeded: when the positive treatment effect is learnt immediately in the first batch and all subsequent batches

[7] As small batch sizes do not work with low limits, it means 63 parametrizations for each setup.

[8] The number of runs depends on the level of noise: for setups with larger noise I run more simulations to get robust results: 10,000 for σ below 10, 20,000 for σ at least 10 but below 20 and 50,000 for larger values of σ.

are assigned to the treatment. It is possible if the noise in the outcomes are small relative to the batch size. This maximum possible welfare is depicted by the dashed line on the chart - if the standard deviation in potential outcomes is not larger than the treatment effect, practically each batch size achieves this maximum.

A similar pattern is visible in the bias (right panel) as well: if the noise is sufficiently low and the batch size is large enough, there is no bias. Obviously, if the treatment effect is perfectly learnt in the first batch, the asymptotic sampling resulting in the bias does not kick in. Figure A19 in Appendix shows the average share of treated in the second batch across batch sizes for each setup. It confirms that full learning in first batch can explain the observed patterns in welfare and bias.

Figure 14 shows the bias in the group means for each setup. It lets us observe a previously stated result, that the mean estimate of *both* outcome is negatively biased. For our illustrative setup, it was not true, as the uncertainty was not high enough and the adaptivity was not quick enough. The reason why the treatment outcome estimate needs a higher level of noise to show some bias is intuitive: As the treatment effect is positive, asymmetric sampling leads to more assignments to the treatment, so any error will be compensated by a large number of additional new observations. However, as the uncertainty increases, the Thompson sampling gets more inclined to learn on the wrong pattern and arrive at the false conclusion that the control outcome is larger. In these cases, a negative error in the treatment outcome's estimate could lead to under-sampling treatment later, which results in the same type of bias that we already discussed in section 3.

It is also worth noting, that as both outcome estimate suffer from negative bias, they partially compensate each other in the treatment effect estimate, so the bias in the treatment effect estimator is smaller than the bias in the control mean.

5.3 Welfare-Estimation Trade-off

The results of the previous subsection are in line with the main message of this paper: welfare and estimation goals are working against each other: mainly, quicker adaptivity leads to higher outcome but also higher bias, for each level of σ. This observation only works differently for two special regions: for high levels of noise, extreme adaptivity hurts both goals, whereas for low levels of noise, adaptivity can be increased until a certain point, exploiting the welfare gain but without introducing any bias.

I suggested limiting as a working method for bias correction in section 4. I showed that small amounts of limitation result in unbiased treatment effect estimates with highly improved MSE for only a low price in achieved welfare, and this disproportionality allows for the extension of

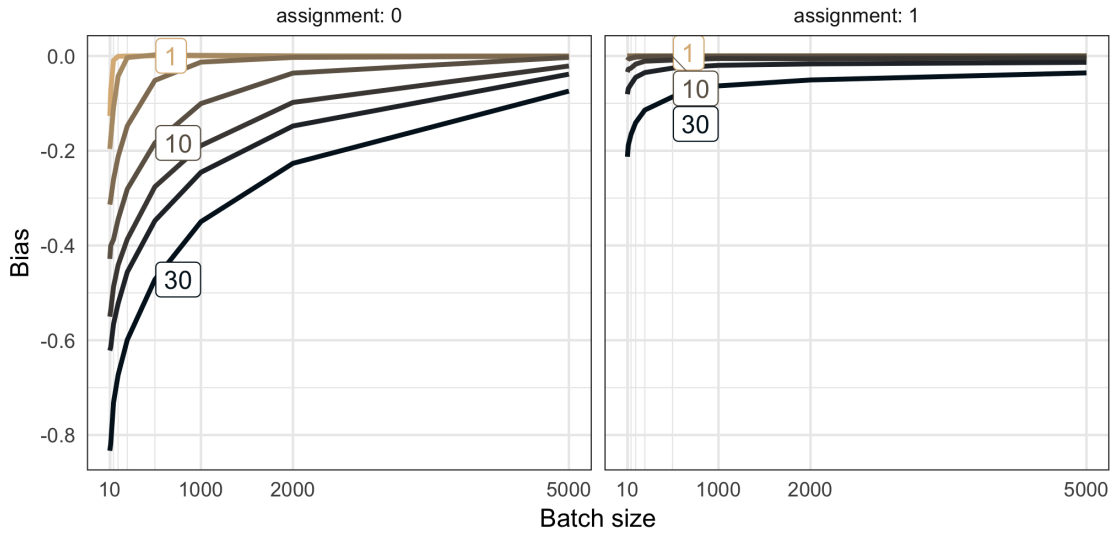


Figure 14: Bias in group means for different levels of uncertainty. Larger level of noise and quicker adaptivity leads to larger negative bias. For low noise scenarios, moderate adaptivity is enough to prevent bias. Number of simulations = 10-50,000.

the set of available choices for the policy-maker in the welfare-estimation space.

Previously, Figure 11 illustrated that limiting more means a choice that favors estimation over welfare: it gives up some of the welfare gain in order to win on precision. Table 2 shows the batch size - limit combinations that achieve the highest welfare for each setup. Obviously, for small noise the best is to limit the least possible (to preserve unbiasedness, quicker adaptivity would mean higher limitation, e.g. for batch size of 10, the smallest possible limitation is 10%). For larger noise, choosing a somewhat higher limit with correspondingly quicker adaptivity could increase the expected welfare.

	$\sigma \leq 15$		$\sigma = 20$		$\sigma = 30$	
	Batch size	Limit	Batch size	Limit	Batch size	Limit
1	200	0.5%	100	1%	100	1%
2	100	1%	100	2%	50	2%
3	200	1%	50	2%	100	2%

Table 2: Combinations of batch size and limit that yield the highest welfare with an unbiased estimate for various levels of uncertainty. Number of simulations = 10-50,000.

Figure 15 shows the performance of the different strategies in the welfare-estimation space for each setup. Similarly to Figure 10, it only shows the frontier for the limited IPWE strategy that is formed by the best combinations of batch size and limit. Obviously, as the problem gets harder (as the uncertainty grows), each strategy performs worse (are farther away from the top right

corner). My previous result is strengthened: adaptivity with limitation almost always extends the feasible set of welfare-MSE pairs. For high noise, my suggested strategy even extends upon the unlimited bandit that were excluded because the estimate is biased. Only in low-noise setups is this extension ambiguous. However, in these setups the problem to solve is easy, and the whole question is of less importance. The treatment effect can be learnt perfectly right in the first batch, so an unlimited bandit could deliver an unbiased estimate next to near-optimal welfare (see Figure 13).

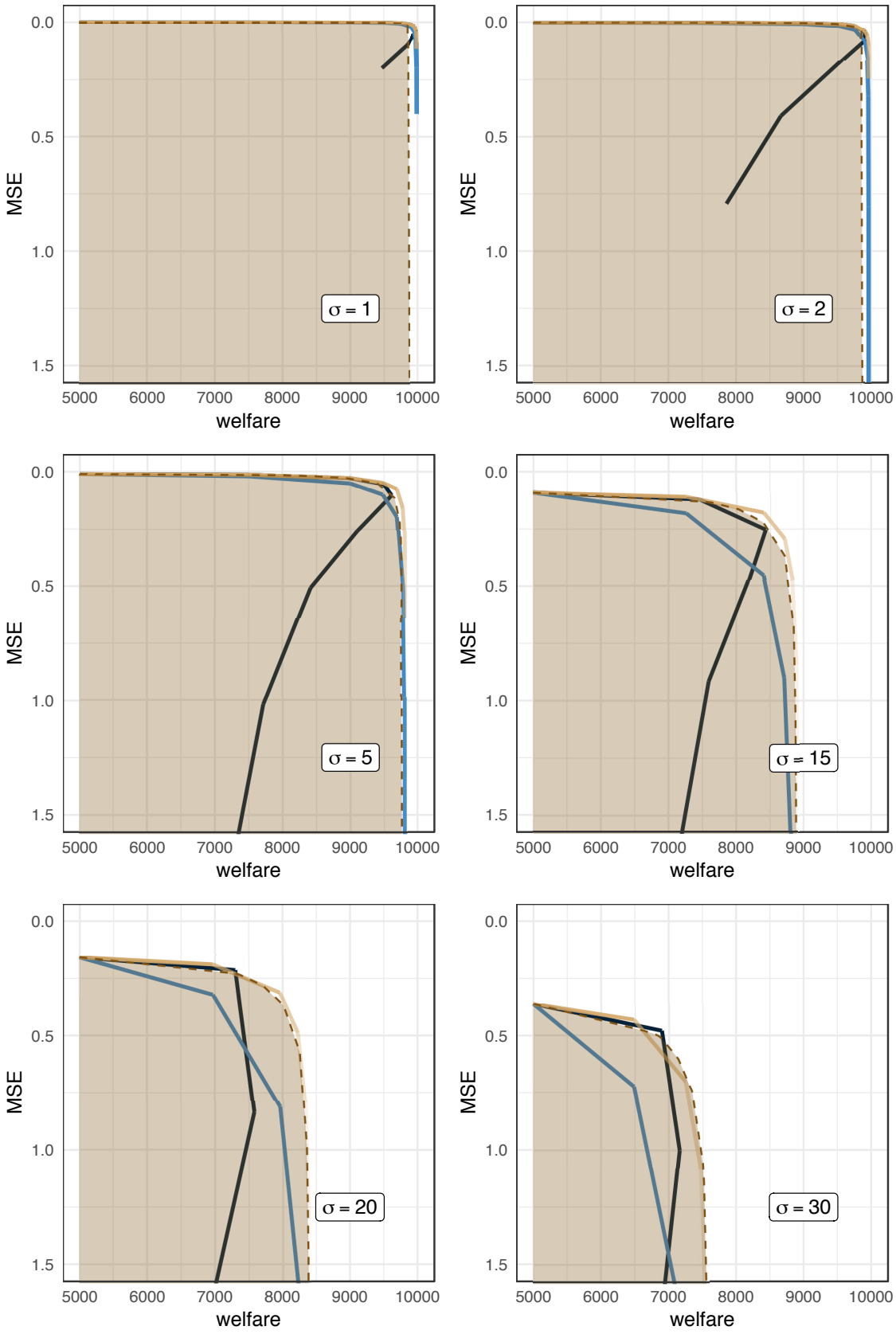


Figure 15: Performance of different strategies in the welfare-estimation space, for different levels of noise. The limited IPWE strategy always extends the set of choices, especially if the problem is hard (the noise is large). Number of simulations = 10-50,000.

6 Monte Carlo Simulation for Horizon

I also consider different lengths for the horizon[9]. Note that this is similar to changing the noise and batch size appropriately: e.g. a 4 times larger sample size is equivalent to a setup with 2 times larger σ with 4 times larger batches (e.g. holding the number of batches fixed). Simulating the illustrative case ($\sigma = 10$) for different lengths makes the comparison easier.

The right panel of Figure 16 validates the theoretical result, that the regret of Thompson sampling with any batch size grows slower than the regret of the exploit-then-commit (ETC) strategy typical in the treatment choice literature.

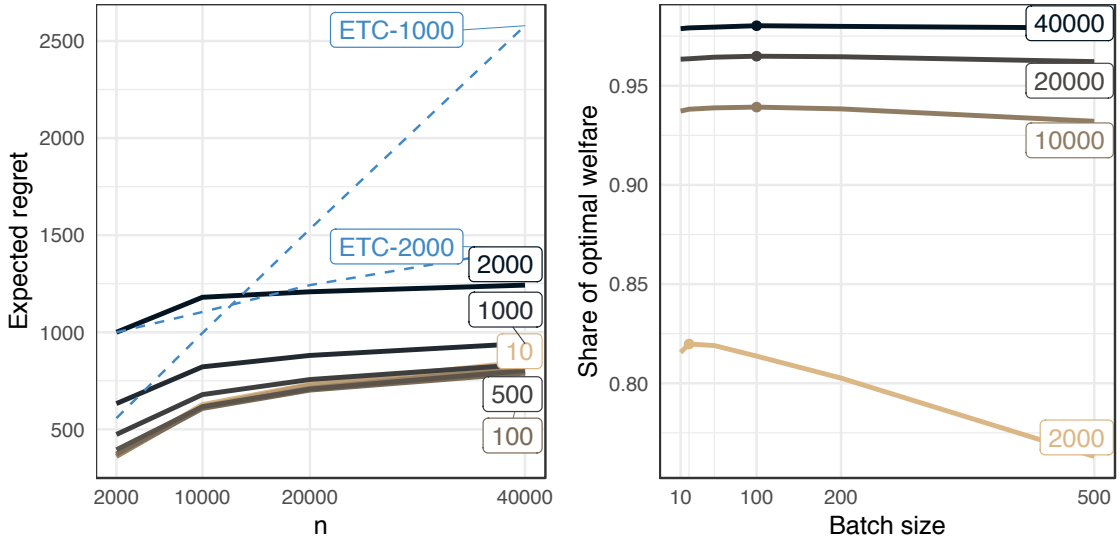


Figure 16: Welfare performance of bandit algorithm with various levels of adaptivity across different horizons. The regret of Thompson sampling grows slower with n than for the explore-then-commit strategy common in the econometric practice. Longer horizons lessen the importance of the choice of batch size. Number of simulations = 10,000.

The left panel of the chart focuses on the choice of batch size by different horizons. If the horizon is shorter, smaller batch sizes are better: quicker adaptivity means less opportunity cost at the beginning. Extreme adaptivity can still lead to over-fitting and thus, lower welfare. As the horizon gets longer, larger batch sizes fare better. This result might be explained by the fact that in the longer run, one has more time to invest in learning as there will be more time to gather the interests. Note also, that for shorter horizon, smaller batch size means the same number of batches. E.g. for $n = 2000$, the best batch size of 20 means 100 batches, the same, as the optimal batch size of 100 for the $n = 10,000$ case. The most decisions should be made in the longest horizon setup (400 batches deliver the best result for $n = 40,000$). It is also worth noting, that

[9]The simulated values are the followings: 2000, 10,000, 20,000, and 40,000.

the importance of the batch size gets less important as the horizon grow: smaller batch sizes reach about the same level of expected welfare.

Figure 17 depicts the performance of different strategies in the welfare-estimation space. The limited IPWE strategy extends the available set of choices, especially if the horizon is shorter. Note that decreasing the horizon is making the learning problem harder, similarly to increasing the noise. Therefore, it is not surprising that the chart for the longest horizon resemble more for the small noise setups of Figure 15.

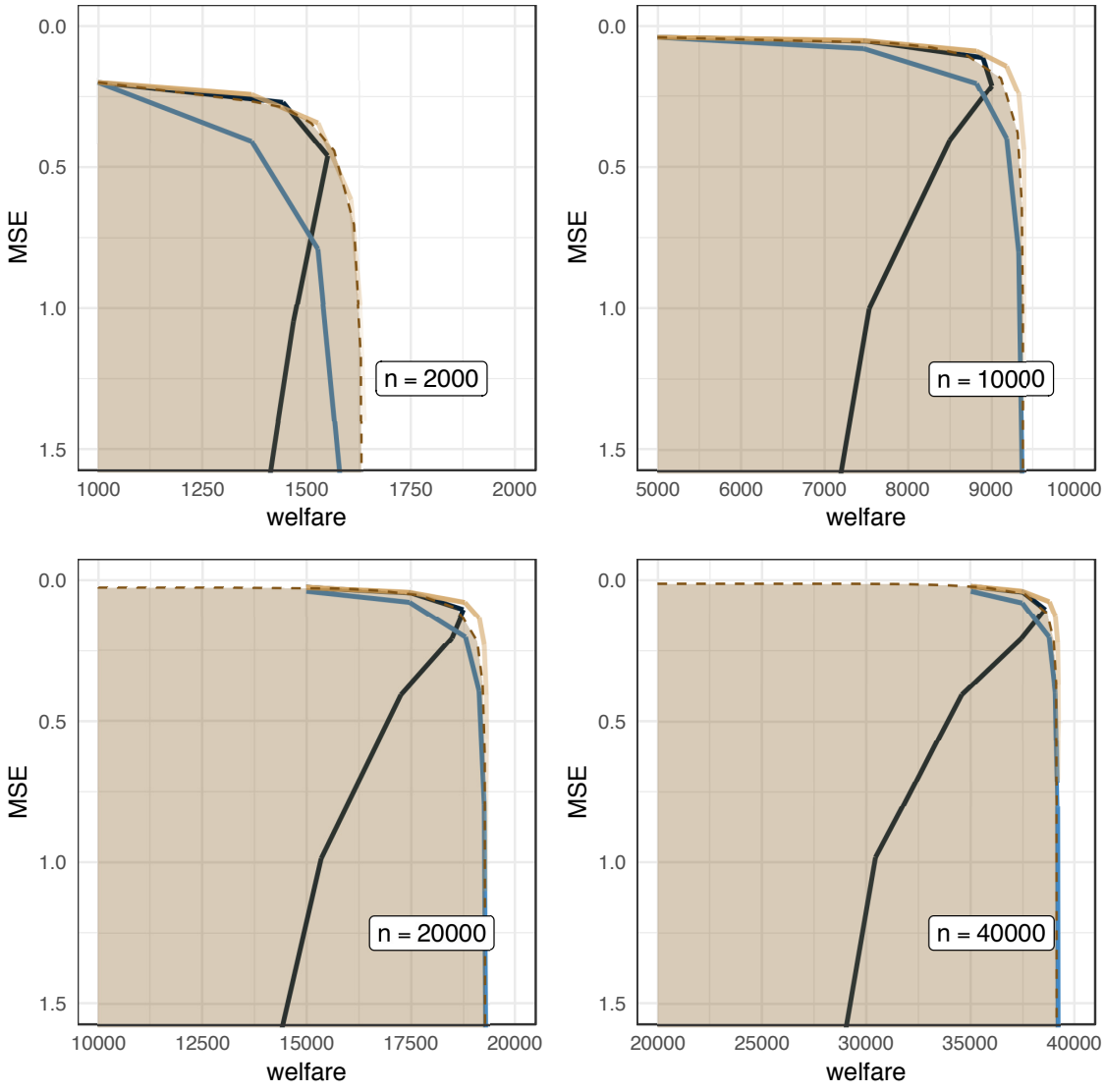


Figure 17: Performance of different strategies in the welfare-estimation space, for different horizons. Limited IPWE always extends the set of possible choices, especially if the problem is hard (n is small). Number of simulations = 10,000.

7 Concluding remarks

In our digital world, collecting data and base our decisions on them are getting technologically feasible. Therefore, online experimentation is getting more and more popular. In this paper, I dealt with this problem from a new perspective. Instead of focusing either on welfare maximization or estimation, I take a more practical viewpoint by considering both goals together. I borrow ideas from program evaluation and apply them on multi-armed bandits to improve upon the established methods valued by both welfare and estimation metrics.

Running a systematic Monte Carlo study, I highlight an important trade-off between welfare and estimation: experimentation strategies that result in good estimators (such as randomized controlled trial) suffer from huge opportunity cost, whereas the bandit algorithm that optimizes for welfare leads to biased treatment effect estimate. Some straight-forward solutions (e.g. explore-then-commit, bandit with estimation on randomized subsample) form transitions between the two extremes, so they provide good choices for policy-makers who have both welfare and estimation goals.

My contribution is threefold: First, I characterize the behavior of a well-known bandit heuristic, the Thompson sampling, across different setups. The standard treatment effect estimator on adaptively collected data suffers from amplification bias, and this bias increases in the relative size of the treatment effect and in the speed of adaptivity of the algorithm (smaller batches). The traditional bias correction method of inverse propensity weighting (IPW) does not work, it can even exacerbate the bias. Second, I highlight the welfare-estimation trade-off for established solutions. Finally, I suggest an easy-to-implement trick to correct the bias: limiting the adaptivity of the data collection by requiring sampling from all arms. Using inverse propensity weighting on data that arise from limited adaptivity results in an unbiased treatment effect estimate, whereas it preserves almost all of the welfare gain stemming from adaptivity.

If you face an easy problem where the relative size of the treatment effect is large, quick adaptivity along with small (or even no) limiting is the best choice to reach both high welfare and a reasonable estimator. If the noise is larger, choosing a higher batch size (skipping some decisions) is a better idea, as it could improve the expected outcome (similarly to how regularization improves prediction accuracy if the noise is large). Running a bandit algorithm with limiting has a major advantage over the explore-then-commit strategy. While the latter could beat the frontier defined by the best batch size and limit combinations in certain setups, one should choose the sample for exploration optimally to realize this result. However, this sample should be chosen in advance where we do not know the relative treatment effect, nor the horizon. In contrast, when running an adaptive experiment, one can change the batch size and limiting parameters throughout the whole process, and adjust them according to the actual knowledge about the

environment – without risking unbiasedness.

My simulation considered only a very simple setup. Real world scenarios often include fat tail distributions, or much more than just one treatment. I stick to the simple setup to concentrate on the basic mechanisms of adaptive data collection. The main result of the welfare-estimation trade-off should hold for a much broader set of environments. I suppose that regularization with higher limits and larger batch sizes gets more important for fat tail distributions. However, this question should be answered by future research.

I expect that adaptive experiments are becoming more popular in every field, including economics. Understanding its mechanisms is essential to be able to use this tool correctly. This paper hopefully could contribute to this purpose.

Appendix

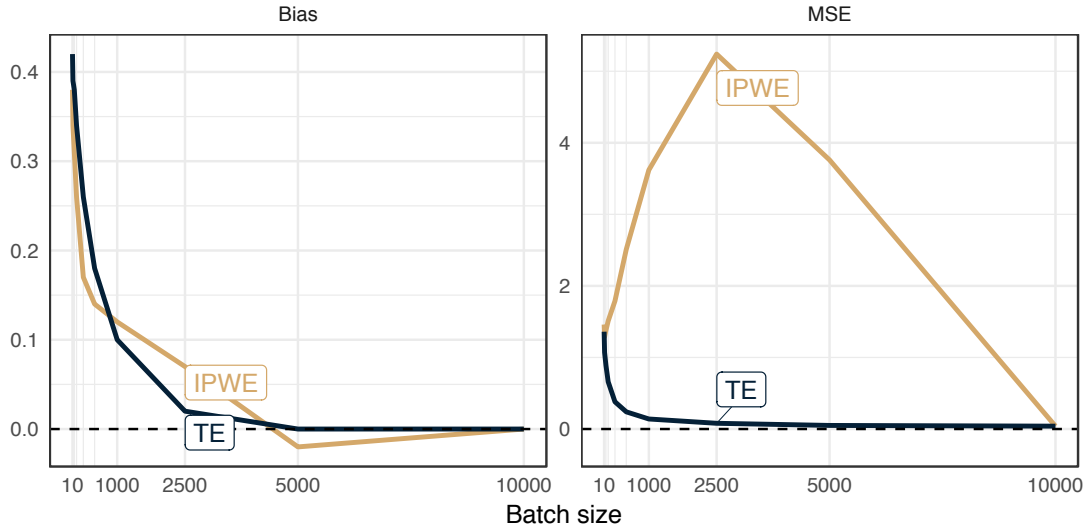


Figure A18: Estimation performance by batch size. Quicker adaptivity (smaller batch size) leads to higher welfare only at the cost of worse (more biased and more volatile) estimators. IPWE can mitigate the bias for small batches, but generally, the improvement is small and it also means a much higher mean squared error.

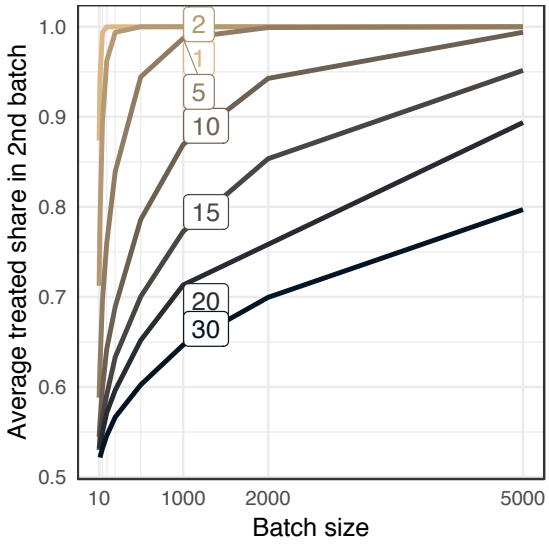


Figure A19: Average treated share in the second batch. If the noise is small and the adaptivity is slow enough, full learning occurs. These situations do not cause any bias, and they end up with the highest possible welfare (see the left panel of Figure 13). Number of simulations = 10-50,000.

References

Athey, S. and Wager, S. (2019), Efficient Policy Learning.
URL: *https://arxiv.org/abs/1702.02896*

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

Dimakopoulou, M., Zhou, Z., Athey, S. and Imbens, G. (2018), 'Estimation Considerations in Contextual Bandits'.
URL: *http://arxiv.org/abs/1711.07077*

Graepel, T., Quinonero Candela, J., Borchert, T. and Herbrich, R. (2010), Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine, *in* 'Proceedings of the 27th International Conference on Machine Learning (ICML)', pp. 13–20.

Hahn, J., Hirano, K. and Karlan, D. (2011), 'Adaptive experimental design using the propensity score', *Journal of Business and Economic Statistics* **29**(1), 96–108.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York Inc., New York, NY, USA.

Hirano, K. and Porter, J. R. (2009), 'Asymptotics for Statistical Treatment Rules', *Econometrica* **77**(5), 1683–1701.

Kasy, M. (2016), 'Why experimenters might not always want to randomize, and what they could do instead', *Political Analysis* **24**(3), 324–338.

Kasy, M. and Sautmann, A. (2019), 'Adaptive Experiments for Policy Choice'.
URL: *https://maxkasy.github.io/home/files/papers/adaptiveexperimentspolicy.pdf*

Kitagawa, T. and Tetenov, A. (2018), 'Who should be treated? Empirical welfare maximization methods for treatment choice', *Econometrica* **86**(2), 591–616.

Korda, N., Kaufmann, E. and Munos, R. (2013), Thompson Sampling for 1-Dimensional Exponential Family Bandits, *in* 'Advances in Neural Information Processing Systems 26 (NIPS)', pp. 1448–1456.

Lai, T. L. and Robbins, H. (1985), 'Asymptotically Efficient Adaptive Allocation Rules', *Advances in Applied Mathematics* **6**(1), 4–22.

Lattimore, T. and Szepesvári, C. (2019), *Bandit Algorithms*, Cambridge University Press.
URL: *https://banditalgs.com/2018/07/27/bandit-algorithms-book/*

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

Nie, X., Tian, X., Taylor, J. and Zou, J. (2018), Why Adaptively Collected Data Have Negative Bias and How to Correct for It, *in* 'Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)'.
URL: *http://arxiv.org/abs/1708.01977*

Perchet, V., Rigollet, P., Chassang, S. and Snowberg, E. (2016), 'Batched bandit problems', *Annals of Statistics* **44**(2), 660–681.

Russo, D., Van Roy, B., Kazerouni, A., Osband, I. and Wen, Z. (2017), 'A Tutorial on Thompson Sampling', *Foundations and Trends® in Machine Learning* **11**(11), 1–96.
URL: *http://arxiv.org/abs/1707.02038*

Scott, S. L. (2010), 'A modern Bayesian look at the multi-armed bandit', *Applied Stochastic Models in Business and Industry* **26**, 639–658.

Slivkins, A. (2019), *Introduction to Multi-Armed Bandits*.
URL: *http://slivkins.com/work/MAB-book.pdf*

Thompson, W. R. (1933), 'On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples', *Biometrika* **25**(3-4), 285–294.

Villar, S. S., Bowden, J. and Wason, J. (2015), 'Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges', *Statistical Science* **30**(2), 199–215.