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

Preliminary Draft

János K. Divényi [†]
Central European University

January 24, 2020

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 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.

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.

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 consider situations similar to the online shop example above, where the decision-maker assigns individuals to different treatments with two goals in mind: (1) maximizing profit (or welfare) and (2) estimating the treatment effect. There are two treatments (status quo and innovation, control and treatment) and individuals arriving in groups or batches should be assigned to one of them. 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. The size of the batches, ie. the frequency of allocation decisions is controlled by the decision-maker.

I run Monte Carlo simulations to understand the welfare and estimation behavior of different strategies in this setup. I study a well-known multi-armed 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).

What we know so far The welfare 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 assignment policy (which is usually infeasible). The difference between these two measures is the expected regret. Each bandit can be characterized by their worst-case regret (within a given set of environments formed by the distribution of rewards and the length of the horizon). The seminal paper of Lai and Robbins (1985) derived an asymptotic lower bound on regret that any bandit algorithm should suffer.

Korda et al. (2013) prove that Thompson sampling is asymptotically optimal on Gaussian rewards with known variance. Perchet et al. (2016) extends their result to batched bandits, where individuals arrive in groups (or batches) instead of one-by-one. The traditional solution in econometrics to experiment first and form an appropriate assignment rule later is suboptimal (see e.g. Lattimore and Szepesvári 2019).

There are much less result that considers estimation after bandits. 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 that can eliminate the bias.

Villar et al. (2015) compare various bandit algorithms in terms of outcome and also estimation performance in a simulated clinical trial. They show biased treatment effect estimations simulating many different multi-armed bandit algorithms.

My contribution To my knowledge, this is the first paper that considers welfare and estimation goals parallel and compares different strategies in the welfare-estimation space. I have three main contributions to the literature:

First, I characterize the welfare and estimation behavior of Thompson sampling and the traditional treatment effect estimator on adaptively collected data. I show that, generally, smaller batch size (ie. deciding more often) increases the expected welfare. However, too quick adaptivity (really small batch size) leads to increased probability of "getting lost" that results in a welfare cost that outweighs the gains from smaller opportunity cost. Quicker adaptivity also increases the negative bias in means (for which I provide an intuitive explanation) that results in a larger amplification bias in the treatment effect estimate. These results highlight an important trade-off: strategies that achieve high welfare (adaptive algorithms) lead to highly biased treatment effect estimates - whereas running a randomized controlled trial on the whole sample (the gold standard for measuring the effect) suffers from a huge opportunity cost (resulting from assigning too many individuals to the inferior treatment).

Second, I prove that inverse propensity weighting (IPW) – traditionally used for bias correction – is equivalent to taking the simple averages of the batch averages (if the propensity weights are estimated). I show that in this setup, IPW does not work – in fact, it can even exacerbate the bias.

Finally, I suggest an easy-to-implement bias correction method: limiting the propensity scores away from the extremes that practically moderates the adaptivity of the data collection by requiring sampling from both arms in each batch. This assignment rule allows for unbiased inverse propensity weighted 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 some established strategies (such as the standard "explore first, exploit later" or explore-then-commit strategy).

Related recent literature A recent paper of Hadad et al. (2019) deals with a similar problem: they suggest data-adaptive weighting schemes to correct the standard treatment effect estimator on adaptively collected data, also ensuring asymptotic normality to make statistical inference possible. They deal only with estimation, and do not consider welfare.

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. In contrast, I focus on the general characteristics of the standard treatment effect estimator that are apparent even if the effect itself is constant.

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's setup is especially close to mine 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.

This paper The paper is structured as follows. Section 2 gives a formal setup for the problem. Section 3 characterizes the basic welfare and estimation properties of the bandit assignment rule using the standard treatment effect estimate and shows the welfare-estimation trade-off. Section 4 discusses different methods for correcting the bias: inverse-propensity weighting, first batch treatment effect and propensity score limiting. Section 5 demonstrates the results of the systematic Monte Carlo simulation which illustrate the behavior of the previously discussed strategies in different scenarios. Section 6 concludes.

2 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 treatment effect. Therefore, the population is characterized by $\{Y_i(0)\}_{i=1}^{n}$. For simplicity, I assume $Y(0)$ is Gaussian with known variance.

Individuals arrive randomly in equal-sized batches denoted by B and indexed by $j \in \{1, ..., m\}$. The batch size is under the control of the decision-maker[1] and 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.

[1] It is natural to assume that the decision-maker has some control over the batch size. Even if the arrival of individuals is dictated by an external process, one can still increase the batch size by collapsing original batches. How frequently the decision-maker decides about allocation is a decision itself.

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.

Let us denote the observed history (assignments and outcomes) up until the kth batch by $H^{(k)} = \{Y_i, W_i\}_{i \in \bigcup_{j=1}^{k} B_j}$. Therefore, the whole history of n individuals is $H^{(m)}$.

The decision-maker has two goals: she cares about the total outcome of individuals[2], and she also wants to estimate the treatment effect τ with an unbiased, precise estimator. She decides about two things in parallel:

1. **assignment rule** A function that maps the history to a probability that expresses the share of the next batch assigned to the treatment: $\pi\left(H^{(k)}\right) = \mathbb{P}\left(W_i = 1 | i \in B_{k+1}\right) = p_{k+1}$.
   Goal (welfare): $\max \sum_{i=1}^{n} Y_i$

2. **estimation method** A function that maps the whole history (observed data of the population) to a number that expresses the treatment effect: $\hat{\tau}\left(H^{(m)}\right)$.
   Goal (estimation): $\min \mathbb{E}\left[(\hat{\tau} - \tau)^2\right]$ subject to $\mathbb{E}[\hat{\tau}] = \tau$

I will call a combination of an assignment rule and an estimation method a **strategy**.

To illustrate adaptive assignment rules that blend exploitation with exploration I use an old heuristic, the Thompson Sampling (Thompson 1933). It suggests to assign each individual to treatment by the probability that corresponds to your actual beliefs that the treatment outcome is the highest[3]. I implement this rule as follows:

[2] It might be the profit of a firm assuming the outcome contains the cost of the treatment, or the total welfare assuming a utilitarian welfare function.

[3] 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 (recall the known-variance assumption)[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

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

3. Calculate the probability that the treatment mean is higher than the control mean (let us denote it with $r^{(k)}$). Technically, this can be achieved by sampling from the corresponding distributions.
4. Split the next batch according to this probability: $p_{k+1} = r^{(k)}$
5. 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 Characterization of welfare and estimation properties

3.1 Parametrization

I assume – without loss of generality – a positive average treatment effect with unit value ($\tau = 1$). The population consists of $n = 10,000$ individuals, the potential outcomes are Gaussian with $\sigma = 10$. 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. The minimum batch size is 10 (where $m = 1000$), and I simulate the following choices for the decision-maker: $n_B \in \{10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000\}$. The maximum value corresponds to a simple random split on the whole sample.

In this setup, 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.

I run $20,000$ simulations for each assignment rule. The runs differ only in the sequence of 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 Welfare

One would expect that smaller batch size (more batches, quicker adaptivity) leads to higher welfare, as it extends the possibilities of the policy maker. Also, as the first batch is a simple random split, the maximum welfare an adaptive rule could achieve in the best case is $10,000 - \frac{n_b}{2}$. Smaller batch sizes give the chance of reacting more quickly to a positive treatment effect, hence, suffering less opportunity cost.

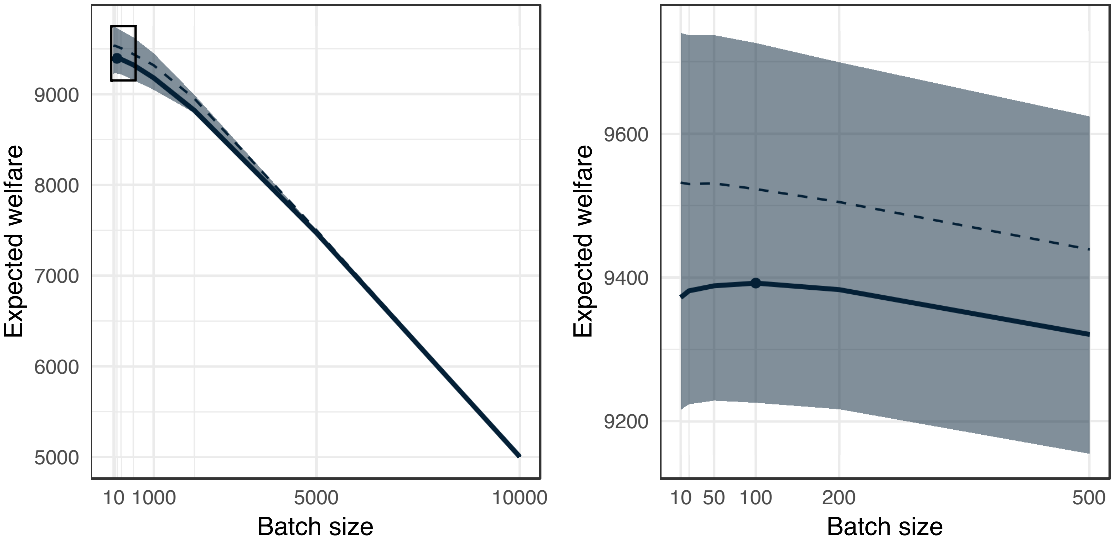

Figure 1: Expected welfare by batch size (shaded: interquartile range, dashed line: median, point: maximum). The right panel focuses on small batch sizes. Smaller batches (quicker adaptivity) generally leads to higher welfare, but only after a certain point: really small batch size can harm. Number of simulations = 20,000.

Figure 1 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 $n_B = 100$. The reason for this is that being more adaptive means deciding based on more volatile estimates that increases the probability of "getting lost", and adapting to the wrong pattern. Figure A17 in the Appendix shows that the distribution of welfare is much more volatile for smaller batch sizes. Under a certain threshold of batch size, the loss on volatility outweighs the gain on opportunity cost.

Figure 2 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 smaller batch sizes mean that 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 1 originates from these two contradicting processes.

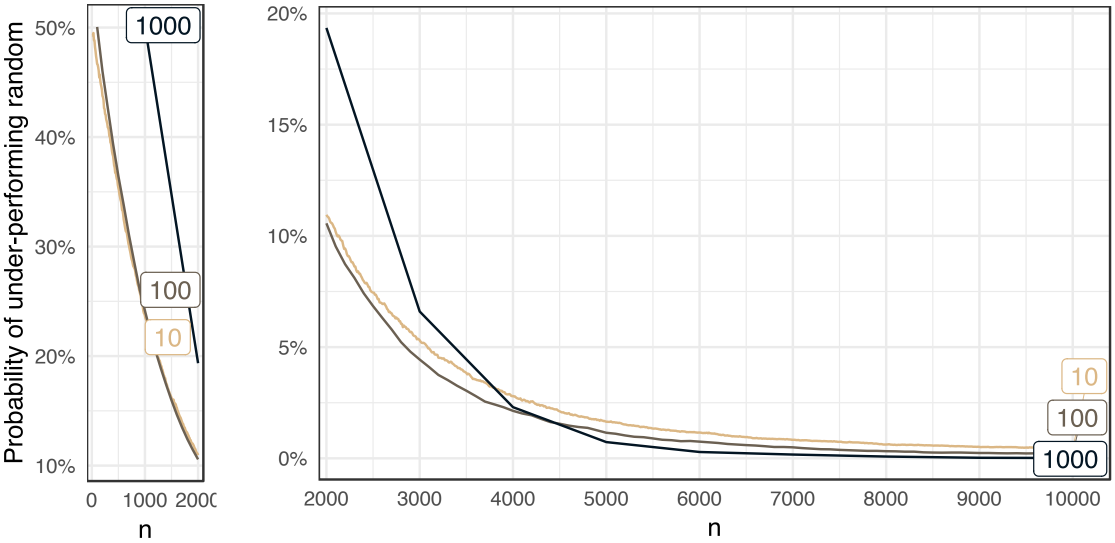

Figure 2: Evolution of bandit algorithms for some batch sizes across arriving individuals, measured by the probability of under-performing a simple random split at each point of the process. Quicker adaptivity results in smaller opportunity cost at the beginning (left panel), but leads to higher probability of getting wrong at later stages (right panel).
Number of simulations = 20,000.

The behavior of the batch size parameter lets us raise an interesting analogy from the machine learning literature: regularization (see e.g. Hastie et al. 2001) is a technique that discourages learning a too complex or flexible model (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 constrains the set of choices and 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 algorithm is a general solution, working well in different setups whose parameters (mainly τ and n) are ex-ante 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.

3.3 Estimation

The standard method to estimate the treatment effect is to compare the observed averages of the individuals in both groups:

$$\hat{\tau}_0 \quad = \quad \frac{\sum_{i=1}^{n} Y_i W_i}{\sum_{i=1}^{n} W_i} - \frac{\sum_{i=1}^{n} Y_i (1 - W_i)}{\sum_{i=1}^{n} (1 - W_i)} \tag{1}$$

According to the theoretical results of Nie et al. (2018) the averages are negatively biased estimator for the true expected values of the outcomes. Figure 3 characterizes the bias for different choices of batch size. It confirms the negative bias result and shows two additional interesting result: (1) quicker adaptivity leads to a more volatile estimate with larger bias and (2) the control mean contains a larger (negative) bias that is more volatile than the treatment mean. The latter result follows from the fact that the treatment effect is positive so we end up with much more treatment observations (recall that the expected welfare equals to the number of individuals assigned to the treatment). As a result, the treatment effect estimator suffers from amplification bias but because of partial compensation, the bias in the s smaller than the bias in the control mean (Figure A18 in Appendix shows the distribution of τ_0 for different batch sizes).

The negative 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

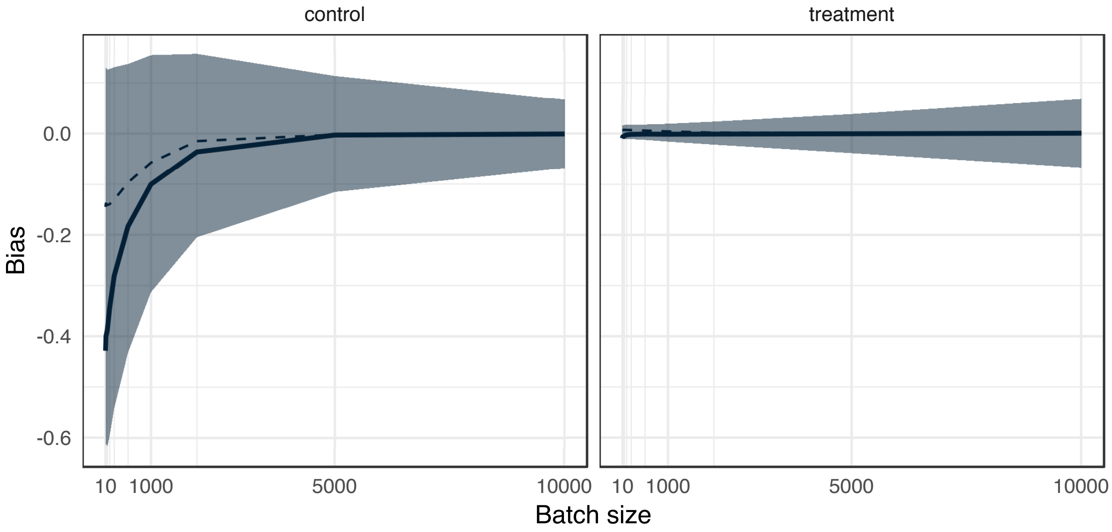

Figure 3: Bias in group mean estimates by batch size (shaded: interquartile range, dashed line: median, point: maximum). Quicker adaptivity results in larger negative bias that is much more expressed for the control (as we end up with more treatment observations). Number of simulations = 20,000.

experiment, than in the case when the first batch results in a positive control estimate.

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 of $\hat{\tau}_0$: 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 2). 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 regarding 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.

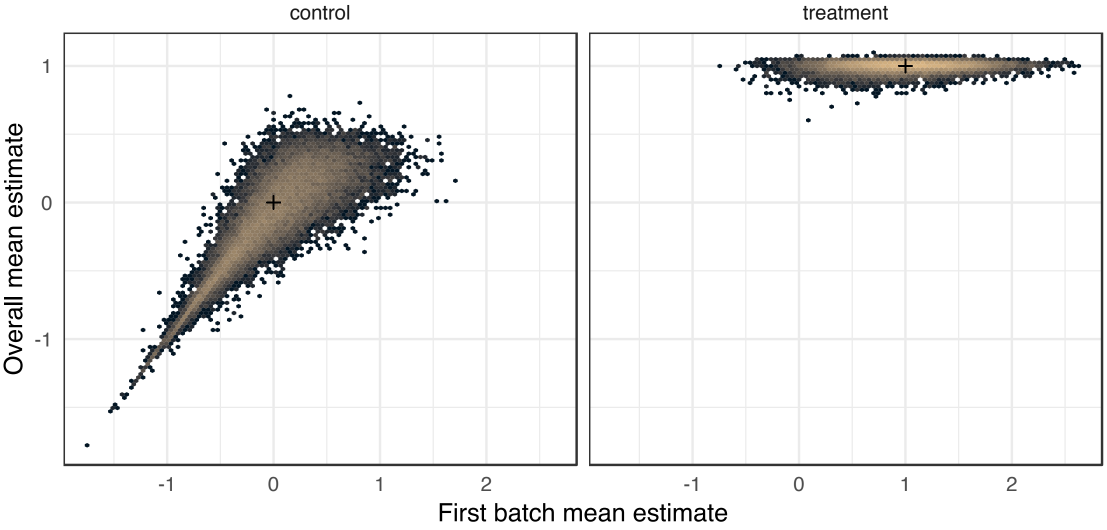

Figure 4: Estimated means using only the first batch ($n_B = 1000$) versus the estimated means on the whole sample (density: darker regions mean higher density). The importance of the first batch estimate is clear, especially for the control outcome: an underestimated group mean from the first batch remains uncompensated in the overall estimate. Number of simulations = 20,000.

$$\hat{\tau}_0 = \sum_{j=1}^{m} \underbrace{\frac{\sum_{i\in B_j} Y_i W_i}{\sum_{i\in B_j} W_i}}_{\substack{\text{batch treated}\\ \text{average}}} \underbrace{\frac{\sum_{i\in B_j} W_i}{\sum_{i=1}^{n} W_i}}_{\substack{\text{share of batch}\\ \text{within all treated}}} - \sum_{j=1}^{m} \underbrace{\frac{\sum_{i\in B_j} Y_i(1-W_i)}{\sum_{i\in B_j}(1-W_i)}}_{\substack{\text{batch control}\\ \text{average}}} \underbrace{\frac{\sum_{i\in B_j}(1-W_i)}{\sum_{i=1}^{n}(1-W_i)}}_{\substack{\text{share of batch}\\ \text{within all control}}} \tag{2}$$

3.4 Welfare-Estimation Trade-off

My previous results suggest an interesting trade-off: quicker adaptivity generally results in higher expected outcome (welfare goal) but leaves us with a more biased and more volatile treatment effect estimate (estimation goal). Using the maximum batch size of $10,000$ is equivalent to running a randomized controlled trial (RCT) on the whole sample: being the gold standard for measuring an effect it results in a reasonable estimate, but also a much lower expected welfare.

To compare the performance of different strategies in this space I plot the expected welfare (x axis) against the mean squared error[4] of the estimator (reversed y axis) as in Figure 5. To highlight the decision-maker's constraint of unbiasedness, biased estimates are shown with hollow circles whose transparency is proportional to the size of bias. The best strategy would be a strong point at the top right corner: with a total welfare of $10,000$ and an unbiased treatment effect estimate

[4] Recall that MSE = bias^2 + variance.

with zero MSE. Obviously, such a strategy does not exist.

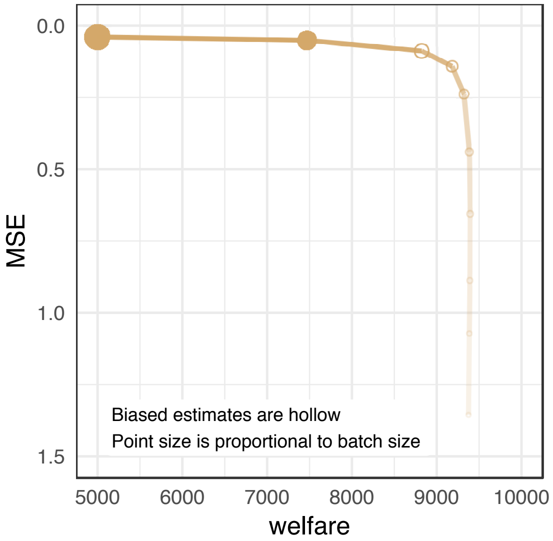

Figure 5: Performance of bandit assignment rule with different batch size choices in the welfare-estimation space (using the standard treatment effect estimation). Quicker adaptivity leads to higher welfare but also larger bias and larger MSE. Number of simulations = 20,000.

Each strategy on the figure combines the adaptive allocation rule with $\hat{\tau}_0$, the only difference is the choice of n_B. A decision-maker who only cares about the estimation goal would choose the top left point of full RCT. Moving towards more adaptive rules brings significant welfare gains for a slow increase in the variance of the estimator. However, the bias needs to be corrected.

4 Bias correction

4.1 Inverse Propensity Weighting (IPW)

A standard technique to correct bias in the treatment effect estimator is inverse propensity weighting (also mentioned by Nie et al. 2018, Dimakopoulou et al. 2018). I prove in Equation 3 that using IPW with estimated[5] propensity score (the actual share of a batch assigned to the treatment) is equivalent to using simple average of the batch averages (without weighting as in $\hat{\tau}_0$). Following from the fact that each group average is an unbiased estimate for the corresponding group mean, this method takes the averages of multiple unbiased estimates and thus gets rid of the compositional effect and takes the averages of multiple unbiased estimates. As individuals arrive in batches, individual propensity scores depend only on the individual's batch: $p_i = \mathbb{P}(W_i = 1) = p_j$ for $i \in B_j$.

[5] Other works, such as Hadad et al. (2019), use true propensity scores instead. This requires that one stores the allocation probabilities as well. For me, $\{Y_i, W_i\}$ suffice.

$$\begin{aligned} \hat{\tau}_{IPW} &= \frac{1}{n}\left(\sum_{i=1}^{n}\frac{Y_iW_i}{p_i} - \sum_{i=1}^{n}\frac{Y_i\left(1-W_i\right)}{1-p_i}\right)\\ &= \frac{1}{n}\sum_{j=1}^{m}\left(\sum_{i\in B_j}\frac{Y_iW_i}{p_j} - \sum_{i\in B_j}\frac{Y_i\left(1-W_i\right)}{1-p_j}\right)\\ &= \frac{1}{n}\sum_{j=1}^{m}\left(\sum_{i\in B_j}\frac{Y_iW_in_B}{\sum_{i\in B_j}W_i} - \sum_{i\in B_j}\frac{Y_i\left(1-W_i\right)n_B}{\sum_{i\in B_j}\left(1-W_i\right)}\right)\\ &= \frac{1}{m}\sum_{j=1}^{m}\underbrace{\frac{\sum_{i\in B_j}Y_iW_i}{\sum_{i\in B_j}W_i}}_{\substack{\text{batch treated}\\ \text{average}}} - \frac{1}{m}\sum_{j=1}^{m}\underbrace{\frac{\sum_{i\in B_j}Y_i(1-W_i)}{\sum_{i\in B_j}(1-W_i)}}_{\substack{\text{batch control}\\ \text{average}}} \end{aligned}\tag{3}$$

However, IPW does not seem to be effective: instead of eliminating the bias, it can even exacerbate the problem (Figure A19 in Appendix shows the distributions of $\hat{\tau}_{IPW}$ for different batch sizes). The volatility of the estimator is also much higher.

The reason for this 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 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). I illustrate this process for $n_B = 1000$. Table 1 summarizes the expected value of the estimator by how many batches contained any control assignee: the more batch is without controls (everyone is assigned to the treatment) the more over-estimated is the effect. As the natural consequence of this selection, runs with controls in every batch (the majority) result in an under-estimated treatment effect.

Table 1: Comparison of $\hat{\tau}_{IPW}$ by number of batches with control assignment

# of batches with controls	1	2	3	4	5	6	7	8	9	10
$\mathbb{E}[\hat{\tau}_{IPW}]$	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 ($n_B = 1000$) underestimate the treatment effect while runs with batches without controls overestimate the treatment effect, using the average of averages ($\hat{\tau}_{IPW}$) 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.

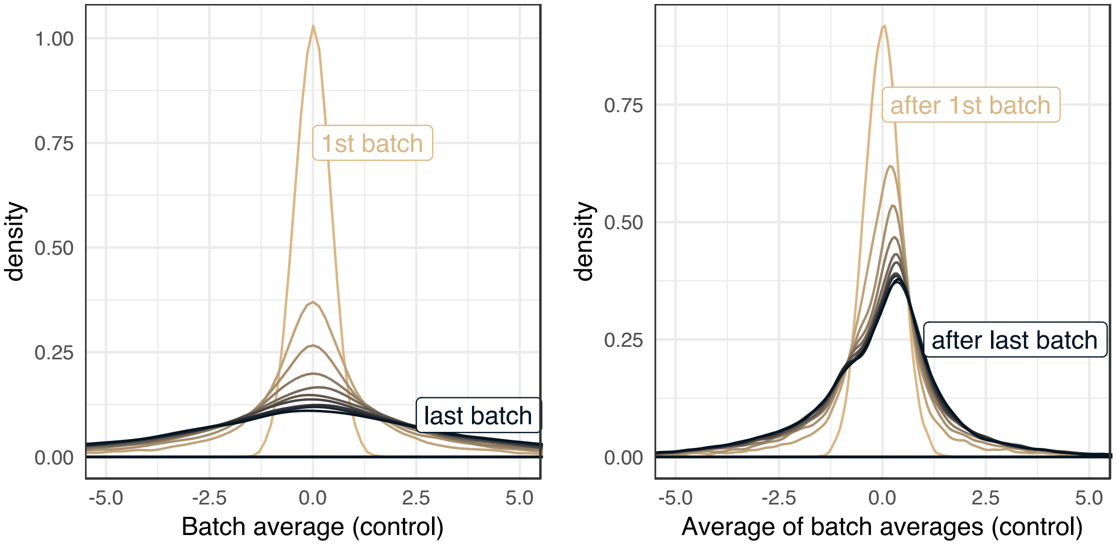

Figure 6: Batch average for the control mean across batches ($n_B = 1000$). Each batch in itself is unbiased. Average of batch averages is getting biased due to selection.
Number of simulations = 20,000.

4.2 Using the first batch only

One can overcome the problem with inverse propensity weighting by using only the data collected in the first batch. I call this as First Batch Estimator ($\hat{\tau}_{FB}$):

$$\hat{\tau}_{FB} \quad = \quad \frac{\sum_{i \in B_1} Y_i W_i}{\sum_{i \in B_1} W_i} - \frac{\sum_{i \in B_1} Y_i (1 - W_i)}{\sum_{i \in B_1} (1 - W_i)} \tag{4}$$

This estimator is unbiased, so the strategy of Thompson sampling assignment rule combined with the first batch estimation method (TS-FB) works. However, it loses on efficiency as it drops a large fraction of observations, especially for small batch sizes (Figure A20 in Appendix shows the distributions of $\hat{\tau}_{FB}$ for different batch sizes).

To better understand the efficiency cost relative to the welfare gain of this strategy, I visualize its performance on the welfare-estimation plot (Figure 7). As a benchmark, I add the traditional strategy in economics where the assignment rule is not adaptive: first, concentrate on the estimation goal and run an RCT on an experimental sample, and then, focus on the outcome and form a deterministic rule based on the 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 rule of Explore-then-commit (ETC):

Explore-then-commit (ETC)

1. Split the first batch equally between treatment and control[a] .
2. Estimate the average treatment effect by comparing the treatment and control averages calculated on the collected data[b]:

$$\hat{\tau}^{(1)} = \hat{\mu}_1^{(1)} - \hat{\mu}_0^{(1)} = \frac{\sum_{i \in B_1} Y_i W_i}{\sum_{i \in B_1} W_i} - \frac{\sum_{i \in B_1} Y_i (1 - W_i)}{\sum_{i \in B_1} (1 - W_i)}$$

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

$$p_k = \arg\max_w \left\{ \hat{\mu}_w^{(1)} \right\} \text{ for } k \geq 2$$

[a] Typically, the size of the batch is calculated by assuming a minimum size for the treatment effect and deriving a required sample size that yields enough power given a predetermined false positive rate (or significance level).

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

Adaptive data collection using $\hat{\tau}_{FB}$ clearly dominates the Explore-then-Commit (ETC) strategy (using $\hat{\tau}_0$) for decision-makers valuing welfare more, but it loses when MSE is more important. The closest choices to the optimal top right point are $n_B \in \{1000, 2000\}$ for both strategies.

4.3 Limiting the propensity scores

With a slight modification of the assignment rule the efficiency problem of the TS-IPW strategy can be improved (while preserving the bias-corrected estimate). As I showed in section 4.1, the reason why τ_{IPW} is biased after adaptive data collection is that the algorithm does not assign to both groups in each batch, and this unanimous assignment asymmetrically depends on previous observations. A simple solution for this issue is to ensure that people are assigned to both groups in each batch, that is to limit the (realized) propensity score away from the extremes of zero and one. Although this method needs the modification of the data collection process, in the digital

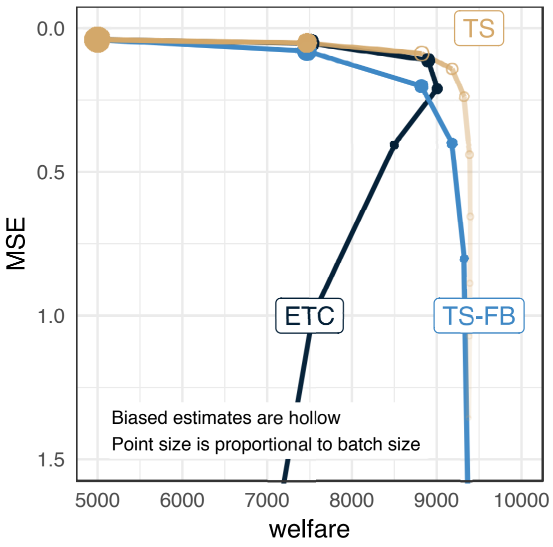

Figure 7: Performance of different strategies in the welfare-estimation space (varying batch size). Generally, quicker adaptivity leads to higher welfare but also larger MSE. ETC with moderate batch size works well, but smaller batch size harms not only MSE but also welfare. TS-FB approximates the standard TS strategy with higher MSE but ensuring an unbiased estimate. Number of simulations = 20,000.

world this is typically not very costly. Also, this solution is easy-to-implement.

Limited Thompson Sampling (LTS)

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.

The smallest possible limitation (e.g. 1% for the batch size of 100) would yield an unbiased $\hat{\tau}_{IPW}$ 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 TS and TS-FB strategies) and a random split of the full sample (50% limit, ETC with $n_B = 10000$, full RCT).

Figure 8 shows the effect of limitation on welfare and estimation goals simulating 8 different limit levels[6]. As expected, higher limit means lower welfare and more precise $\hat{\tau}_{IPW}$ estimate[7].

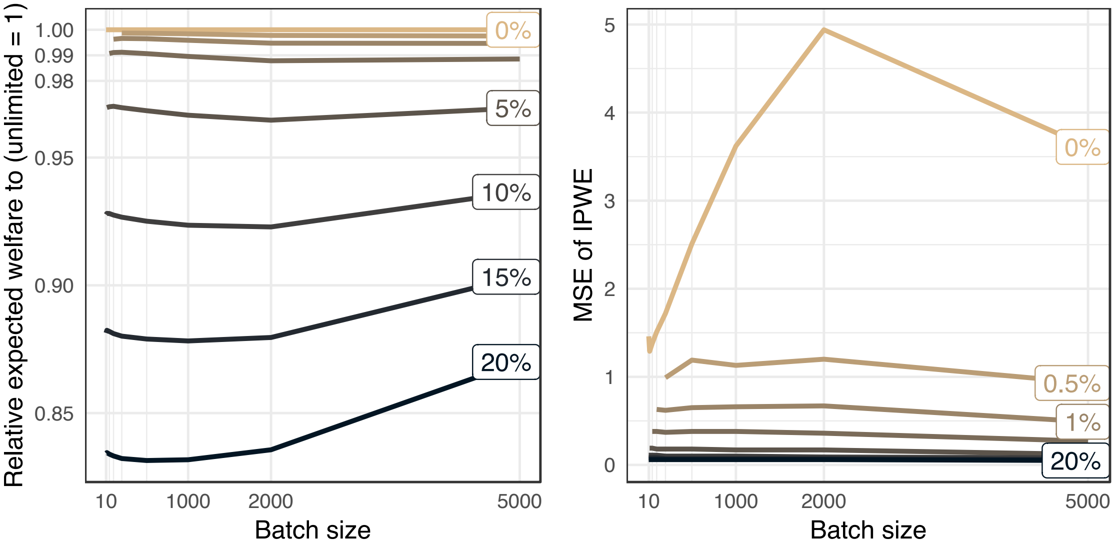

Figure 8: Welfare and estimation performance of the LTS-IPW strategy by batch size. Higher limits incur higher welfare cost (left panel) but bring more precision (right panel). The loss and gain by the amount of limit are disproportionate. Number of simulations = 20,000.

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 $n_B = 2000$ - 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 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.

[6] 0%, 0.5%, 1%, 2%, 5%, 10%, 15% and 20%.

[7] Limitation also decreases the bias of the $\hat{\tau}_0$, but due to the inherent weighting in Equation 2, some bias remains until the limit reaches the level of the simple random split.

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 9 shows the performance of LTS-IPW with different limits. Lower limitation can achieve higher welfare with an appropriate batch size, but only for a growing cost on MSE. The lines are close to horizontal, showing that smaller batch sizes can achieve higher expected welfare for practically no estimation cost. Different points of this chart depict different parametrizations (n_B, L) of LTS-IPW 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) - not necessarily simulated in this exercise.

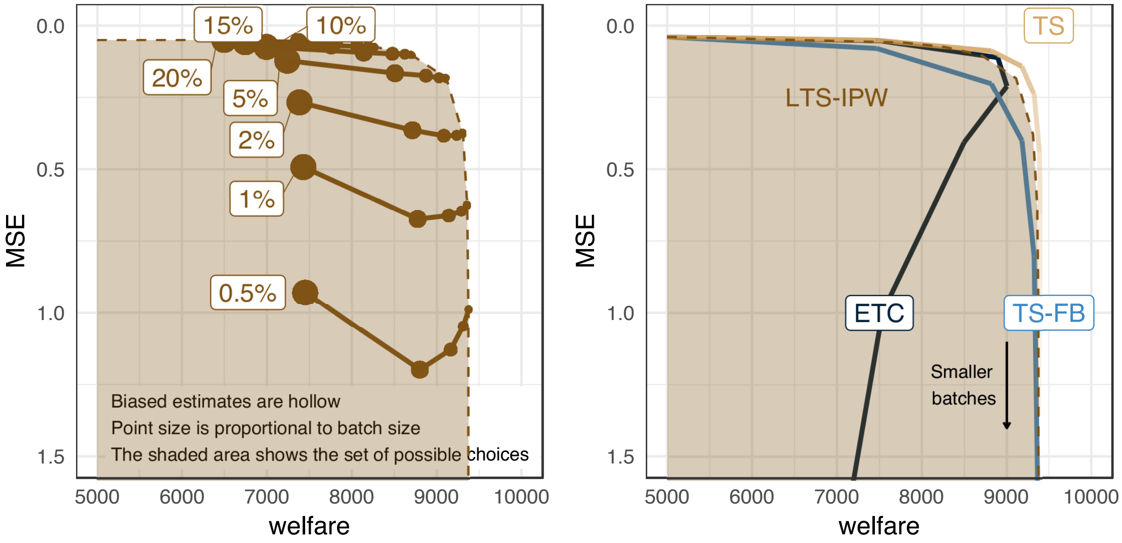

Figure 9: Performance of different strategies in the welfare-estimation space. The right panel shows the outcomes for LTS-IPW with different limits along with the Performance Frontier. The left panel shows only this frontier along with the previous strategies: LTS-IPW extends the possibilities by approximating the TS strategy while also ensuring an unbiased estimate. Number of simulations = 20,000.

The right panel of figure shows only the frontier for the LTS-IPW strategy, along with our previous strategies. Limitation with inverse propensity weighting clearly extends the possibilities of the decision-maker: It gets the closest to the TS strategy but also allows for an unbiased estimate, and dominates TS-FB and also ETC for $n_B < 2000$. If the decision-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

5.1 Uncertainty

Parametrization I investigate the behavior and performance of different strategies with different levels of uncertainty (σ) holding the treatment effect constant at unit value, so σ expresses the noise-to-signal ratio. As the important measure in this problem 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 8 different values for σ with $n = 10,000$[8]. Each setup is simulated with 10 values of batch size and 8 values of limit[9], $10 - 50$ thousand runs for each[10].

Welfare Figure 10 summarizes the results of the expected total welfare and the bias in $\hat{\tau}_0$ 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.

Unlike in the setup of the previous section ($\sigma = 10$), the quickest adaptivity results in the highest expected welfare for low levels of noise ($\sigma < 5$). For these setups, the danger of overfitting is low, so regularizing by increasing the batch size does not help, only incurs a higher opportunity cost.

There is another interesting pattern to note: For welfare, each line approaches the one with the smallest σ as batch size increases, some also reach it. This means that less uncertainty does not lead to higher outcome under a certain value of σ if batches are large enough. 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 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.

[8] $\sigma \in \{1, 2, 5, 10, 15, 20, 25, 30\}$

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

[10] 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.

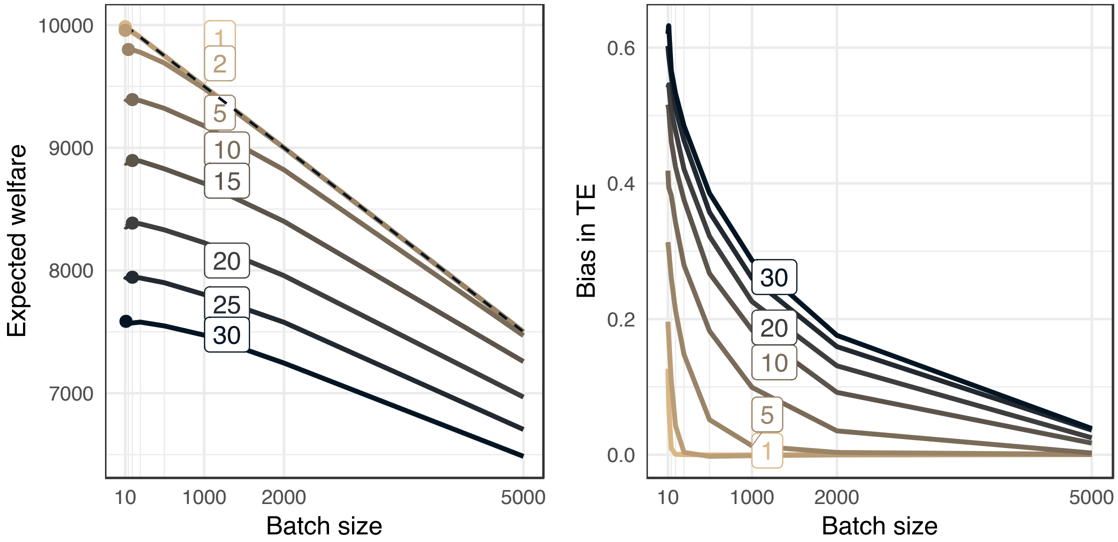

Figure 10: Expected total welfare (left panel) and bias in $\hat{\tau}_0$ (right panel) with different batch sizes (on x axis) for different levels of noise (labelled). The dashed line on left shows the maximum welfare that each strategy could achieve, the points show the batch sizes with the maximum welfare for a given σ. Number of simulations = 10-50,000.

Estimation 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 asymmetric sampling that causes the bias does not kick in. Figure 11 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.

Welfare-Estimation Trade-off The previous results 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 works differently only for two special regions: (1) for high levels of noise, extreme adaptivity hurts both goals, whereas (2) for low levels of noise, adaptivity can be increased until a certain point gathering the welfare gain but without introducing any bias.

I suggested limiting as a working method for bias correction in section 4.3. 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 the set of available choices for the decision-maker in the welfare-estimation space.

Figure 12 shows the performance of the different strategies in the welfare-estimation space for each setup. Similarly to Figure 7, it only shows the frontier for the TS-IPW strategy that is formed by the best combinations of batch size and limit. Obviously, as the problem gets harder

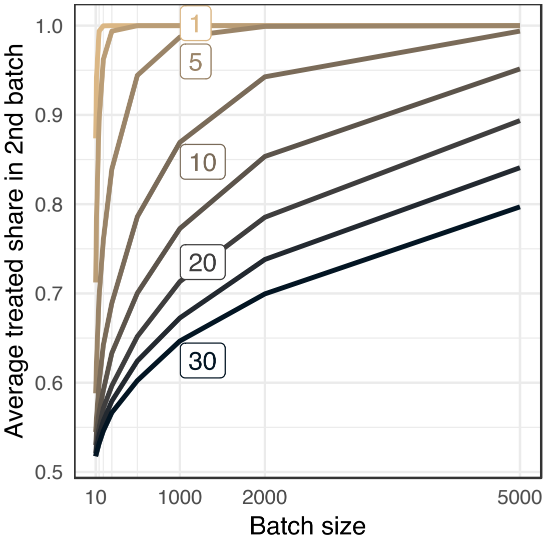

Figure 11: Average treated share in the second batch. If the noise is small and the adaptivity is slow enough, full learning occurs. These situations do not cause any bias, and they end up with the highest possible welfare (see the left panel of Figure 10). Number of simulations = 10-50,000.

(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 TS 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 10).

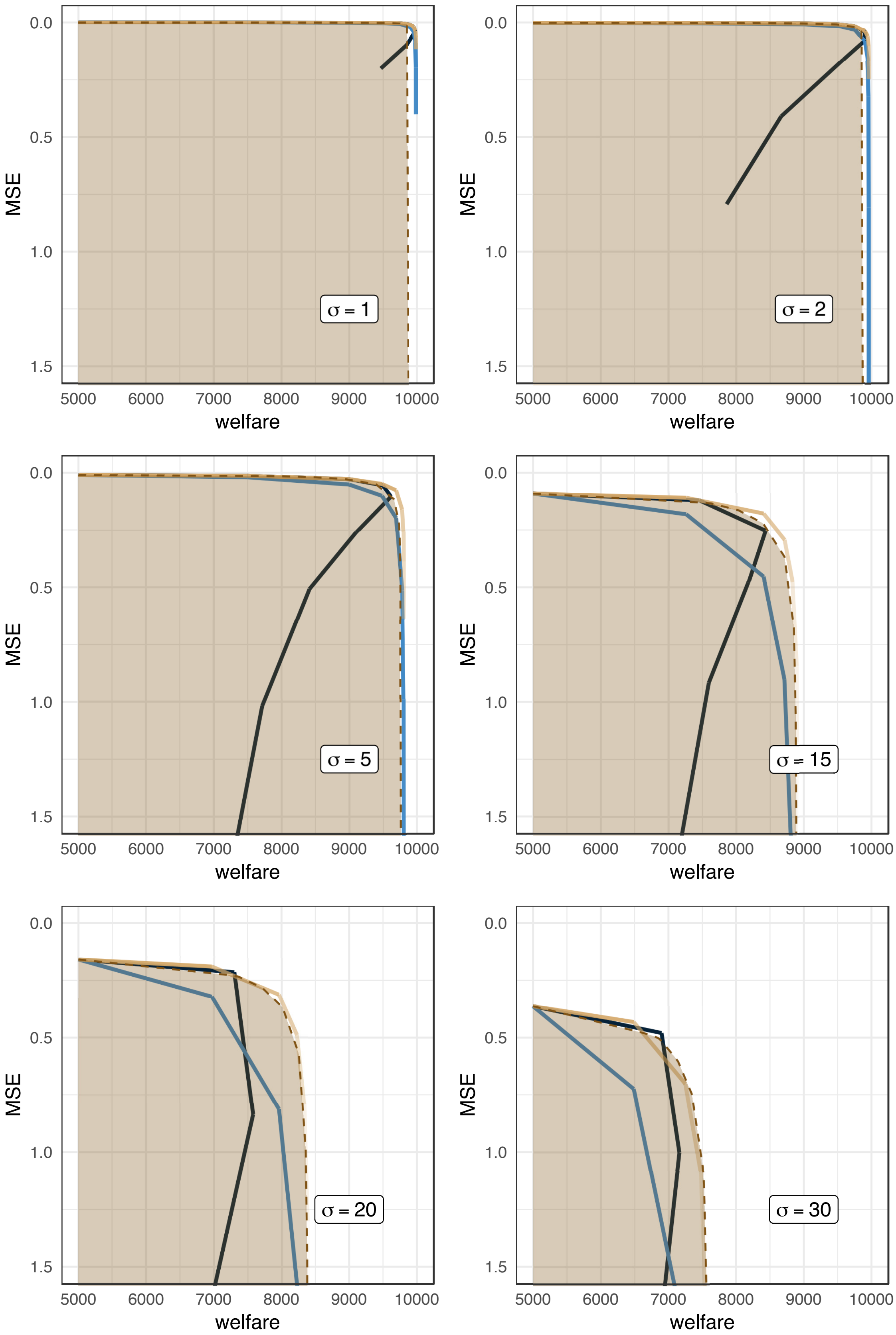

Figure 12: Performance of different strategies in the welfare-estimation space, for different levels of noise. The TS-IPW strategy always extends the set of choices, especially if the problem is hard (the noise is large). Number of simulations = 10-50,000.

In practice it is important to know which combinations form the frontier that extends the possibilities. For welfare, it is obvious, that smaller limits are expected to fare better. However, a small limit excludes small batch sizes as we need control assignees in every batch to ensure unbiasedness. So, it is not straightforward how to choose the best strategy. Figure 13 shows the expected welfare for all batch size - limit combinations, for different levels of uncertainty. There are three interesting results to note:

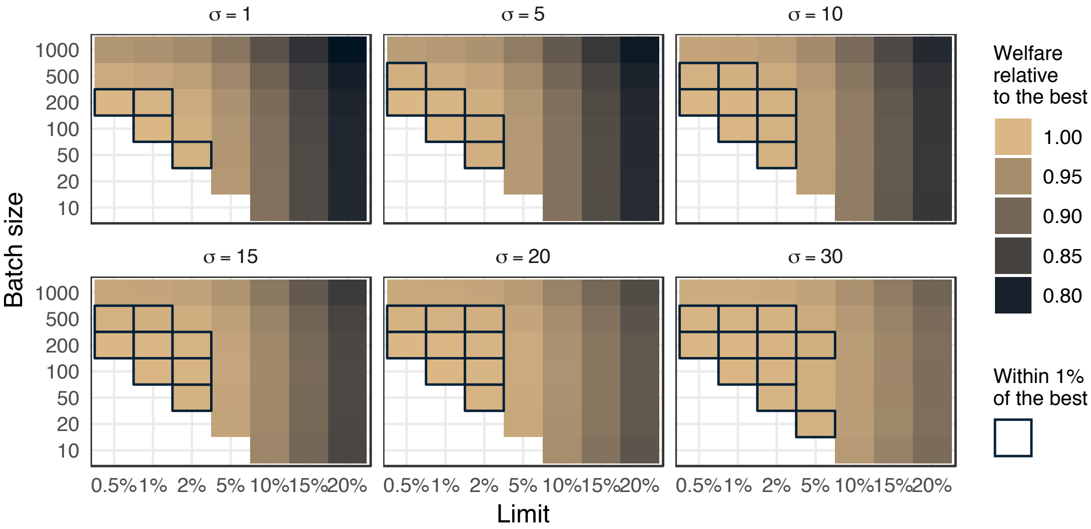

Figure 13: Expected welfare of different combinations of n_B and L, for different levels of σ. Welfare is normalized by the best result within the scenarios. Number of simulations = 10-50,000.

1. Quicker adaptivity is generally better, but not beyond $n_B = 50$. Too small batch size requires too large limit to preserve unbiasedness that adversely affects welfare. Also, the opportunity cost they could possibly win is no more than the size of the batch which is obviously small for small batches.

2. One can increase limit and decrease batch size to achieve about the same welfare. For large noise cases, many combinations result in the same level of welfare.

3. Limiting does not eliminate the problem of over-fitting: too quick adaptivity has a detrimental effect on expected welfare (excluding the low noise scenarios).

Figure 14 shows the same chart for the estimation goal, plotting the MSE of different combinations. As in this case, the important comparison is the estimated treatment effect itself, I use the levels of MSE: a value above 1 means an error that is larger than what is measured.

1. Intuitively, larger noise means larger MSE, across each combinations.

2. Smaller adaptivity and larger limits improve MSE. More interestingly, limiting matters more than batch size: in terms of estimation precision, increasing the limit is more effective than increasing the batch size.

3. The combination that results in the smallest MSE while still achieving the maximal welfare is: $\{n_B = 50, L = 2\%\}$ for $\sigma = 1$ while $\{n_B = 200, L = 5\%\}$.

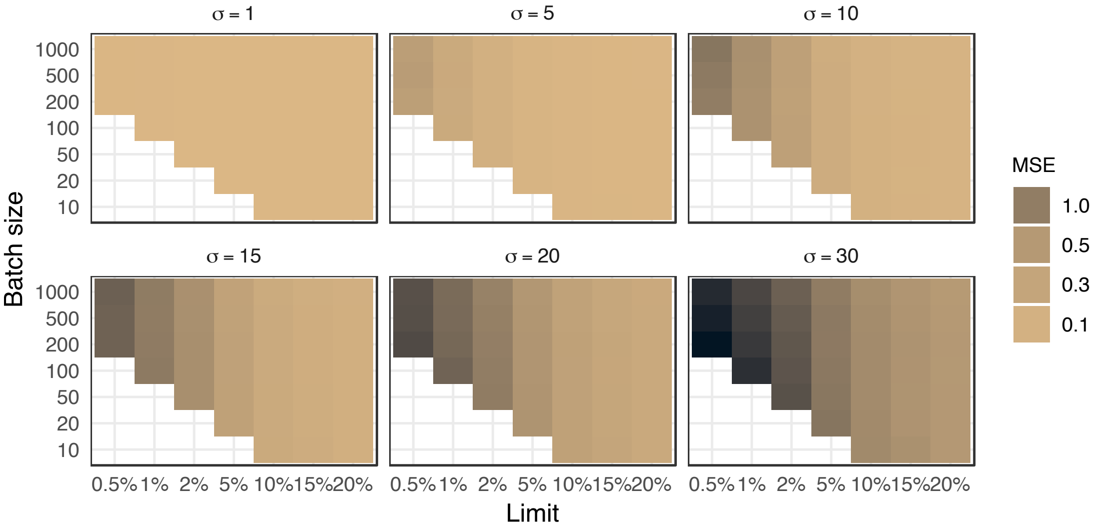

Figure 14: MSE of different combinations of n_B and L, for different levels of σ. Recall, MSE above the unit level means an error that is larger than what is measured.
Number of simulations = 10-50,000.

Generally, we can conclude to two practical recommendations:

1. Too quick adaptivity ($n_B < 50$) is never optimal: these strategies are dominated by other choices.

2. Unless the noise is really small ($\sigma = 1$), more limitation has practically no welfare cost while it improves MSE a lot.

5.2 Horizon

I also consider different lengths for the horizon[11]. 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

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

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 15 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) rule typical in the treatment choice literature.

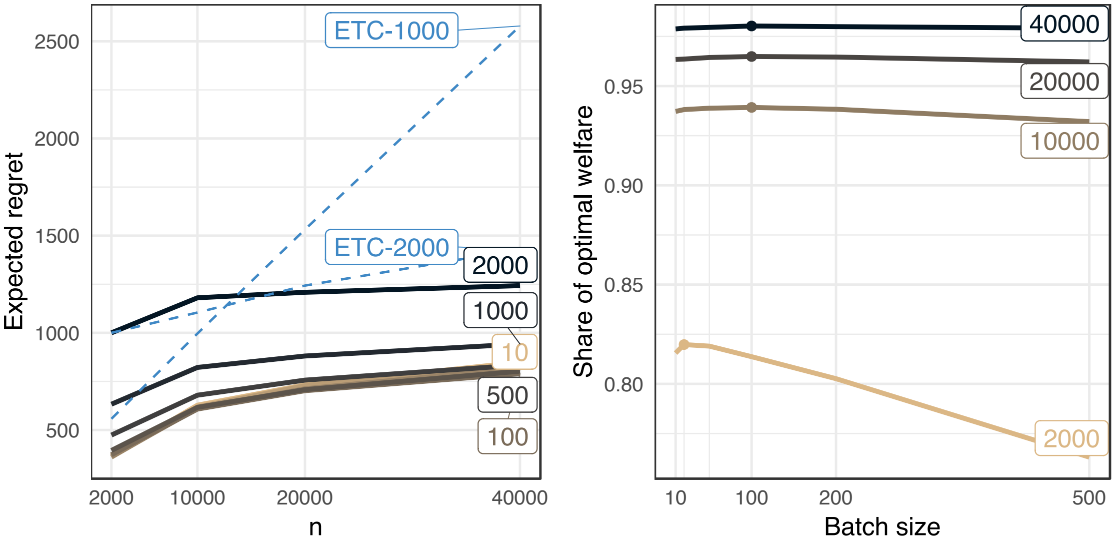

Figure 15: Welfare performance of bandit algorithm with various levels of adaptivity across different horizons. The regret of Thompson sampling grows slower with n than for the explore-then-commit rule common in the econometric practice. Longer horizons lessen the importance of the choice of batch size. 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 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 16 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 12.

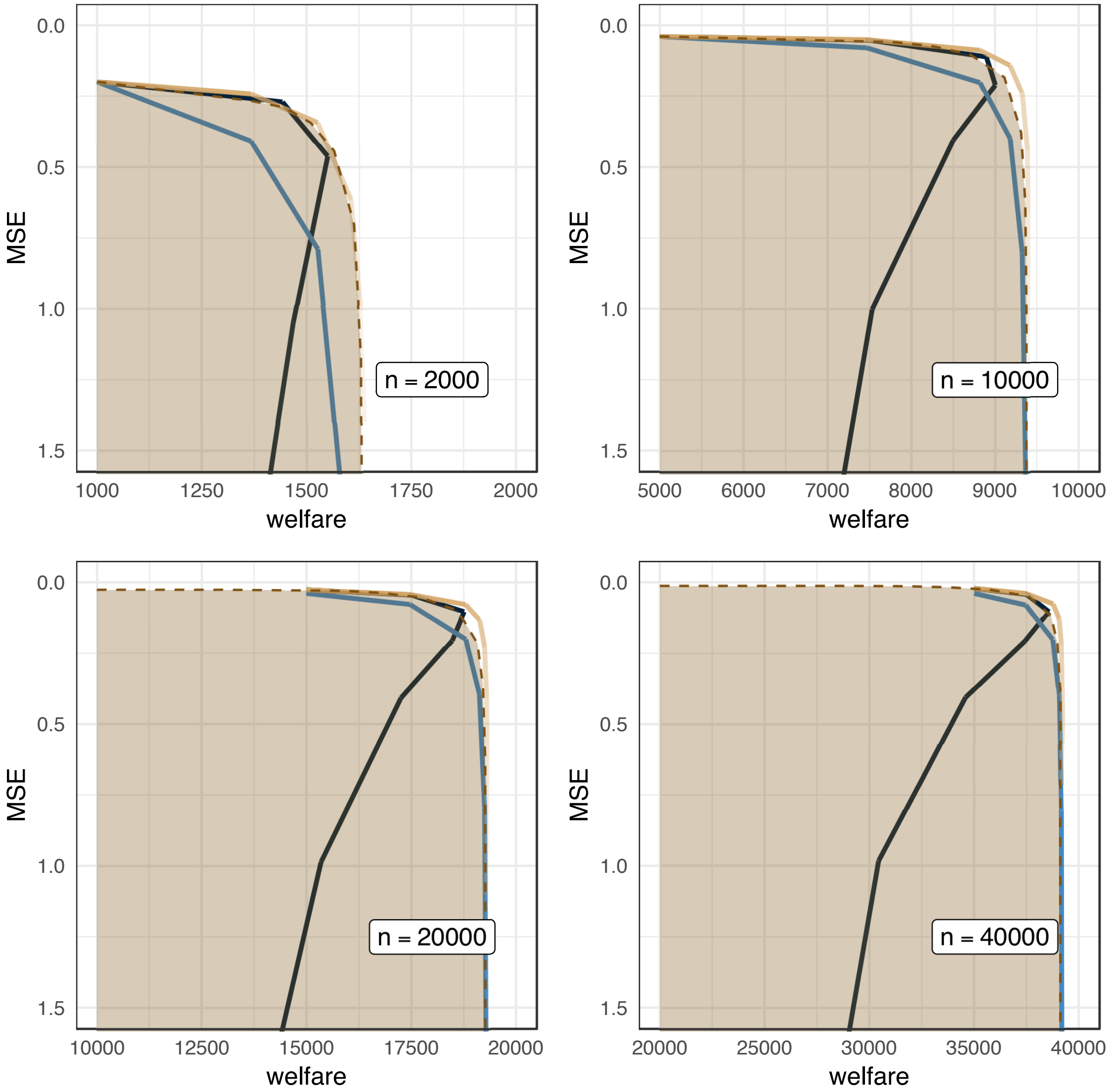

Figure 16: Performance of different strategies in the welfare-estimation space, for different horizons. LTS-IPW always extends the set of possible choices, especially if the problem is hard (n is small). Number of simulations = 10,000.

6 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 strategies (e.g. explore-then-commit, bandit with estimation on randomized subsample) form transitions between the two extremes, so they provide good choices for decision-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). Limiting more has small welfare cost while it can highly improve the precision of the estimator.

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

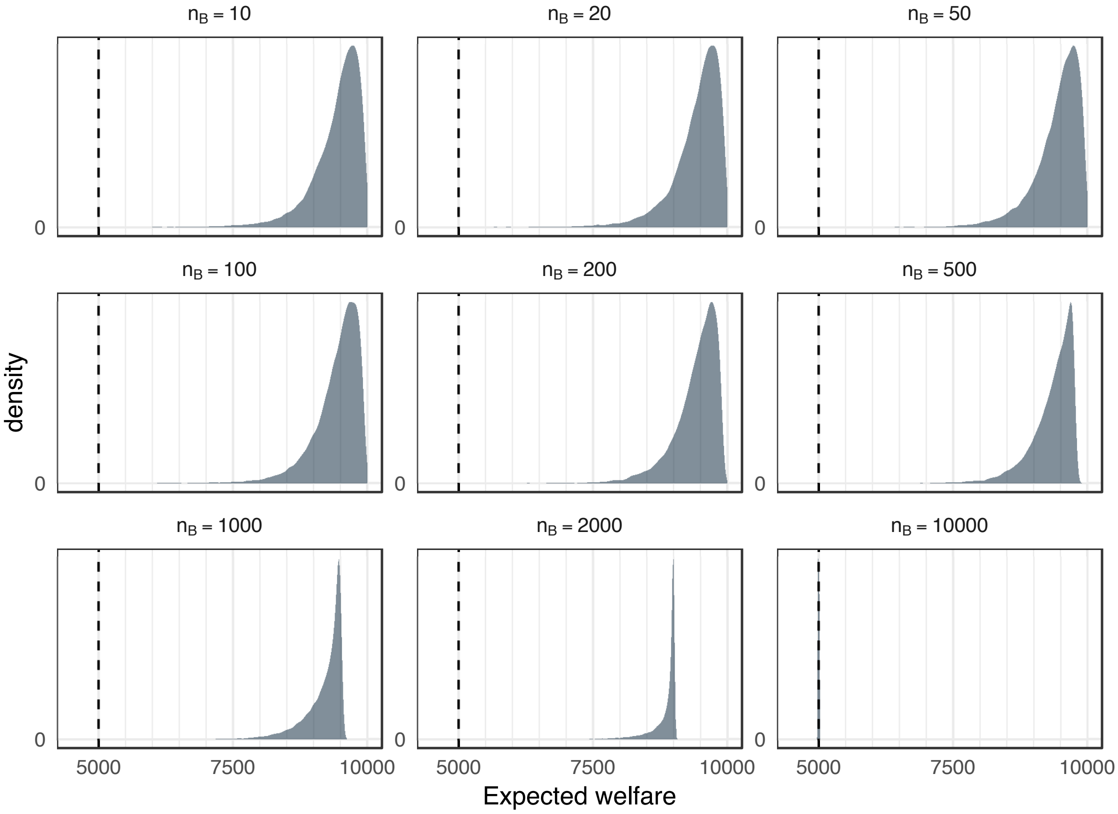

Figure A17: Distribution of expected welfare by batch size. Quicker adaptivity (smaller batch size) leads to higher achievable welfare but also higher variance.

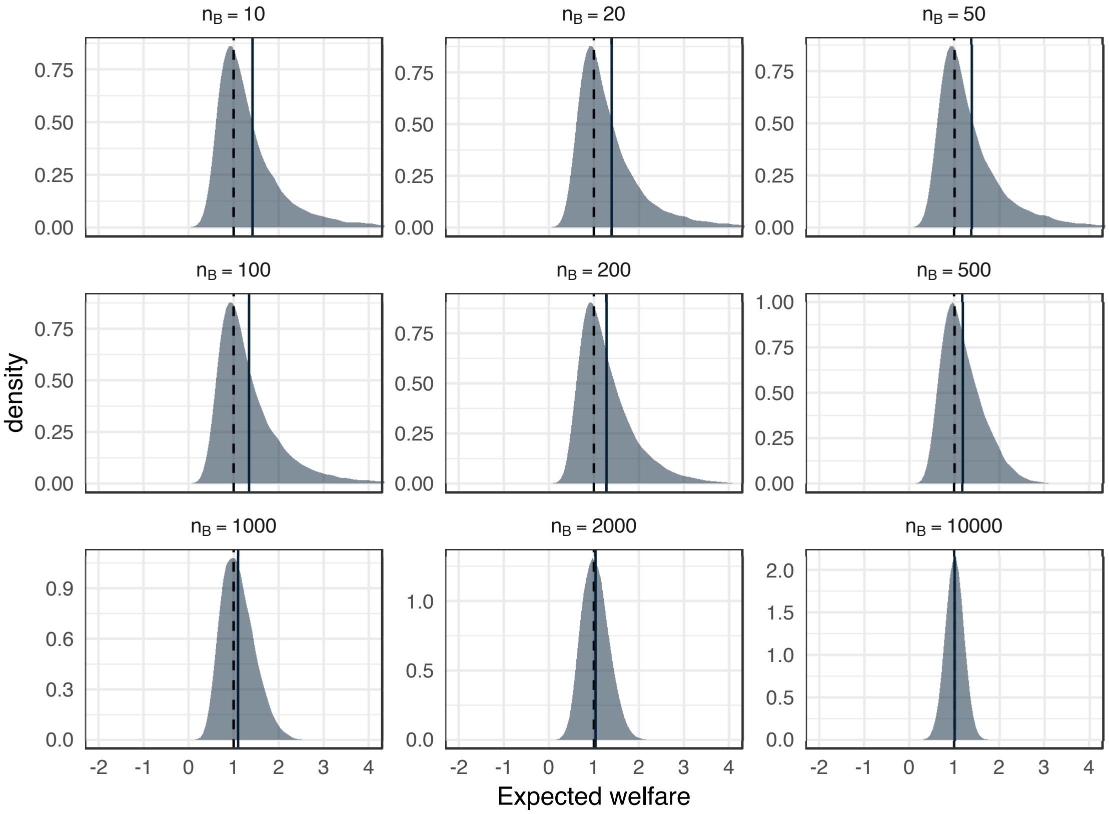

Figure A18: Distribution of $\hat{\tau}_0$ by batch size (dashed line: true treatment effect, solid line: expected value of the estimates). Quicker adaptivity (smaller batch size) leads to a more volatile estimate with larger bias.

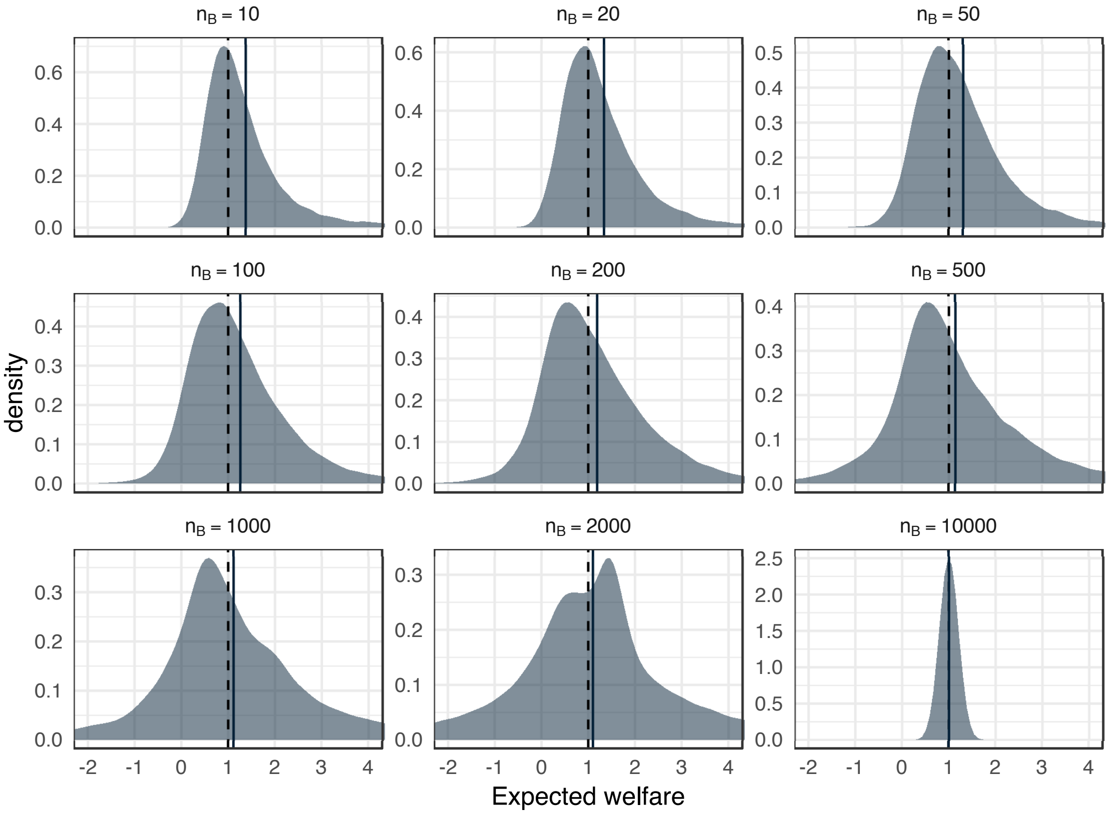

Figure A19: Distribution of $\hat{\tau}_{IPW}$ by batch size (dashed line: true treatment effect, solid line: expected value of the estimates). Quicker adaptivity (smaller batch size) leads to larger bias. The variance is larger compared to $\hat{\tau}_0$, especially for larger batch sizes.

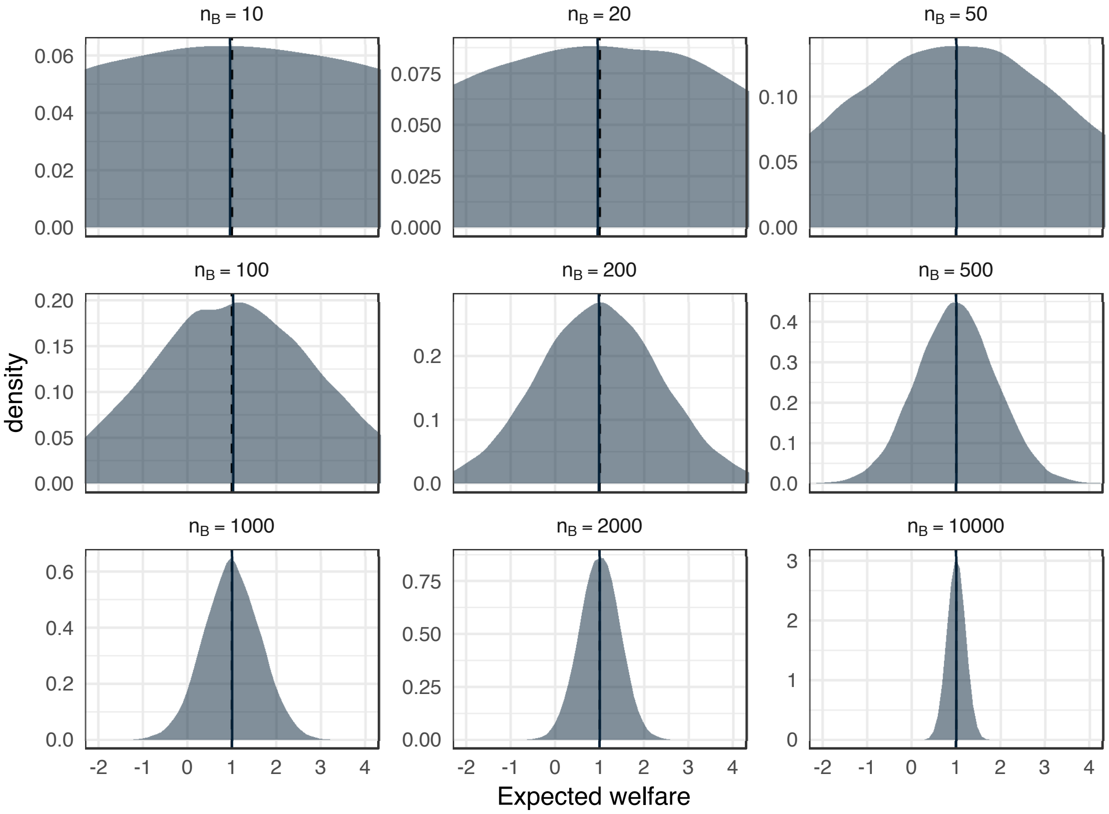

Figure A20: Distribution of $\hat{\tau}_{FB}$ by batch size (dashed line: true treatment effect, solid line: expected value of the estimates). The estimator is unbiased but really volatile, especially for smaller batch sizes.

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.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S. and Athey, S. (2019), Confidence Intervals for Policy Evaluation in Adaptive Experiments.
URL: *http://arxiv.org/abs/1911.02768*

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.