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

János K. Divényi †
Central European University

September 29, 2020

Abstract

It is well understood 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 has been much less 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 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 (IPW) 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 expected welfare and having a good estimate of 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, Miklós Koren, Sergey Lychagin, Arieda Muço, Jenő Pál, Sándor Sóvágó, Anthony Strittmatter, 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 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.

Recent papers (Agrawal and Goyal 2012, 2013, Korda et al. 2013) prove that Thompson sampling is asymptotically optimal in terms of regret in various settings. 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 welfare-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, if adaptivity is too quick adaptivity (the batch size is below a certain cutoff) the welfare cost of higher volatility 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 ef-

fect 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 assesses the simulation results in a practically relevant setting using data from the well-known National Job Training Partnership Act (JTPA) study. Section 7 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 (I show in Section 5.3 that the Gaussian assumption is only technical, the main results stand for skewed and fat-tailed distributions as well as long as they have finite means).

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.
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 wants to maximize profit (or welfare) based on outcomes, 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: $\pi\left(H^{(k)}\right) = \mathbb{P}\left(W_i = 1 | i \in B_{k+1}\right) = p_{k+1}$. The choice of assignment rule incorporates the choice of batch size as well: $n_B = |B_k|$.
2. **estimation method** A function that maps the whole history (observed data of the population) to a number that expresses the treatment effect: $\hat{\tau}\left(H^{(m)}\right)$.

I will call a combination of an assignment rule and an estimation method a **strategy**. The decision-maker chooses a strategy to pursue both of her goals. Throughout this paper I use two simple objective functions to measure these goals:

1. **welfare goal** $\max \sum_{i=1}^{n} Y_i$[2]

[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] Assuming the outcome contains the cost of treatment, it is the profit of a firm. Assuming a utilitarian social welfare function, it is the total welfare.

2. **estimation goal** $\min \mathbb{E}\left[(\hat{\tau}-\tau)^2\right]$ subject to $\mathbb{E}[\hat{\tau}]=\tau$ [3]

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[4]. I implement this rule as follows (for a chosen batch size):

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

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

3. 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] Recall the bias-variance decomposition: $\mathbb{E}\left[(\hat{\tau}-\tau)^2\right] = \left(\mathbb{E}\left[\hat{\tau}\right]-\tau\right)^2 + \mathbb{E}\left[\hat{\tau}^2\right] - \mathbb{E}^2\left[\hat{\tau}\right]$ where the last two terms give the variance of the estimator. So minimizing the mean-squared error is just minimizing the variance if the estimator is unbiased.

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

3 Demonstration 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.

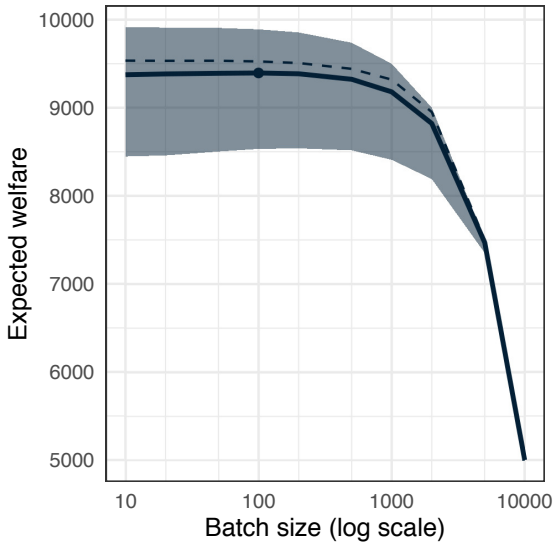
I run $20,000$ simulations for each assignment rule. The runs differ only in the sequence of how the individuals arrive; they all use the same population of 10,000 with the average of potential outcomes equaling to 0 and 1, respectively.

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.

However, the simulation results only partially justify this expectation. Figure 1 shows the expected welfare by batch size: 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 around $n_B = 100$. The reason for this is that being more adaptive means deciding based on more volatile estimates that increases the probability of adapting to the wrong pattern (in this case, "learning" a negative treatment

Figure 1: Expected welfare by batch size



Notes: The figure shows the expected welfare by batch size using a logarithmic scale to focus on the interesting region. The shaded area shows the 90% confidence interval, the dashed line depicts the median, the point highlights the batch size with maximum expected welfare. Smaller batches (quicker adaptivity) generally lead to higher welfare, but only until a certain point: really small batch size can harm. Number of simulations = 20,000.

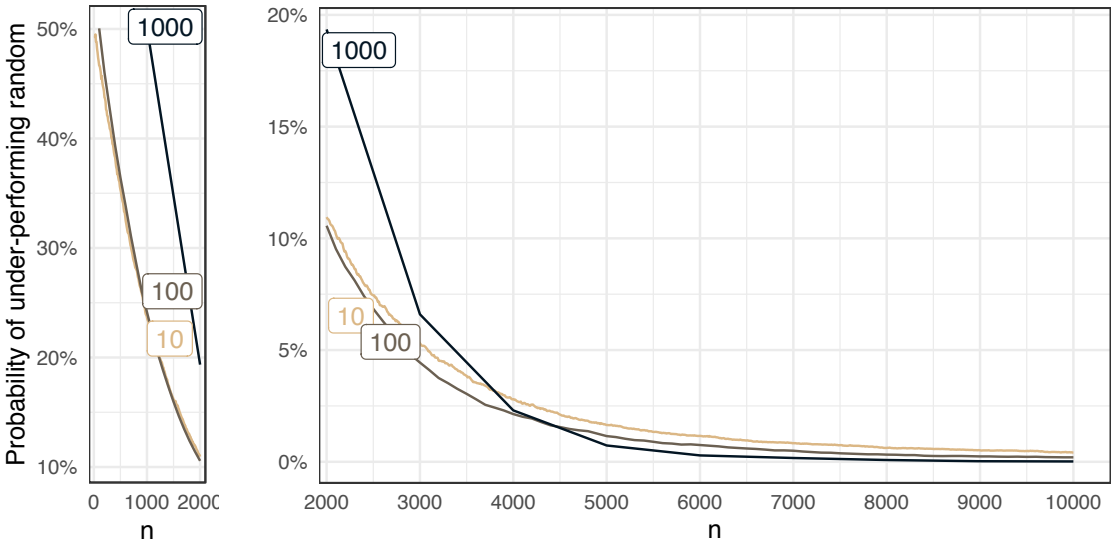
effect)[5]. Under a certain threshold of batch size, the loss on volatility seems to outweigh the gain on opportunity cost[6].

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.

[5]Figure A.20 in the Appendix shows the whole distribution of welfare for each batch size: the achieved welfare (that is equivalent to the number of individuals assigned to the treatment) is much more volatile for smaller batch sizes

[6]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.

Figure 2: Evolution of bandit algorithms



Notes: Each point depicts the probability that the bandit algorithm under-performs a simple random split after the first n arriving individuals (evaluated across the simulation runs). 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 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, avoiding too small batches helps only if the noise is high, or equivalently, if the treatment effect is small.

3.3 Estimation

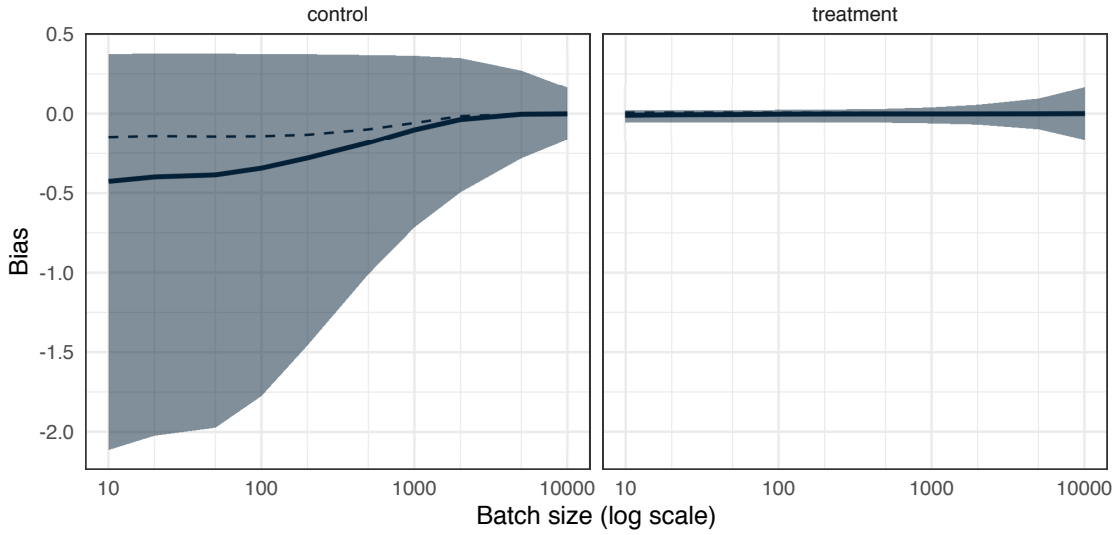
The standard method to estimate the treatment effect is to compare the observed averages of the individuals in both groups:

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

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 A.21 in Appendix shows the distribution of τ_0 for different batch sizes).

Figure 3: Bias in group mean estimates by batch size

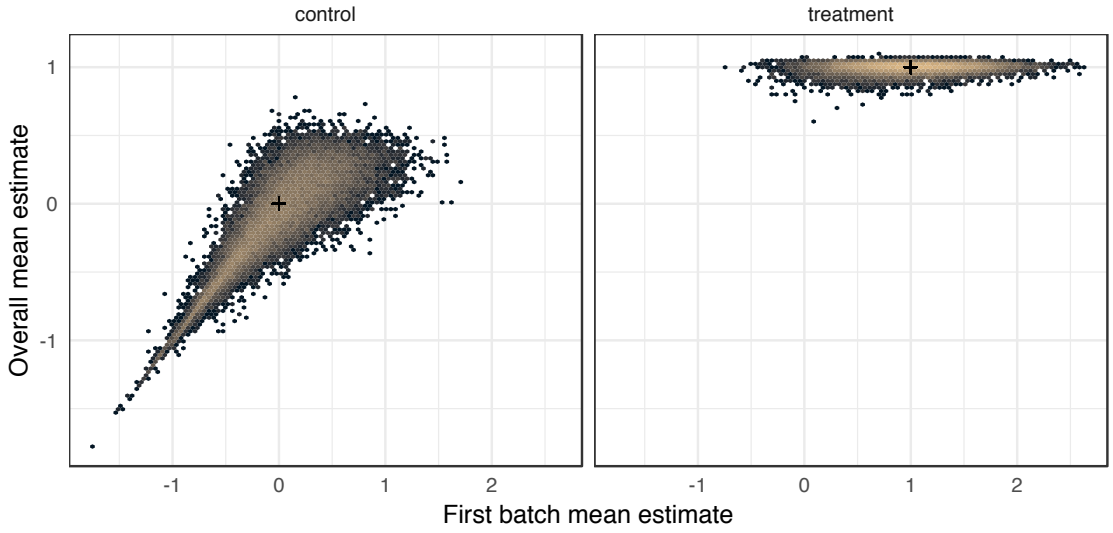


Notes: The figure shows the bias in the group mean estimates by batch size using a logarithmic scale to focus on the interesting region. The shaded area shows the 90% confidence interval, the dashed line depicts the median. Quicker adaptivity results in larger negative bias that is much more expressed for the control group (as we end up with more treatment observations). Number of simulations = 20,000.

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: $\mathbb{E}[\hat{\mu}_0^{(1)}] = \mu_0$. However, the actual estimate contains some estimation error: $\hat{\mu}_0^{(1)} = \mu_0 + \varepsilon_0^{(1)}$. If this error is negative – $\varepsilon_0^{(1)} < 0$ – 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 – $\varepsilon_0^{(1)} > 0$ – 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: Density of mean estimates, using the first batch versus the whole sample



Notes: First batch mean estimate is evaluated on individuals arriving in the first batch ($n_B = 1000$). 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 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.

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

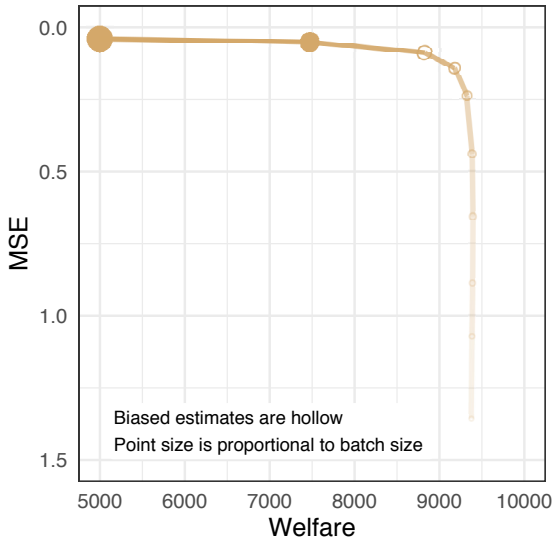
3.4 Welfare-Estimation Trade-off

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 of the estimator (reversed y axis), the two objective functions of the decision-maker (see 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 with zero MSE. Obviously, such a strategy does not exist.

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.

Figure 5: Performance of the bandit assignment rule in the welfare-estimation space



Notes: Each dot shows the achieved welfare and the mean squared error of the standard treatment effect estimator of the bandit assignment rule with a given batch size. Smaller batch size (quicker adaptivity) leads to higher welfare but also larger bias and larger MSE. Number of simulations = 20,000.

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[7] 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 $i \in B_j$.

[7]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.

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

However, IPW does not seem to be effective: instead of eliminating the bias, it can even exacerbate the problem (Figure A.22 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

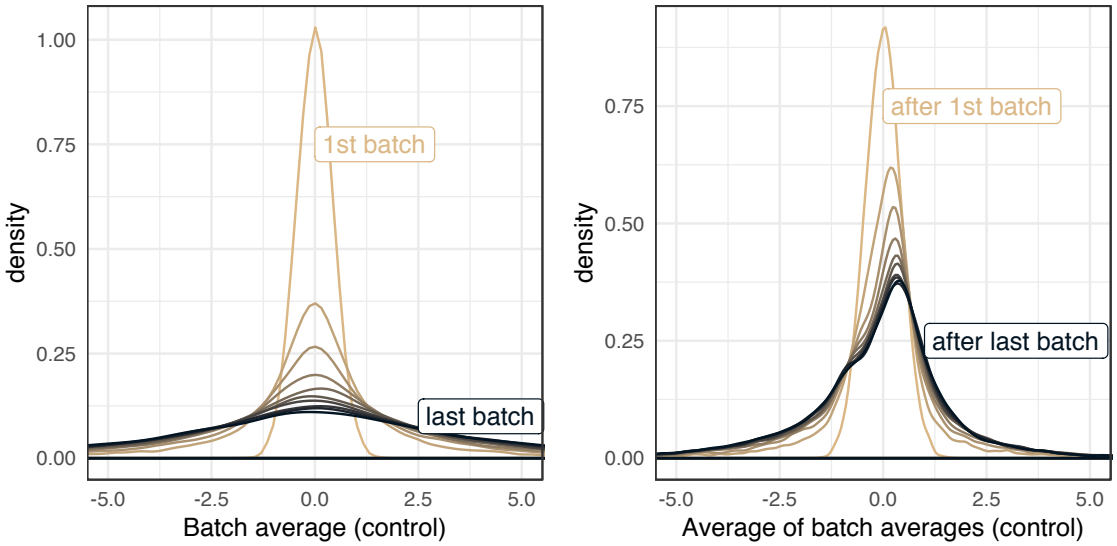
# of batches with controls	1	2	3	4	5	6	7	8	9	10
$\mathbb{E}[\hat{\tau}_{IPW}]$	1.99	1.85	1.76	1.79	1.71	1.46	1.64	1.32	1.22	0.80
Probability	2.0%	3.2%	3.5%	3.5%	3.8%	4.4%	5.2%	6.8%	11.4%	56.3%

Notes: 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 corre-

sponding 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 ($n_B = 1000$)



Notes: 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 ($\hat{\tau}_{FB}$):

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

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 A.23 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).

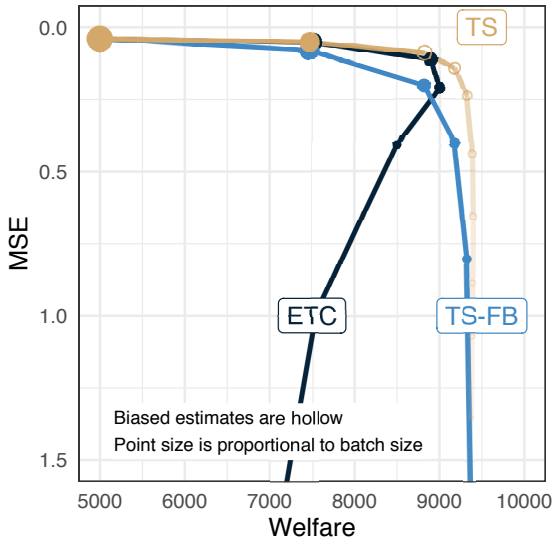
[b] Comparing the averages corresponds to the Conditional Empirical Success Rule of Manski (2004).

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

Figure 7: Performance of different strategies in the welfare-estimation space



Notes: Each dot shows the achieved welfare and the mean squared error of the corresponding treatment effect estimator for a given strategy with a given 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.

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 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[8]. As expected, higher limit means lower welfare and more precise $\hat{\tau}_{IPW}$ estimate[9].

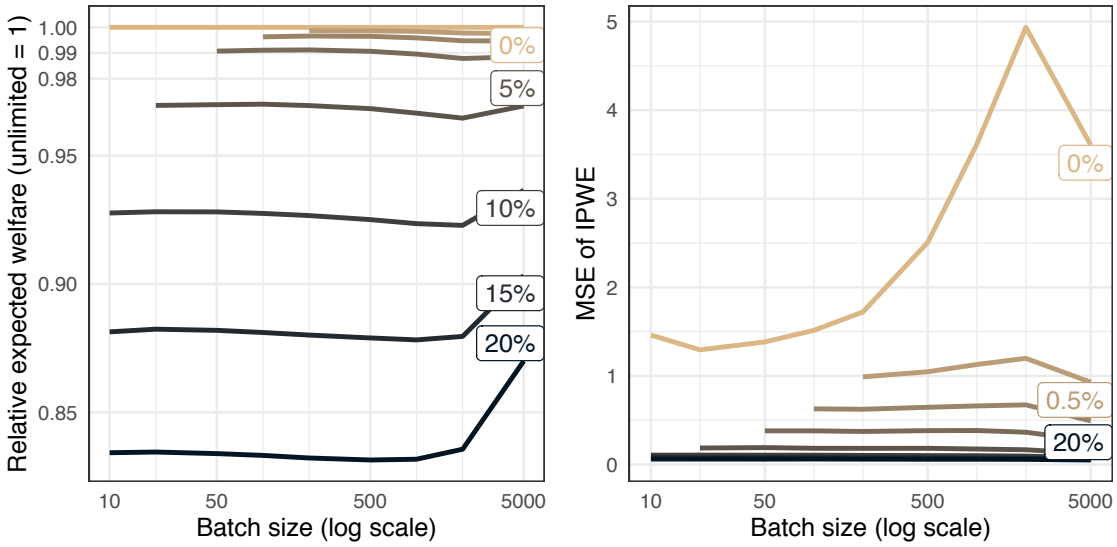
The loss in welfare and the gain in precision is disproportionate: while the loss is linear in the amount of limitation, the gain is not: using a 1% limit, MSE drops dramatically for each batch size (by as much as 80% for $n_B = 2000$ - see right panel) while it costs no more than 1% of welfare (left 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

[8] 0%, 0.5%, 1%, 2%, 5%, 10%, 15% and 20%.

[9] 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.

Figure 8: Welfare and estimation performance of the LTS-IPW strategy



Notes: The left panel shows the relative welfare achieved by the limited bandit rule compared to the unlimited one for various limit choices, by batch size. The right panel compares the MSE of the inverse-propensity-weighted estimators on the resulting data. Higher limits incur higher welfare cost but bring more precision. The loss and gain by the amount of limit are disproportionate. Number of simulations = 20,000.

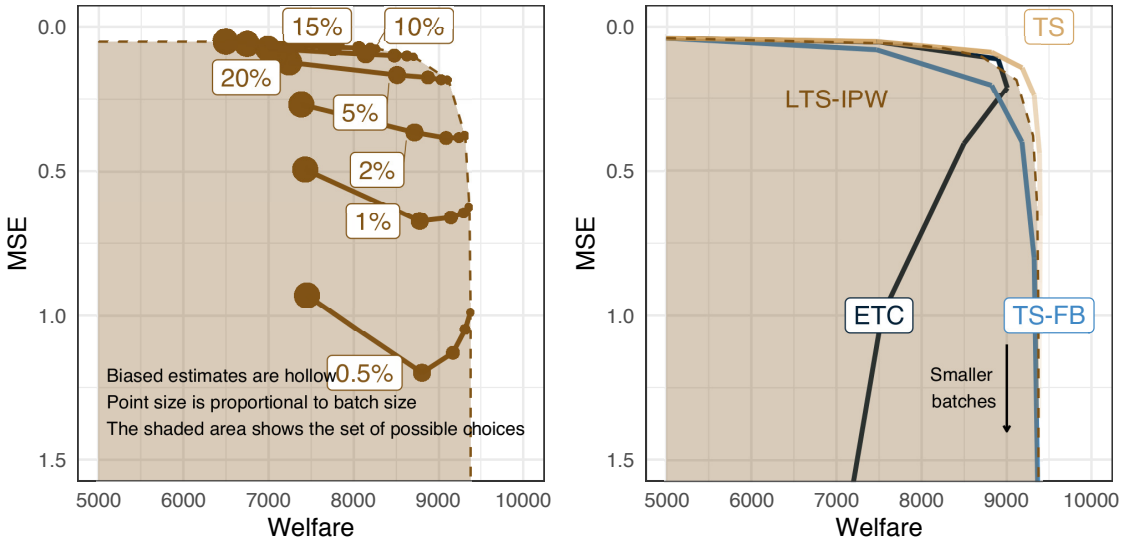
can thus improve welfare for some runs. Second, limitation obviously does not affect the simple random split of the first batch. For larger batch sizes, the share of the first batch is higher, thus, the limitation cost is relatively lower.

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 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



Notes: The right panel shows the achieved welfare and the MSE of the inverse-propensity-weighted estimator of the limited bandit rule, by various limits and batch sizes. The dashed line connects the best available choices (Performance Frontier). The left panel shows only this frontier compared to 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

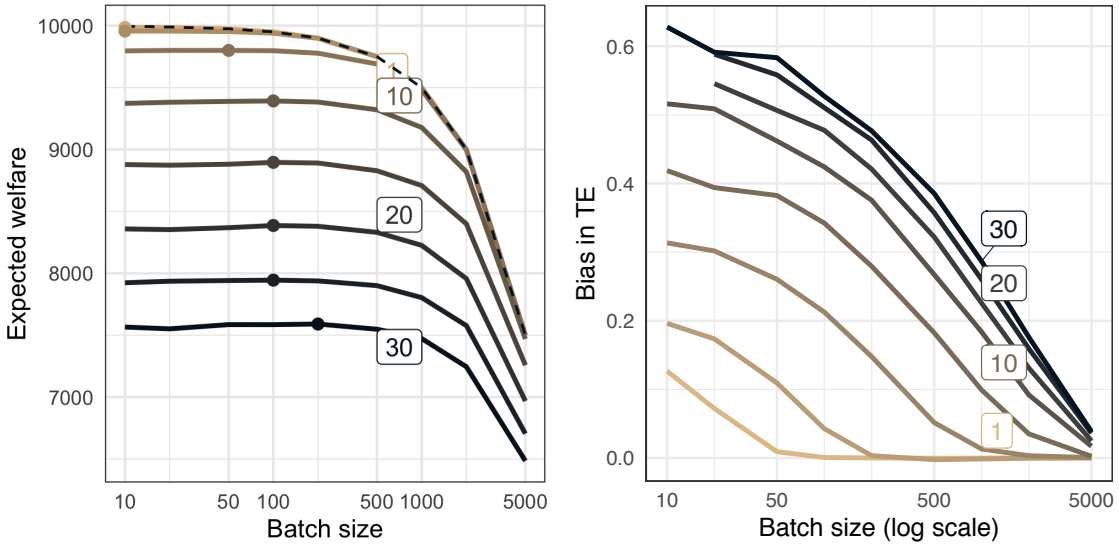
5.1 Uncertainty

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 $n = 10,000$[10]. Each setup is simulated with 10 values of batch size and 8 values of limit[11], $10 - 50$ thousand runs for each[12].

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.

Figure 10: Expected total welfare and bias



Notes: The left panel shows the expected welfare achieved by the (unlimited) bandit rule with different batch sizes (along the x axis) by different levels of noise (labelled). The dashed line highlights the maximum welfare that each strategy could achieve, and the points depict the batch sizes with the maximum welfare for a given σ. The right panel compares the bias in the standard treatment effect estimators. Larger noise results in lower welfare and larger bias. Number of simulations = 10-50,000.

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 over-fitting 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

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

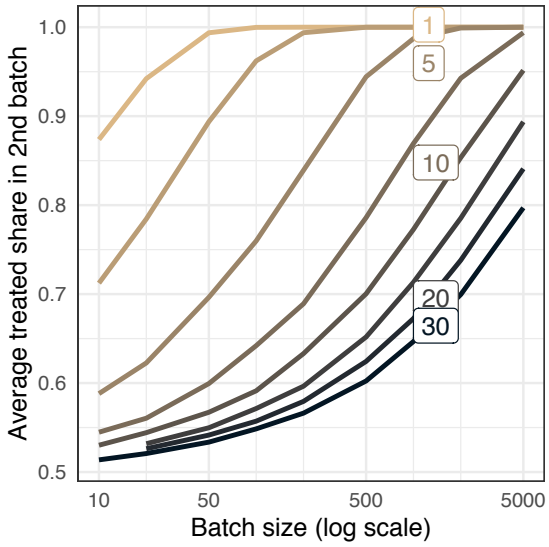
[11] As small batch sizes do not work with low limits, it means 63 parametrizations for each setup.

[12] 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.

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. Table A.1 in the Appendix contains the results for each scenario.

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. Table A.2 and A.3 in the Appendix contain the expected bias and MSE values for each scenario.

Figure 11: Average treated share in the second batch



Notes: The figure shows the expected share of individuals assigned to the treatment in the second batch for various batch sizes, under different noise levels. 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.

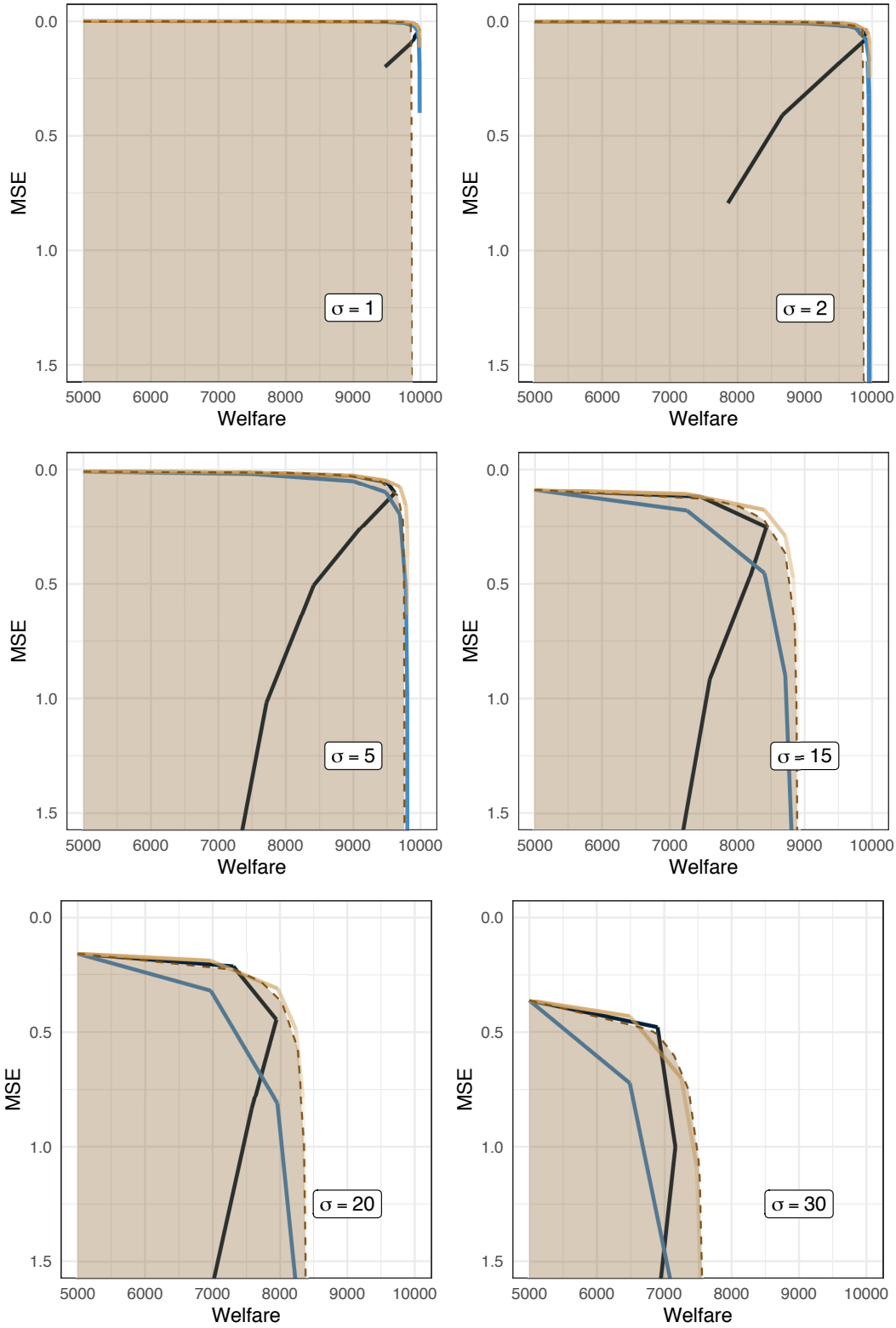
Welfare-Estimation Trade-off The previous results are in line with the main message of this paper: welfare and estimation goals are working against each other. Mainly, quicker

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 (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).

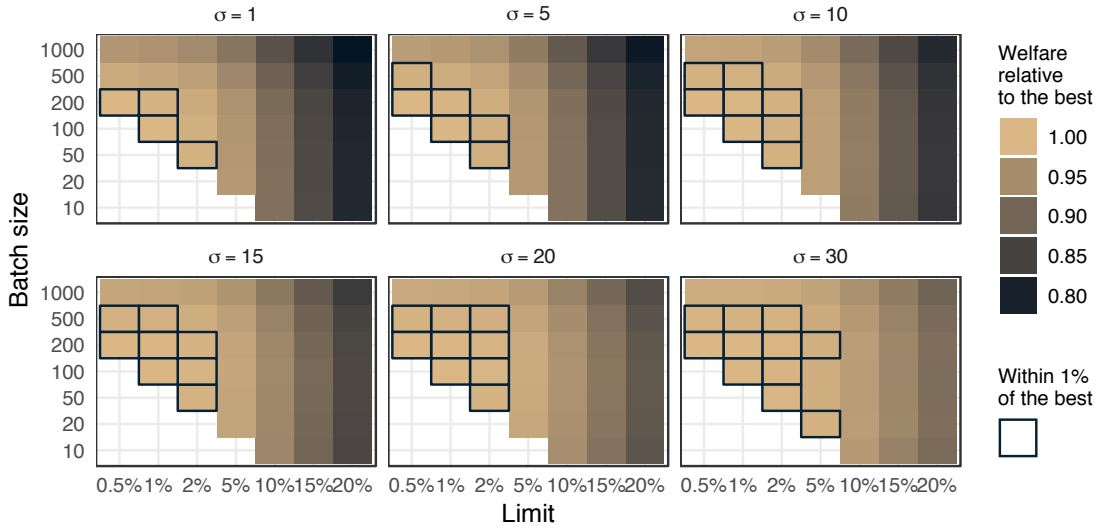
Figure 12: Performance of different strategies in the welfare-estimation space



Notes: Each panel is a replication of the left panel of Figure 9 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



Notes: Each panel shows the expected welfare relative to the best strategy for each batch size and limit combinations, for different levels of noise. The best strategies are highlighted within each scenario. 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.

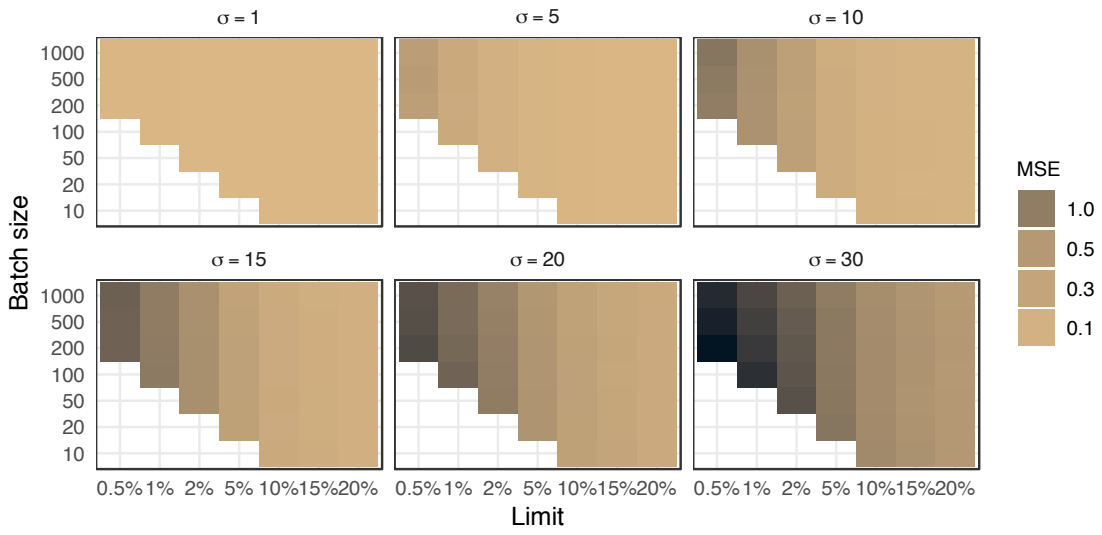
2. One can increase limit and decrease batch size to achieve about the same welfare. For large noise cases, many combinations result in the same level of welfare. Note, however, that this level is smaller in absolute value than in low-noise scenarios (recall Figure 1).

3. Limiting does not eliminate the problem of over-fitting: too quick adaptivity has a detrimental effect on expected welfare if the noise is high (e.g. for $\sigma = 20$ the achieved welfare is smaller with $n_B = 10$ than with $n_B = 50$ even with larger limits).

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 $\{n_B = 200, L = 5\%\}$ for $\sigma \geq 5$.

Figure 14: MSE of different combinations of n_B and L



Notes: Each panel shows the MSE of the inverse-propensity-weighted treatment effect estimator of the limited bandit for each batch size and limit combinations, for different levels of noise. Recall, MSE above the unit level means an error that is larger than what is measured. Number of simulations = 10-50,000.

To better understand the behavior of different strategies, it is worth considering the limiting cases of uncertainty:

1. **no-noise scenario** $\sigma \to 0$ For low-noise cases, smaller limits reach higher welfare while the MSE remains stable, so as $\sigma \to 0$ it is reasonable to $L \to 0$. Also recall that the standard treatment effect estimator on unlimited bandit data is unbiased for sufficiently large batches (see Figure 10), where the sufficiently large batch size

decreases in noise. Last, smaller batch size reaches higher welfare, and for low noise levels we should not worry about over-fitting either. All of these suggest that we should run an unlimited bandit with the smallest batch size ($n_B = 2$) for the no-noise scenario (the estimation method does not matter as $\sigma = 0 \Rightarrow \hat{\tau}_0 = \hat{\tau}_{IPW}$). This strategy is equivalent to the intuitive solution of this problem: assign one observation to both groups and then assign everyone based on the comparison of these outcomes.

2. **no-treatment-effect scenario** $\sigma \to \infty$ For high-noise cases, high limits are needed to keep MSE at moderate level. As noise increases, so decreases the achievable welfare (see Figure 10) and expands the set of batch size and limit choices that result in about the same welfare as the best combination. These suggest to use the maximum limit of 0.5 for the limiting no-treatment-effect scenario which strategy is equivalent to the simple random split. Again, this is an intuitive solution as zero treatment effect means there is no potential welfare to gain from being adaptive so it would only incur losses on the estimation goal.

Generally, we can conclude to following practical recommendations: choose the limit based on the welfare-MSE trade-off and then use the smallest possible batch size. This choice of the batch size gets less relevant as the noise increases.

5.2 Horizon

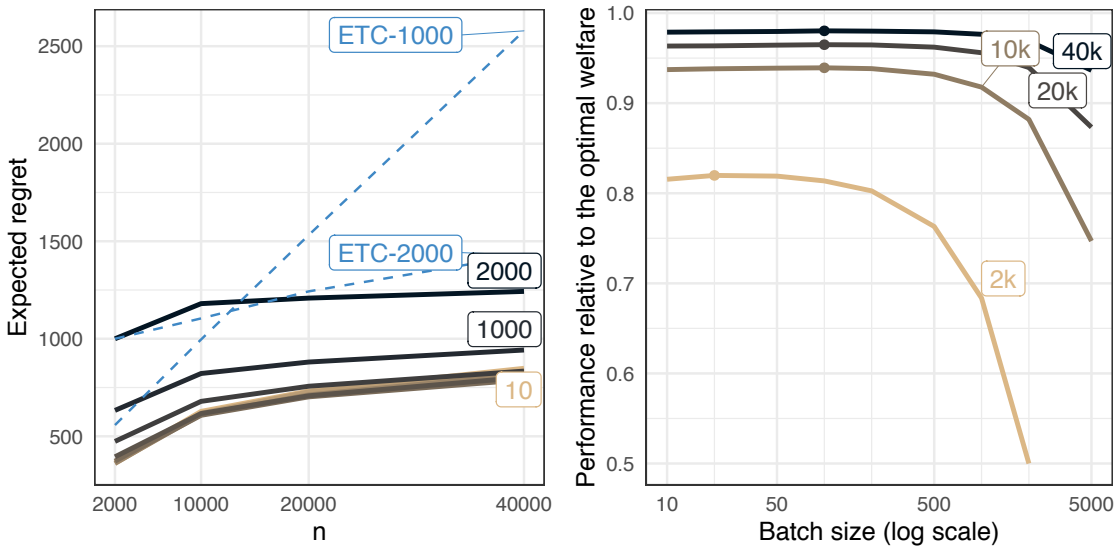
I also consider different lengths for the horizon[13]. Note that this is similar to changing the noise and batch size appropriately: e.g. a 4 times larger sample size is equivalent to a setup with 2 times larger σ with 4 times larger batches (e.g. holding the number of batches fixed). Simulating the illustrative case ($\sigma = 10$) for different lengths makes the comparison easier.

The right panel of Figure 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.

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

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

Figure 15: Welfare performance of bandit algorithm with various levels of adaptivity across different horizons

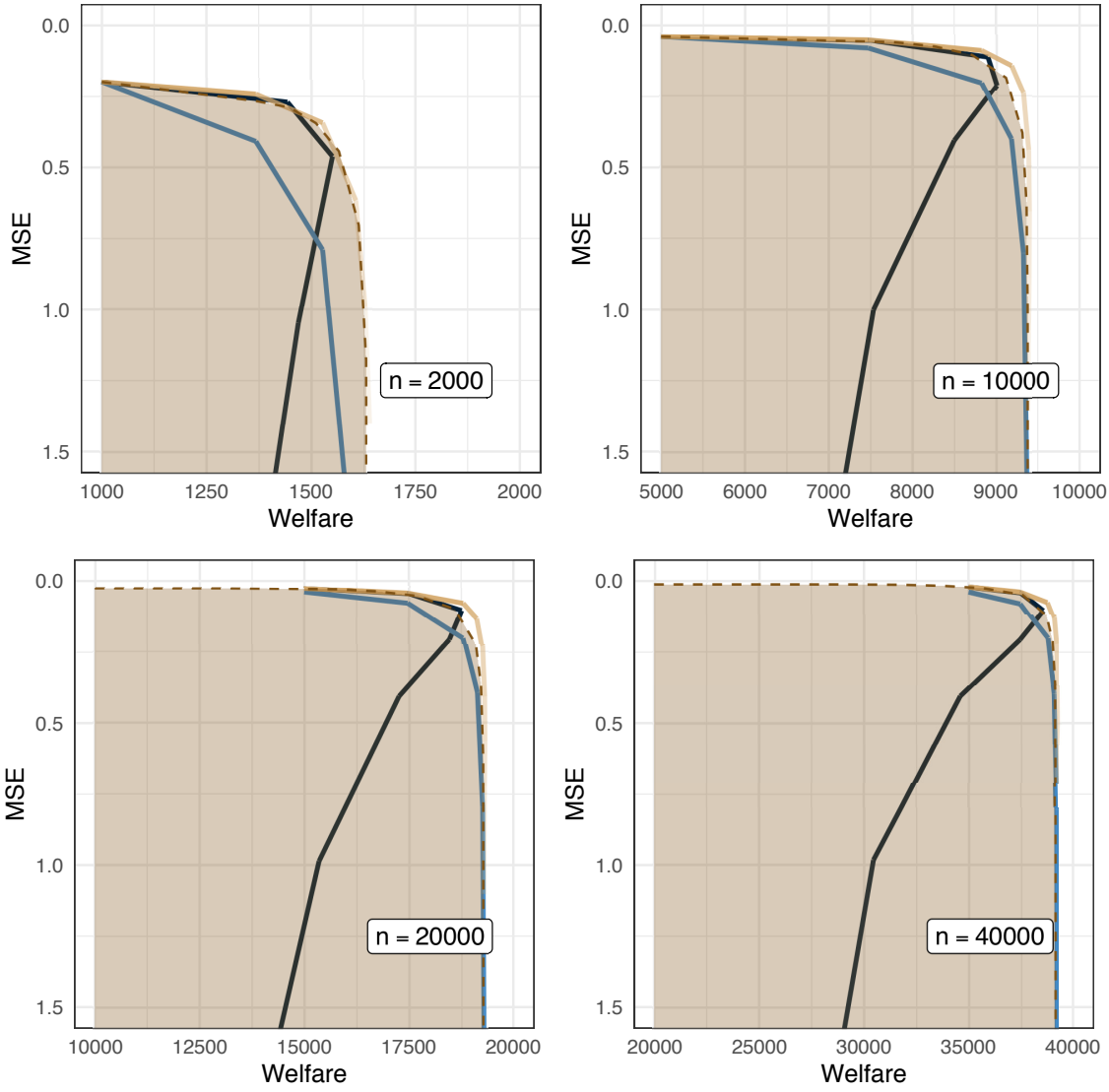


Notes: The left panel shows the expected regret of different strategies by the horizon: the regret of Thompson sampling grows slower with n than for the explore-then-commit rule common in the econometric practice. The right panel shows the expected welfare achieved by different strategies relative to the (infeasible) optimal welfare (treatment-only scenario). Longer horizons lessen the importance of the choice of batch size. Number of simulations = 10,000.

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 allocation 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. Table A.4, A.5 and A.6 in the Appendix contain the expected welfare, bias and MSE values for each scenario.

Figure 16: Performance of different strategies in the welfare-estimation space, for different horizons



Notes: Each panel is a replication of the left panel of Figure 9 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.

5.3 Non-Gaussian Potential Outcomes

All the previous results were built on the Gaussian assumption for the potential outcomes. In this subsection I show how relevant this assumption is by considering less well-behaved distributions as well. I focus on two common behavior: fat tails and skewness. I compare the behavior of the strategies by simulating untreated potential outcomes by four distributions:

1. Normal distribution
2. Student's t-distribution with 4 degrees of freedom (fat tails)
3. χ^2 distribution with 5 degrees of freedom (positive skewness)
4. negative χ^2 distribution with 5 degrees of freedom (negative skewness)

All of the simulated outcomes are standardized to have $\mu_0 = 0$ and $\sigma = 10$ in the population (as in the original setup, see Section 3.1) to allow for a strict comparison by the shape of the distribution.

For the comparison I choose two strategies: Thompson sampling with the standard treatment effect estimator ($\hat{\tau}_0$) and the limited Thompson sampling with 5% limit using the inverse-propensity-weighted estimator ($\hat{\tau}_{IPW}$)[14].

Figure 17 compares welfare performance of the strategies by distribution. The only difference can be detected in the TS strategy with quick adaptivity: the fat-tailed and the negatively-skewed distribution fare worse (but this difference is relatively small). The difference disappears with the limited strategy.

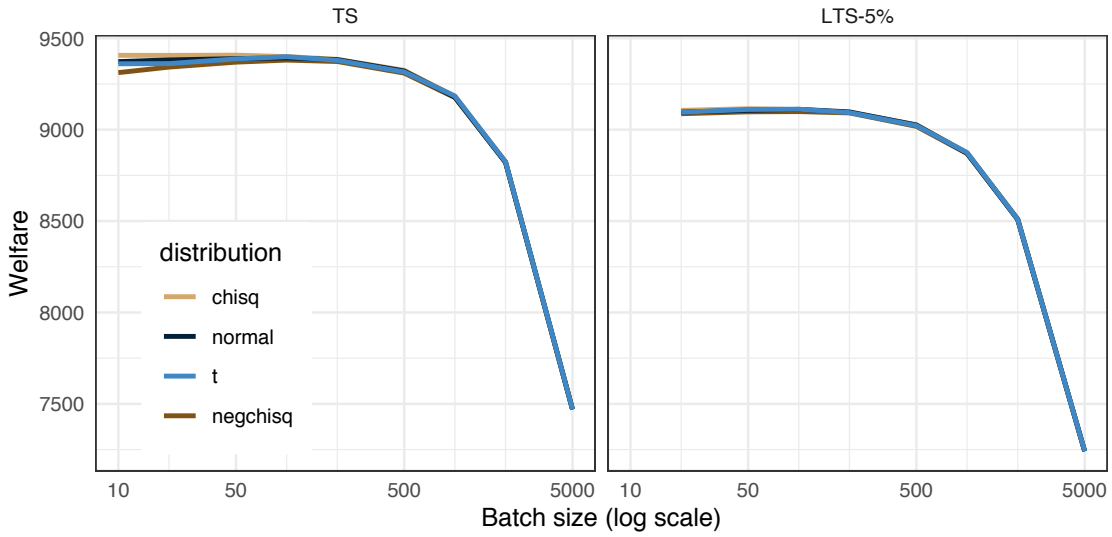
The expected welfare solely depends on each strategy's ability to assign as much individual to the best group (here: to the treatment) as possible. In adaptive allocation rules this ability is determined by how the estimated means compare to each other. As the assignment rule compares averages of the observed outcomes, for large enough sample size the central limit theorem kicks in and this makes the underlying distribution less relevant. In small sample cases, certain shapes of the underlying distribution makes the true means harder to estimate: if it has fat tails or a negative skewness. Interestingly, positive skewness seems to help.

Figure 18 shows the estimation performance of the same strategies using the standard treatment effect estimator on the unlimited bandit data and the inverse-propensity-weighted estimator on the limited bandit data. The general patterns are very similar to that of welfare. Practically, one could detect differences only for the TS strategy with small enough batch sizes. Fat-tailed and negatively skewed distributions of potential outcomes are worse, a positively skewed distribution is better than the standard normal one. However, all of these differences disappear when we apply limiting even if we want to preserve adaptivity.

The result that fat tails make our problem harder is intuitive. The differential result in skewness needs some explanation. As I discussed in Section 3.3, the bias originates from

[14]Note that I do not change the assignment mechanism, so the posterior beliefs about the group means are still formed using normal distributions. This better approximates a real-world situation where the exact distribution of the outcomes are not known.

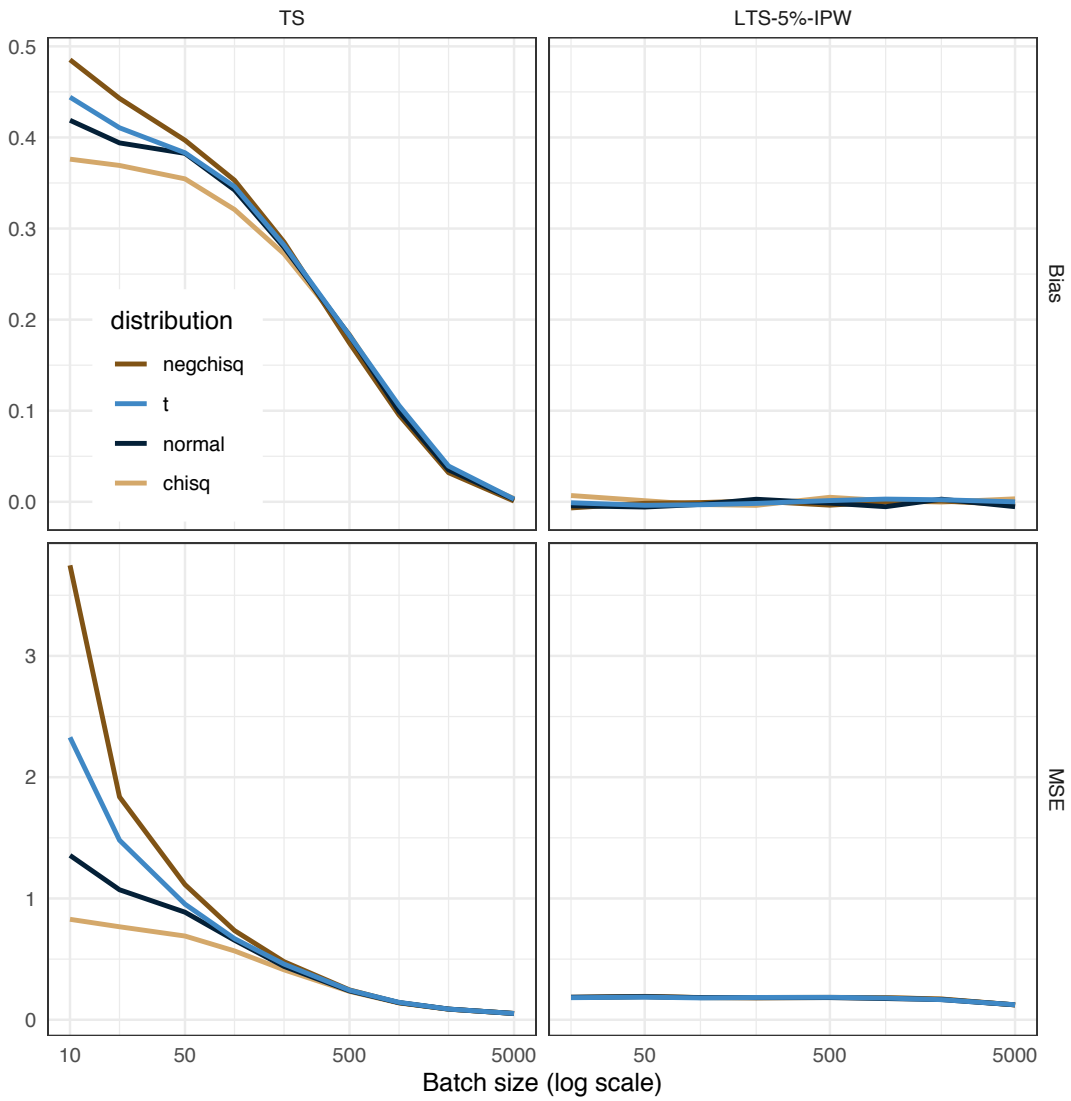
Figure 17: Welfare performance by different strategies compared by the distribution of the potential outcome



Notes: The figure shows the expected welfare of different strategies by batch size for each outcome distribution. TS: Thompson sampling, LTS-5%: limited Thompson sampling with 5% limit. There is no difference in the expected welfare by the distribution of the potential outcomes for the LTS strategy. Number of simulations = 10,000.

the belief about the control mean getting stuck a very low region. For this to happen we should draw from the low end of the distribution. If the distribution of the potential outcomes is such that drawing an observation negatively far from the mean has a higher probability, our problem gets harder. Negative skewness means that the mode is below the mean so the probability of drawing negative outliers is higher. In contrast, positive skewness brings more positive outliers that – due to the asymmetric sampling – only makes our problem easier. Obviously, if the treatment effect is negative, positive skewness is better.

Figure 18: Estimation performance by different strategies compared by the distribution of the potential outcome



Notes: The figure shows the expected bias and MSE of the estimator of different strategies by batch size for each outcome distribution. TS: Thompson sampling using $\hat{\tau}_0$, LTS-5%-IPW: limited Thompson sampling with 5% limit using $\hat{\tau}_{IPW}$. Number of simulations = 10,000.

6 Data-driven simulations

To assess the behavior of different strategies and the welfare-estimation trade-off in a practical setting, I will ran data-driven Monte Carlo simulations using the well known National Job Training Partnership Act (JTPA) study (Bloom et al. 1997). I take the experimental sample that was used by the influential paper of Abadie et al. (2002). This sample has been used many times for illustrative purposes in the treatment choice literature (see

among others Kitagawa and Tetenov 2017). Participants of the JTPA study assigned to the treatment group were offered job training. The outcome of interest it the earnings of the participants in the next 30-months period.

Table 2 shows the main numbers of the experiment. The program seems to be effective. The average earnings of the treatment group is $1,159 higher, even though only 64% of them actually got the training. This shows a positive intention-to-treat effect (ITT), that is my main interest here focusing on treatment assignment rules. The positive ITT more than compensates for the actual cost of the treatment, resulting in a net intention-to-treat effect of $674.

Table 2: Descriptive statistics of JTPA experiment

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

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

JTPA was a one-off experiment lasting for more than a year. For the sake of illustration, I will assume participants could have arrived in batches to simulate how adaptive assignment rules would have behaved with the JTPA-participants. Considering that such programs last over years this might be a relevant thought experiment: one can regard batches as yearly participants of such programs where each year's policy depends on available observations until that point[15].

For data-driven simulation, I relax the distributional assumption and the homogeneous treatment effect assumption. Instead, I simulate potential outcomes by bootstrapping from the available data. This way, I can simulate arbitrary assignments using the original data - as a result, it will not be true any more that only the arrival is random: each simulation run will consist of a different population (bootstrapped from the same original

[15]In such setup, the independent and identically distributed arrival is a very strong assumption: unemployed people in different years are likely to behave differently. Dimakopoulou et al. (2018) investigate the estimation problem in the exploration-exploitation framework in settings where the outcome is heterogenous by the arrival. They suggest a balancing method for contextual bandits to eliminate bias. In this paper I maintain the IID assumption to focus on the bias that arises even without any heterogeneity.

population). Besides this difference, the JTPA data could be translated to my setup by scaling the number of individuals and the treatment effect (average net ITT) corresponding to a scenario of $\sigma = 13.7$[16].

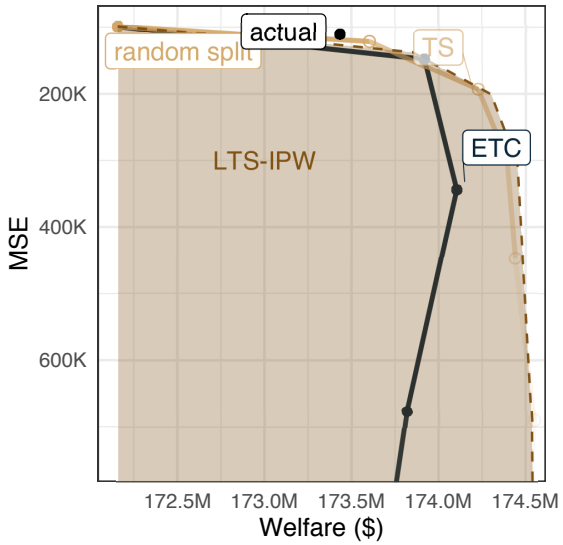
Figure 19 compares the performance of different strategies in the welfare-estimation space. As in the actual study 67% of the participants was assigned to the treatment and the net intention-to-treatment effect happened to be positive, the actual strategy fares very well[17]. However, the adaptive rules can win an additional welfare of as much as \$1M while still providing an unbiased estimate (in exchange for higher variance in the estimate). Furthermore, had the treatment effect happened to be negative, the actual strategy would have suffered a huge loss whereas the adaptive strategy could adapt to that scenario as well. Compared to a neutral 50-50% random split, the welfare gain of the adaptive strategy is much bigger and a large share of it could be realized without much loss on MSE.

The patterns of the figure are really similar to that of the $\sigma = 15$ case in Figure 12, only the uncertainty seems to be larger (limiting beats the unlimited strategy in welfare). This could result from the fact that the treatment effect is no longer constant and there is also variability in the population due to the bootstrap.

[16] First scale the outcome to have $\mu_0 = 0$ and $\mu_1 = 1$. Then scale the standard deviation of this scaled outcome by $\sqrt{10000/n}$.

[17] To assess the MSE of the actual assignment, I simulated a random split with a treatment share of 67%.

Figure 19: Welfare-estimation trade-off for the JTPA experiment



Notes: Each dot shows the achieved welfare and the mean squared error of the standard treatment effect estimator for a given strategy. The shaded are shows the available choices for the limited bandit strategy with inverse-propensity-weighted estimator for an appropriate batch size and limit. The dashed line connects the best possible combinations (Performance Frontier). The LTS-IPW strategy extends the available set of choices. 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 straightforward 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.

References

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

Agrawal, Shipra and Navin Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in "Journal of Machine Learning Research," Vol. 23 Microtome Publishing 2012.

_ **and** _ , "Further optimal regret bounds for thompson sampling," in "Journal of Machine Learning Research," Vol. 31 Microtome Publishing 2013, pp. 99–107.

Athey, Susan and Stefan Wager, "Efficient Policy Learning," 2019.

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

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

Dimakopoulou, Maria, Zhengyuan Zhou, Susan Athey, and Guido Imbens, "Estimation Considerations in Contextual Bandits," 2018.

Graepel, Thore, Joaqin Quinonero Candela, Thomas Borchert, and Ralf Herbrich, "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)" 2010, pp. 13–20.

Hadad, Vitor, David A. Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey, "Confidence Intervals for Policy Evaluation in Adaptive Experiments," 2019.

Hahn, Jinyong, Keisuke Hirano, and Dean Karlan, "Adaptive experimental design using the propensity score," *Journal of Business and Economic Statistics*, jan 2011, *29* (1), 96–108.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning* Springer Series in Statistics, New York, NY, USA: Springer New York Inc., 2001.

Hirano, Keisuke and Jack R. Porter, "Asymptotics for Statistical Treatment Rules," *Econometrica*, 2009, *77* (5), 1683–1701.

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

_ **and Anja Sautmann**, "Adaptive Experiments for Policy Choice," 2019.

Kitagawa, Toru and Aleksey Tetenov, "Equality-minded treatment choice," 2017.

_ **and** _ , "Who should be treated? Empirical welfare maximization methods for treatment choice," *Econometrica*, 2018, *86* (2), 591–616.

Korda, Nathaniel, Emilie Kaufmann, and Remi Munos, "Thompson Sampling for 1-Dimensional Exponential Family Bandits," in "Advances in Neural Information Processing Systems 26 (NIPS)" 2013, pp. 1448–1456.

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

Lattimore, Tor and Csaba Szepesvári, *Bandit Algorithms*, Cambridge University Press, 2019.

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

Nie, Xinkun, Xiaoying Tian, Jonathan Taylor, and James Zou, "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)" 2018.

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

Russo, Daniel, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen, "A Tutorial on Thompson Sampling," *Foundations and Trends® in Machine Learning*, 2017, *11* (11), 1–96.

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

Slivkins, Aleksandrs, *Introduction to Multi-Armed Bandits* 2019.

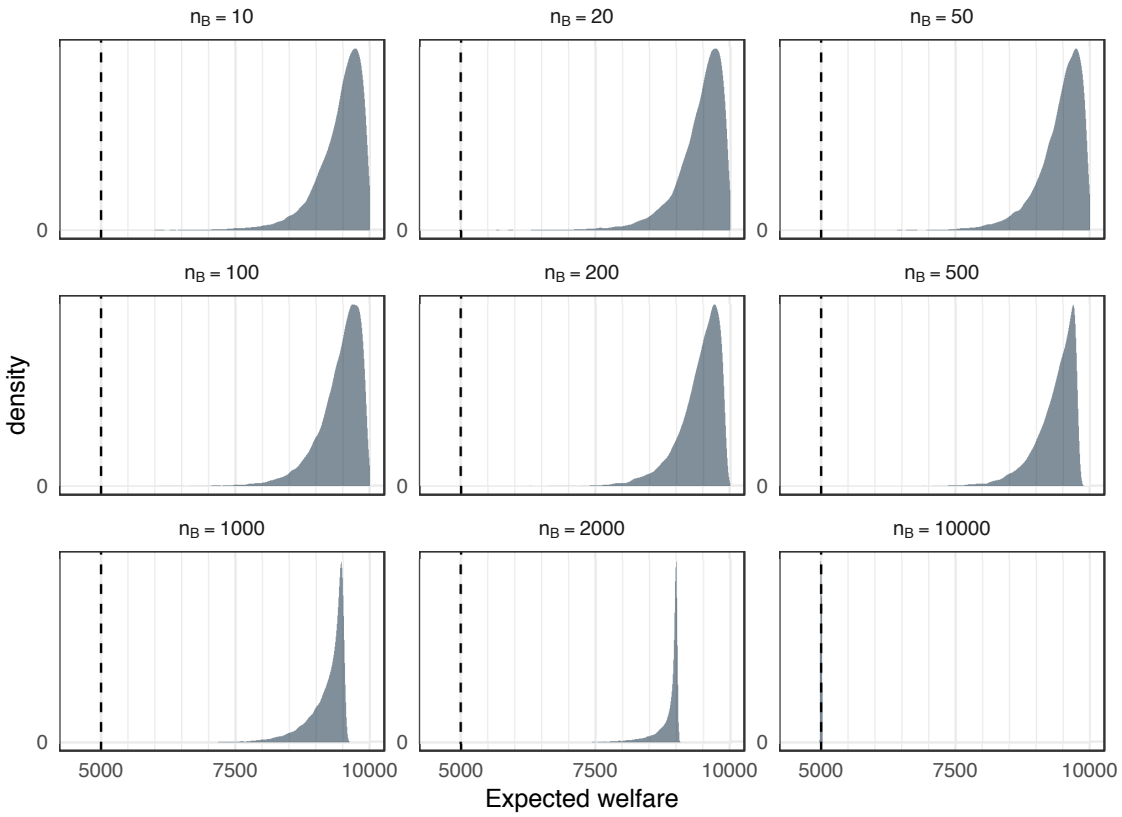
Thompson, William R., "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples," *Biometrika*, 1933, *25* (3-4), 285–294.

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

A Simulation distributions

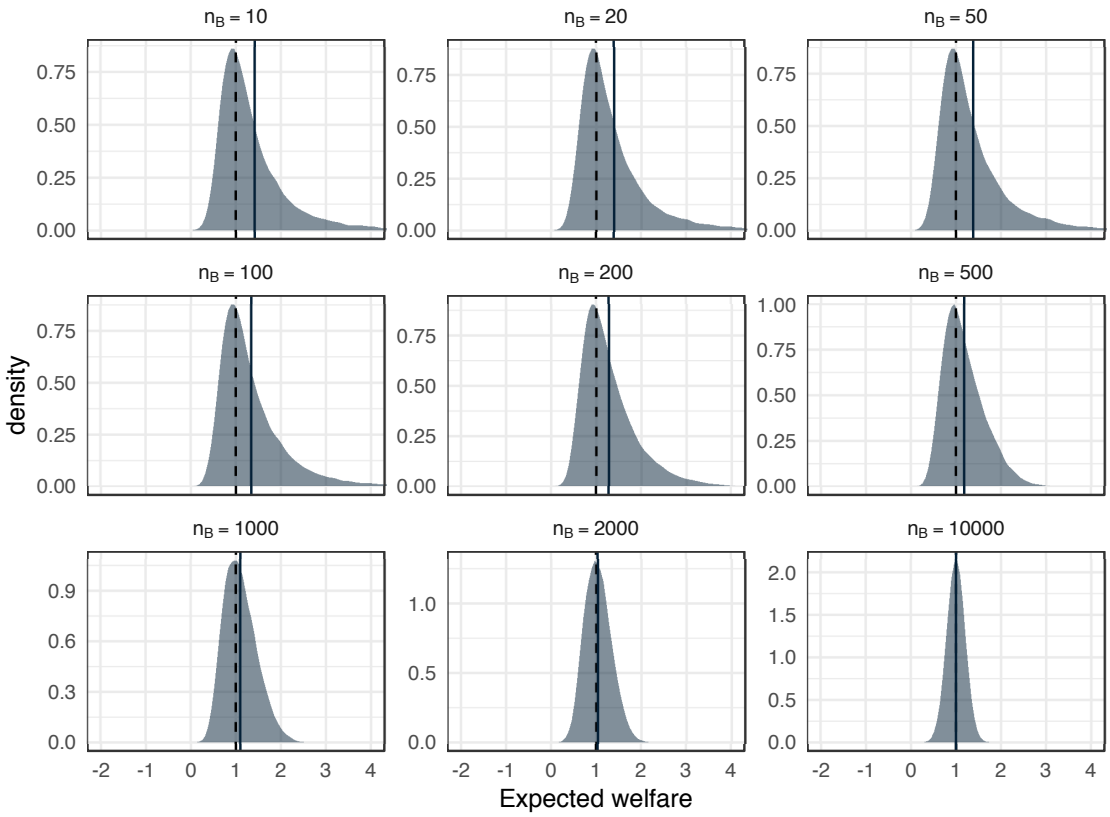
This section presents the whole simulation distributions for expected welfare and various estimators to complement the summary numbers in the main text.

Figure A.20: Distribution of welfare by batch size



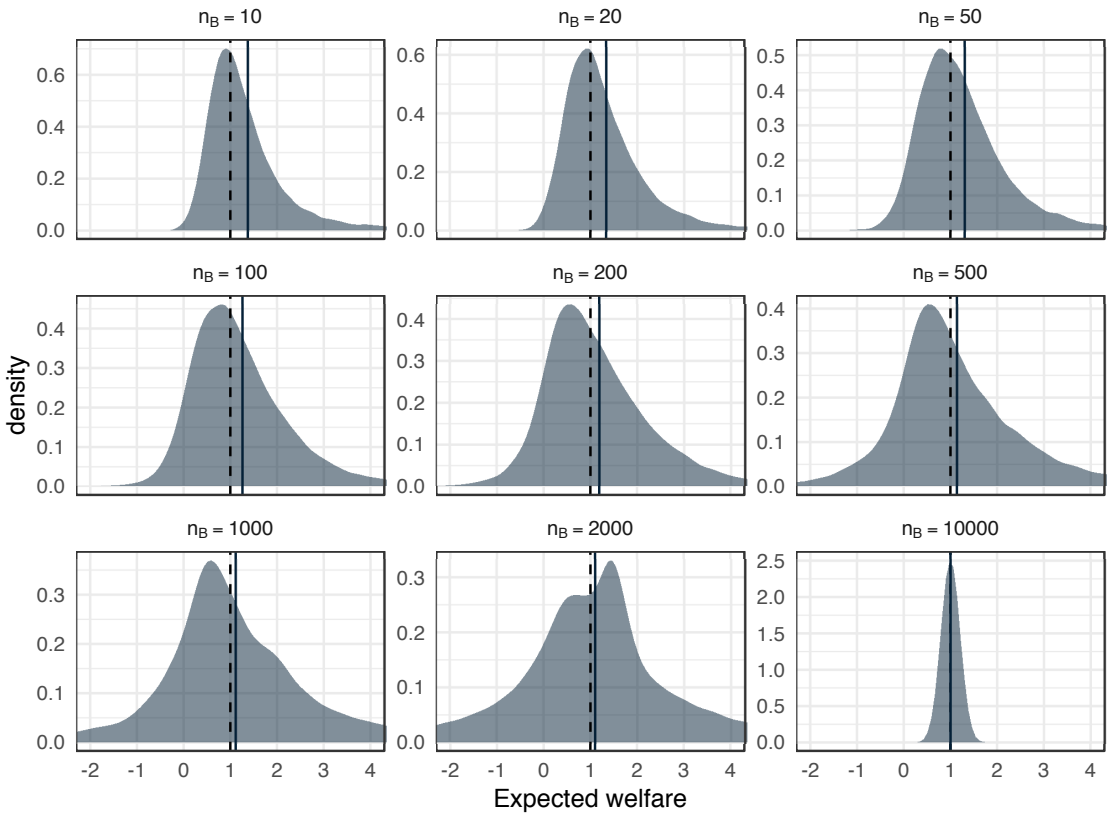
Notes: Each panel shows the distribution of the achieved welfare by the bandit algorithm with the corresponding batch size ($\sigma = 10$). Quicker adaptivity (smaller batch size) leads to higher achievable welfare but also higher variance.

Figure A.21: Distribution of $\hat{\tau}_0$ by batch size



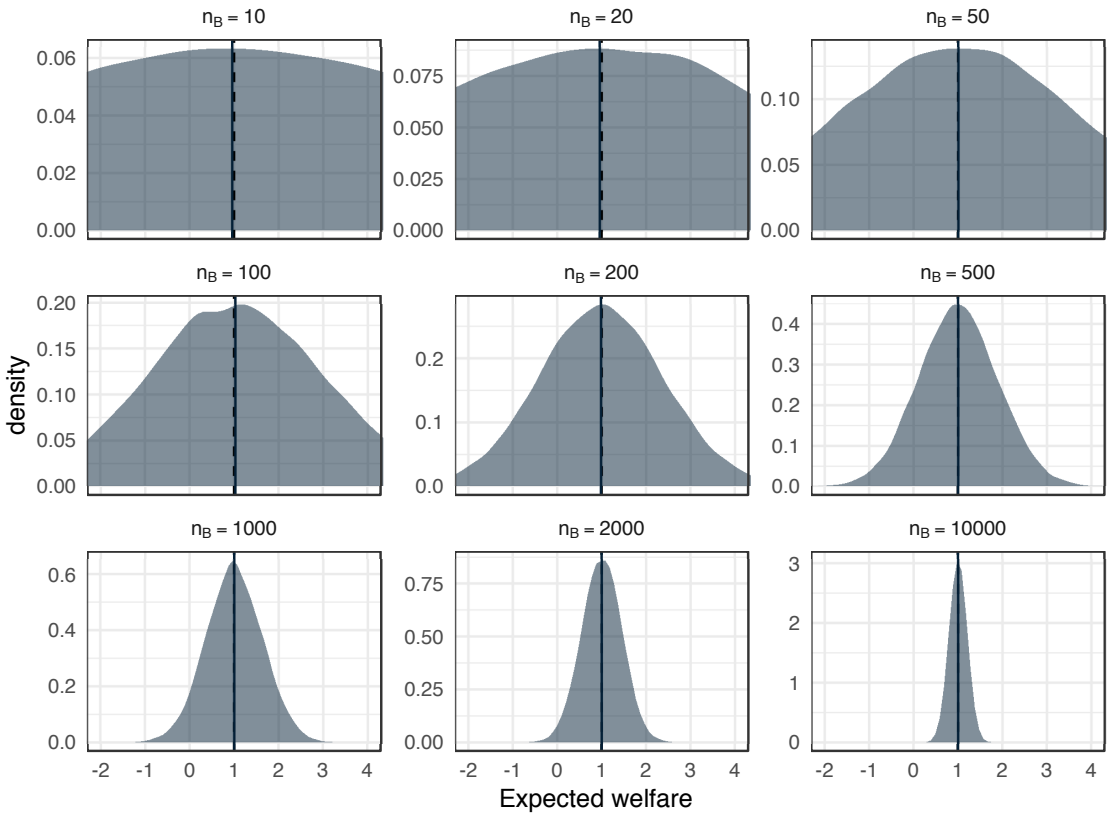
Notes: Each panel shows the distribution of the standard treatment effect estimate for the bandit algorithm with the corresponding batch size ($\sigma = 10$). The dashed line shows the true treatment effect, while the solid line corresponds to the expected value of the estimates. Quicker adaptivity (smaller batch size) leads to a more volatile estimate with larger bias.

Figure A.22: Distribution of $\hat{\tau}_{IPW}$ by batch size



Notes: Each panel shows the distribution of the inverse-propensity-weigthed treatment effect estimate for the bandit algorithm with the corresponding batch size ($\sigma = 10$). The dashed line shows the true treatment effect, while the solid line corresponds to the 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 A.23: Distribution of $\hat{\tau}_{FB}$ by batch size



Notes: Each panel shows the distribution of the treatment effect estimate calculated on the first batch of the bandit algorithm with the corresponding batch size ($\sigma = 10$). The dashed line shows the true treatment effect, while the solid line corresponds to the expected value of the estimates. The estimator is unbiased but really volatile, especially for smaller batch sizes.

B Detailed simulation results

This section presents all the simulation results of expected welfare, expected bias and expected welfare for the various strategies in different setups. The welfare-estimation plots in the main text are based on these numbers.

Table A.1: Expected welfare for different strategies ($n = 10,000$)

	Batch size									
allocation	10	20	50	100	200	500	1000	2000	5000	10,000
$\sigma = 1$										
TS	9987	9985	9974	9950	9900	9750	9500	9000	7500	5000
ETC	9465	9847	9973	9950	9900	9750	9500	9000	7500	5000
LTS-0%					9851	9702	9455	8960	7475	5000
LTS-1%				9851	9802	9655	9410	8920	7450	5000
LTS-2%			9776	9752	9704	9560	9320	8840	7400	5000
LTS-5%		9490	9477	9455	9410	9275	9050	8600	7250	5000
LTS-10%	8995	8991	8980	8960	8920	8800	8600	8200	7000	5000
LTS-15%	8495	8493	8483	8465	8430	8325	8150	7800	6750	5000
LTS-20%	7996	7994	7985	7970	7940	7850	7700	7400	6500	5000
$\sigma = 2$										
TS	9957	9957	9953	9940	9898	9750	9500	9000	7500	5000
ETC	7858	8663	9580	9892	9899	9750	9500	9000	7500	5000
LTS-0%					9850	9702	9455	8960	7475	5000
LTS-1%				9846	9801	9655	9410	8920	7450	5000
LTS-2%			9767	9748	9703	9560	9320	8840	7400	5000
LTS-5%		9479	9471	9452	9410	9275	9050	8600	7250	5000
LTS-10%	8986	8984	8976	8958	8920	8800	8600	8200	7000	5000
LTS-15%	8489	8487	8479	8464	8430	8325	8150	7800	6750	5000
LTS-20%	7991	7990	7983	7969	7940	7850	7700	7400	6500	5000
$\sigma = 5$										
TS	9797	9800	9801	9798	9778	9691	9482	8998	7500	5000
ETC	6186	6710	7713	8412	9098	9615	9494	9000	7500	5000
LTS-0%					9750	9656	9442	8959	7475	5000
LTS-1%				9736	9714	9615	9399	8919	7450	5000
LTS-2%			9660	9656	9630	9527	9311	8839	7400	5000
LTS-5%		9395	9394	9386	9357	9252	9044	8600	7250	5000
LTS-10%	8922	8925	8921	8912	8883	8785	8597	8200	7000	5000
LTS-15%	8439	8442	8438	8429	8403	8315	8148	7800	6750	5000
LTS-20%	7954	7955	7951	7943	7920	7843	7699	7400	6500	5000
$\sigma = 10$										

Table A.1: Expected welfare for different strategies ($n = 10,000$) *(continued)*

	Batch size									
allocation	10	20	50	100	200	500	1000	2000	5000	10,000
TS	9372	9382	9389	9392	9383	9321	9178	8820	7469	5000
ETC	5626	5840	6375	6946	7533	8496	9004	8896	7498	5000
LTS-0%					9371	9308	9163	8800	7451	5000
LTS-1%				9358	9351	9288	9140	8774	7430	5000
LTS-2%			9302	9309	9301	9234	9082	8713	7384	5000
LTS-5%		9096	9106	9111	9097	9026	8871	8508	7241	5000
LTS-10%	8694	8707	8713	8711	8695	8623	8476	8139	6995	5000
LTS-15%	8260	8278	8280	8276	8258	8193	8060	7758	6747	5000
LTS-20%	7819	7829	7829	7826	7809	7750	7634	7370	6498	5000
$\sigma = 15$										
TS	8878	8873	8881	8896	8891	8828	8709	8400	7258	5000
ETC	5441	5615	5950	6285	6781	7594	8175	8429	7453	5000
LTS-0%					8884	8825	8703	8392	7250	5000
LTS-1%				8875	8871	8812	8691	8379	7238	5000
LTS-2%			8830	8849	8837	8779	8656	8343	7209	5000
LTS-5%		8678	8693	8708	8696	8638	8508	8199	7104	5000
LTS-10%	8356	8374	8386	8392	8377	8318	8196	7907	6900	5000
LTS-15%	7989	8014	8025	8030	8012	7955	7842	7579	6679	5000
LTS-20%	7609	7622	7629	7632	7616	7563	7464	7233	6450	5000
$\sigma = 20$										
TS	8359	8353	8369	8386	8380	8331	8226	7957	6968	5000
ETC	5312	5445	5696	5983	6328	6969	7577	7936	7300	5000
LTS-0%					8375	8327	8223	7954	6965	5000
LTS-1%				8373	8368	8321	8216	7946	6958	5000
LTS-2%			8340	8351	8345	8299	8194	7924	6941	5000
LTS-5%		8222	8243	8253	8244	8193	8090	7824	6869	5000
LTS-10%	7964	7983	8005	8014	7997	7946	7847	7598	6714	5000
LTS-15%	7660	7699	7712	7713	7700	7649	7556	7328	6532	5000
LTS-20%	7340	7364	7377	7375	7364	7319	7235	7032	6334	5000
$\sigma = 25$										
TS	7923	7936	7942	7945	7939	7901	7804	7578	6704	5000
ETC	5248	5373	5591	5804	6088	6632	7120	7531	7104	5000
LTS-0%					7935	7901	7802	7575	6702	5000
LTS-1%				7938	7930	7893	7797	7571	6699	5000
LTS-2%			7924	7922	7913	7878	7782	7556	6688	5000
LTS-5%		7829	7847	7856	7835	7797	7703	7483	6637	5000
LTS-10%	7603	7637	7653	7661	7646	7602	7507	7304	6516	5000
LTS-15%	7349	7401	7412	7413	7398	7355	7268	7080	6368	5000
LTS-20%	7084	7116	7130	7130	7115	7075	6994	6828	6200	5000

Table A.1: Expected welfare for different strategies ($n = 10,000$) *(continued)*

	Batch size									
allocation	10	20	50	100	200	500	1000	2000	5000	10,000
$\sigma = 30$										
TS	7566	7551	7585	7585	7590	7548	7473	7245	6485	5000
ETC	5260	5302	5466	5664	5899	6337	6827	7159	6896	5000
LTS-0%					7577	7548	7471	7244	6484	5000
LTS-1%				7566	7574	7543	7468	7240	6481	5000
LTS-2%			7567	7554	7562	7530	7456	7230	6473	5000
LTS-5%		7520	7506	7497	7501	7464	7394	7172	6434	5000
LTS-10%	7316	7347	7357	7338	7345	7301	7231	7025	6337	5000
LTS-15%	7091	7140	7141	7130	7134	7096	7026	6839	6212	5000
LTS-20%	6871	6895	6901	6889	6894	6855	6790	6623	6067	5000

Notes: TS: Thompson sampling, ETC: Explore-then-commit, LTS-X%: Limited Thompson sampling with X% limitation. Expected welfare is calculated as the average of the sum of outcomes ($\sum_{i=1}^{n} Y$) across the simulation runs. Number of simulations = $10,000$ for $\sigma < 10$, $20,000$ for $10 \geq \sigma < 20$ and $50,000$ for $\sigma \geq 20$.

Table A.2: Bias for different strategies ($n = 10,000$)

	Batch size									
strategy	10	20	50	100	200	500	1000	2000	5000	10000
$\sigma = 1$										
TS	0.127	0.072	0.009	0.001	0.000	0.000	0.000	0.000	0.000	0.000
TS-IPW	0.177	0.169	0.047	0.001	0.000	0.000	0.000	0.000	0.000	0.000
TS-FB	0.018	-0.003	0.003	0.001	0.002	0.000	-0.001	0.000	0.000	0.000
ETC	0.008	0.000	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000
LTS-0.5%					0.002	0.000	0.001	0.000	-0.001	0.000
LTS-1%				0.001	0.000	0.000	-0.001	-0.001	0.001	0.000
LTS-2%			0.001	-0.001	0.000	0.000	-0.001	-0.001	0.000	0.000
LTS-5%		0.000	0.000	0.000	0.000	-0.001	-0.001	0.000	0.000	0.000
LTS-10%	0.000	0.000	0.000	0.000	0.000	-0.001	0.000	0.000	0.000	0.000
LTS-15%	0.000	0.000	0.000	-0.001	0.000	0.000	0.000	0.000	0.000	0.000
LTS-20%	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\sigma = 2$										
TS	0.196	0.174	0.109	0.043	0.003	-0.002	-0.001	0.000	0.000	0.000
TS-IPW	0.219	0.231	0.244	0.195	0.055	-0.002	-0.001	0.000	0.000	0.000
TS-FB	-0.002	-0.005	0.001	0.006	0.000	-0.002	-0.002	-0.001	0.000	0.000

Table A.2: Bias for different strategies ($n = 10,000$) *(continued)*

	Batch size									
strategy	10	20	50	100	200	500	1000	2000	5000	10000
ETC	0.000	-0.003	0.003	0.001	-0.002	-0.002	-0.001	0.000	0.000	0.000
LTS-0.5%					0.003	0.003	0.002	-0.001	0.001	0.000
LTS-1%				0.000	0.000	0.002	-0.003	-0.001	0.002	0.000
LTS-2%			0.001	0.001	-0.001	0.000	-0.002	-0.001	0.000	0.000
LTS-5%		-0.001	0.001	0.001	0.000	-0.001	-0.001	-0.001	0.000	0.000
LTS-10%	-0.001	0.000	0.001	0.000	0.000	-0.001	-0.001	-0.001	0.000	0.000
LTS-15%	-0.001	0.000	0.000	0.000	0.000	-0.001	-0.001	0.000	0.000	0.000
LTS-20%	0.000	0.000	0.001	0.000	0.000	-0.001	0.000	-0.001	0.000	0.000
$\sigma = 5$										
TS	0.313	0.302	0.260	0.213	0.148	0.052	0.013	0.003	0.001	0.001
TS-IPW	0.305	0.302	0.287	0.284	0.255	0.198	0.103	0.030	0.000	0.001
TS-FB	-0.012	-0.010	0.022	-0.007	-0.003	0.004	0.002	0.003	0.000	0.001
ETC	0.009	0.009	0.007	0.000	0.004	0.006	0.004	0.003	0.001	0.001
LTS-0.5%					0.004	0.007	0.005	-0.005	-0.006	0.001
LTS-1%				0.003	0.007	0.000	0.000	-0.005	-0.007	0.001
LTS-2%			0.002	0.005	0.004	-0.005	-0.004	-0.002	-0.004	0.001
LTS-5%		0.000	0.004	0.001	0.001	-0.005	-0.002	0.001	-0.001	0.001
LTS-10%	-0.001	0.003	0.001	0.001	0.000	-0.004	0.000	0.000	-0.001	0.001
LTS-15%	0.000	0.002	0.000	0.001	-0.001	-0.002	0.001	0.001	-0.001	0.001
LTS-20%	0.001	0.002	0.001	0.000	-0.001	-0.001	0.000	0.002	-0.001	0.001
$\sigma = 10$										
TS	0.419	0.394	0.383	0.342	0.279	0.182	0.099	0.035	0.003	0.002
TS-IPW	0.375	0.337	0.306	0.261	0.190	0.140	0.115	0.100	-0.007	0.002
TS-FB	-0.041	-0.043	0.008	0.030	-0.015	0.003	0.004	0.001	-0.001	0.002
ETC	-0.066	-0.023	0.002	0.007	-0.004	0.003	0.003	0.001	0.000	0.002
LTS-0.5%					-0.006	-0.001	-0.004	-0.010	0.001	0.002
LTS-1%				-0.010	-0.003	-0.001	-0.001	-0.005	-0.003	0.002
LTS-2%			-0.010	-0.002	-0.008	0.001	-0.005	-0.006	-0.007	0.002
LTS-5%		-0.004	-0.006	-0.003	0.003	-0.001	-0.005	0.003	-0.005	0.002
LTS-10%	-0.005	-0.004	-0.003	0.002	-0.001	-0.004	-0.001	0.002	-0.004	0.002
LTS-15%	-0.004	-0.005	0.000	0.001	0.001	-0.002	0.000	0.001	-0.003	0.002
LTS-20%	-0.002	-0.004	0.000	0.001	0.001	-0.001	-0.001	0.001	-0.002	0.002
$\sigma = 15$										
TS	0.516	0.509	0.462	0.424	0.376	0.267	0.184	0.092	0.017	0.002
TS-IPW	0.443	0.404	0.315	0.246	0.182	0.096	0.060	0.035	0.046	0.002
TS-FB	0.033	0.085	0.009	-0.019	0.029	-0.008	0.003	-0.002	0.000	0.002
ETC	0.045	0.067	0.015	-0.019	0.018	0.001	-0.001	0.001	0.002	0.002
LTS-0.5%					0.002	-0.002	0.005	0.008	0.007	0.002

Table A.2: Bias for different strategies ($n = 10,000$) *(continued)*

	Batch size									
strategy	10	20	50	100	200	500	1000	2000	5000	10000
LTS-1%				0.011	0.002	0.002	0.008	0.010	0.004	0.002
LTS-2%			-0.001	0.007	0.005	0.004	0.005	0.008	0.003	0.002
LTS-5%		0.003	-0.003	0.005	0.001	0.010	0.005	0.004	-0.001	0.002
LTS-10%	0.003	-0.001	-0.001	0.000	-0.001	0.008	0.001	0.002	-0.001	0.002
LTS-15%	0.001	-0.001	-0.002	-0.001	-0.002	0.005	0.002	0.003	0.001	0.002
LTS-20%	-0.001	0.002	-0.002	0.003	-0.002	0.003	0.001	0.000	0.000	0.002
$\sigma = 20$										
TS	0.541	0.545	0.507	0.478	0.421	0.322	0.226	0.131	0.025	-0.003
TS-IPW	0.437	0.401	0.305	0.220	0.156	0.082	0.039	0.029	0.023	-0.003
TS-FB	-0.040	0.009	-0.015	0.011	-0.020	-0.013	0.005	0.003	-0.002	-0.003
ETC	-0.027	-0.008	-0.012	0.003	-0.016	-0.010	0.000	-0.002	-0.004	-0.003
LTS-0.5%					0.002	-0.002	-0.003	0.005	0.000	-0.003
LTS-1%				-0.002	0.003	-0.001	-0.003	0.005	0.000	-0.003
LTS-2%			0.001	-0.002	0.001	0.005	-0.002	0.006	0.003	-0.003
LTS-5%		0.004	0.003	0.000	0.002	0.002	0.001	0.003	0.002	-0.003
LTS-10%	-0.005	0.000	0.002	0.002	-0.001	0.003	0.000	0.001	0.002	-0.003
LTS-15%	-0.005	0.001	0.002	0.000	0.000	0.001	0.001	0.001	0.002	-0.003
LTS-20%	-0.004	-0.001	0.000	-0.001	-0.001	-0.001	0.000	0.001	0.002	-0.003
$\sigma = 25$										
TS	0.603	0.588	0.558	0.511	0.463	0.358	0.260	0.159	0.036	0.006
TS-IPW	0.484	0.409	0.314	0.217	0.131	0.061	0.027	0.018	-0.006	0.006
TS-FB	0.002	0.025	0.032	0.019	0.007	0.005	-0.003	0.007	-0.004	0.006
ETC	-0.048	0.049	-0.001	0.000	-0.003	0.001	0.006	0.007	0.000	0.006
LTS-0.5%					0.003	0.003	-0.006	-0.009	-0.004	0.006
LTS-1%				0.003	-0.001	-0.001	-0.004	-0.009	-0.007	0.006
LTS-2%			0.010	0.003	-0.001	0.001	-0.003	-0.008	-0.002	0.006
LTS-5%		0.001	0.007	0.003	0.002	0.001	-0.004	-0.003	-0.004	0.006
LTS-10%	-0.002	0.003	0.000	0.004	0.002	0.000	-0.002	-0.001	-0.003	0.006
LTS-15%	0.000	0.003	0.005	0.004	0.001	0.002	-0.002	0.001	-0.002	0.006
LTS-20%	0.002	0.000	0.005	0.004	0.002	0.002	-0.002	0.003	-0.003	0.006
$\sigma = 30$										
TS	0.628	0.591	0.583	0.527	0.477	0.386	0.287	0.176	0.039	-0.002
TS-IPW	0.498	0.401	0.312	0.197	0.117	0.061	0.039	0.021	-0.012	-0.002
TS-FB	-0.069	-0.018	0.016	0.030	0.000	-0.014	0.013	-0.008	-0.001	-0.002
ETC	0.109	0.011	-0.032	-0.002	0.020	0.001	0.010	-0.002	-0.002	-0.002
LTS-0.5%					0.000	0.004	0.002	0.001	-0.011	-0.002
LTS-1%				0.003	-0.003	0.007	0.002	0.000	-0.010	-0.002
LTS-2%			0.002	-0.007	-0.005	0.002	-0.002	-0.003	-0.008	-0.002

Table A.2: Bias for different strategies ($n = 10,000$) *(continued)*

	Batch size									
strategy	10	20	50	100	200	500	1000	2000	5000	10000
LTS-5%		0.014	0.000	-0.004	-0.005	0.002	-0.005	0.000	-0.005	-0.002
LTS-10%	0.010	0.005	0.002	-0.007	0.001	-0.003	-0.001	0.001	-0.003	-0.002
LTS-15%	0.008	0.007	0.000	-0.006	-0.003	-0.002	-0.001	0.001	-0.002	-0.002
LTS-20%	0.006	0.007	0.000	-0.006	0.000	-0.001	-0.003	0.002	-0.002	-0.002

Notes: TS: Thompson sampling with $\hat{\tau}_0$, TS-IPW: Thompson sampling with $\hat{\tau}_{IPW}$, TS-FB: Thompson sampling with $\hat{\tau}_{FB}$, ETC: Explore-then-commit with $\hat{\tau}_0$, LTS-X%: Limited Thompson sampling with X% limitation and $\hat{\tau}_{IPW}$. Bias is calculated as the average difference between the estimate and the true treatment effect across the simulation runs. Number of simulations = $10,000$ for $\sigma < 10$, $20,000$ for $10 \geq \sigma < 20$ and $50,000$ for $\sigma \geq 20$.

Table A.3: MSE for different strategies ($n = 10,000$)

	Batch size									
strategy	10	20	50	100	200	500	1000	2000	5000	10000
$\sigma = 1$										
TS	0.118	0.068	0.036	0.020	0.010	0.004	0.002	0.001	0.001	0.000
TS-IPW	0.185	0.154	0.062	0.022	0.010	0.004	0.002	0.001	0.001	0.000
TS-FB	0.400	0.198	0.081	0.040	0.021	0.008	0.004	0.002	0.001	0.000
ETC	0.201	0.099	0.040	0.020	0.010	0.004	0.002	0.001	0.001	0.000
LTS-0.5%					0.019	0.020	0.018	0.016	0.010	0.000
LTS-1%				0.010	0.010	0.010	0.009	0.008	0.005	0.000
LTS-2%			0.005	0.005	0.005	0.005	0.005	0.004	0.003	0.000
LTS-5%		0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.000
LTS-10%	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000
LTS-15%	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000
LTS-20%	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000
$\sigma = 2$										
TS	0.248	0.192	0.104	0.059	0.037	0.017	0.009	0.004	0.002	0.002
TS-IPW	0.319	0.346	0.384	0.379	0.201	0.027	0.009	0.004	0.002	0.002
TS-FB	1.576	0.807	0.323	0.160	0.081	0.032	0.016	0.008	0.003	0.002
ETC	0.793	0.411	0.162	0.082	0.040	0.017	0.009	0.004	0.002	0.002
LTS-0.5%					0.079	0.081	0.073	0.065	0.040	0.002
LTS-1%				0.039	0.040	0.039	0.036	0.032	0.021	0.002
LTS-2%			0.020	0.020	0.020	0.020	0.018	0.016	0.011	0.002
LTS-5%		0.008	0.009	0.008	0.009	0.008	0.008	0.007	0.005	0.002

Table A.3: MSE for different strategies ($n = 10,000$) *(continued)*

	Batch size									
strategy	10	20	50	100	200	500	1000	2000	5000	10000
LTS-10%	0.004	0.005	0.005	0.005	0.004	0.004	0.004	0.004	0.003	0.002
LTS-15%	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.002	0.002
LTS-20%	0.003	0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002
$\sigma = 5$										
TS	0.637	0.548	0.388	0.259	0.159	0.076	0.049	0.028	0.013	0.010
TS-IPW	0.701	0.731	0.762	0.902	1.208	1.821	1.697	0.546	0.016	0.010
TS-FB	9.965	4.994	1.994	0.989	0.515	0.197	0.101	0.051	0.020	0.010
ETC	5.109	2.567	1.019	0.507	0.261	0.104	0.055	0.028	0.013	0.010
LTS-0.5%					0.407	0.454	0.427	0.400	0.254	0.010
LTS-1%				0.224	0.221	0.227	0.223	0.204	0.130	0.010
LTS-2%			0.115	0.119	0.119	0.118	0.112	0.106	0.068	0.010
LTS-5%		0.050	0.052	0.051	0.051	0.050	0.048	0.045	0.031	0.010
LTS-10%	0.027	0.028	0.027	0.027	0.027	0.027	0.026	0.024	0.019	0.010
LTS-15%	0.021	0.020	0.020	0.019	0.019	0.019	0.019	0.018	0.015	0.010
LTS-20%	0.015	0.016	0.015	0.015	0.016	0.015	0.015	0.015	0.013	0.010
$\sigma = 10$										
TS	1.355	1.072	0.887	0.656	0.440	0.238	0.141	0.088	0.052	0.040
TS-IPW	1.461	1.294	1.384	1.515	1.722	2.508	3.616	4.936	3.608	0.040
TS-FB	39.354	19.621	8.090	4.034	1.988	0.802	0.401	0.202	0.080	0.040
ETC	19.620	9.940	4.059	2.010	1.001	0.407	0.210	0.111	0.053	0.040
LTS-0.5%					0.989	1.047	1.129	1.199	0.929	0.040
LTS-1%				0.628	0.624	0.646	0.661	0.673	0.492	0.040
LTS-2%			0.380	0.380	0.373	0.382	0.384	0.364	0.268	0.040
LTS-5%		0.186	0.191	0.183	0.182	0.183	0.175	0.166	0.123	0.040
LTS-10%	0.105	0.106	0.105	0.105	0.103	0.100	0.099	0.095	0.075	0.040
LTS-15%	0.083	0.075	0.076	0.077	0.075	0.073	0.073	0.070	0.059	0.040
LTS-20%	0.062	0.061	0.061	0.062	0.060	0.059	0.060	0.058	0.051	0.040
$\sigma = 15$										
TS	2.571	2.109	1.566	1.173	0.842	0.477	0.292	0.178	0.109	0.091
TS-IPW	2.666	2.407	2.188	2.220	2.376	2.788	3.776	5.158	5.206	0.091
TS-FB	90.993	44.865	17.786	8.885	4.576	1.777	0.900	0.452	0.181	0.091
ETC	45.523	22.524	8.949	4.504	2.271	0.916	0.480	0.253	0.121	0.091
LTS-0.5%					1.484	1.486	1.520	1.625	1.294	0.091
LTS-1%				1.053	1.028	1.011	1.021	1.027	0.796	0.091
LTS-2%			0.684	0.676	0.674	0.661	0.656	0.645	0.483	0.091
LTS-5%		0.373	0.373	0.366	0.366	0.359	0.345	0.334	0.253	0.091
LTS-10%	0.226	0.222	0.224	0.221	0.221	0.219	0.209	0.203	0.163	0.091
LTS-15%	0.180	0.165	0.166	0.164	0.163	0.164	0.158	0.154	0.131	0.091

Table A.3: MSE for different strategies ($n = 10,000$) *(continued)*

	Batch size									
strategy	10	20	50	100	200	500	1000	2000	5000	10000
LTS-20%	0.137	0.136	0.137	0.135	0.134	0.136	0.130	0.129	0.115	0.091
$\sigma = 20$										
TS	4.045	3.241	2.359	1.846	1.301	0.791	0.490	0.312	0.191	0.160
TS-IPW	4.121	3.481	3.039	2.952	2.943	3.234	3.647	4.817	4.828	0.160
TS-FB	160.604	79.991	31.699	15.892	7.963	3.204	1.585	0.812	0.321	0.160
ETC	79.498	39.556	15.826	7.942	4.007	1.641	0.835	0.446	0.215	0.160
LTS-0.5%					1.943	1.844	1.821	1.873	1.480	0.160
LTS-1%				1.463	1.384	1.337	1.317	1.310	0.991	0.160
LTS-2%			1.016	0.992	0.972	0.946	0.912	0.885	0.666	0.160
LTS-5%		0.596	0.587	0.573	0.573	0.556	0.539	0.513	0.394	0.160
LTS-10%	0.375	0.373	0.367	0.370	0.371	0.358	0.351	0.333	0.275	0.160
LTS-15%	0.307	0.284	0.286	0.282	0.284	0.277	0.271	0.262	0.227	0.160
LTS-20%	0.238	0.237	0.236	0.238	0.236	0.230	0.228	0.221	0.201	0.160
$\sigma = 25$										
TS	5.833	4.810	3.600	2.714	1.956	1.181	0.757	0.486	0.299	0.249
TS-IPW	5.831	5.002	4.337	4.033	3.712	3.741	4.020	4.813	4.535	0.249
TS-FB	251.082	124.151	49.360	24.801	12.551	5.030	2.480	1.239	0.497	0.249
ETC	124.807	62.439	24.887	12.560	6.382	2.570	1.310	0.689	0.335	0.249
LTS-0.5%					2.490	2.297	2.154	2.144	1.604	0.249
LTS-1%				1.927	1.804	1.713	1.636	1.552	1.172	0.249
LTS-2%			1.397	1.340	1.300	1.251	1.201	1.109	0.843	0.249
LTS-5%		0.840	0.824	0.806	0.811	0.785	0.754	0.701	0.543	0.249
LTS-10%	0.563	0.548	0.544	0.543	0.544	0.527	0.512	0.484	0.401	0.249
LTS-15%	0.466	0.425	0.429	0.425	0.429	0.420	0.409	0.393	0.341	0.249
LTS-20%	0.362	0.359	0.361	0.360	0.361	0.358	0.351	0.339	0.305	0.249
$\sigma = 30$										
TS	9.055	7.077	5.094	3.642	2.670	1.693	1.085	0.701	0.432	0.363
TS-IPW	9.056	7.285	6.168	4.952	4.528	4.539	4.597	5.099	4.519	0.363
TS-FB	364.392	178.634	72.500	35.187	18.079	7.213	3.605	1.810	0.723	0.363
ETC	180.242	89.158	35.970	18.149	9.123	3.720	1.891	1.001	0.480	0.363
LTS-0.5%					3.065	2.808	2.590	2.392	1.862	0.363
LTS-1%				2.493	2.289	2.151	2.018	1.845	1.395	0.363
LTS-2%			1.835	1.748	1.693	1.622	1.528	1.393	1.051	0.363
LTS-5%		1.145	1.114	1.081	1.059	1.060	1.007	0.920	0.727	0.363
LTS-10%	0.772	0.765	0.755	0.743	0.751	0.731	0.705	0.660	0.561	0.363
LTS-15%	0.646	0.604	0.604	0.602	0.606	0.587	0.573	0.548	0.487	0.363
LTS-20%	0.516	0.515	0.513	0.513	0.518	0.504	0.494	0.479	0.441	0.363

Table A.3: MSE for different strategies ($n = 10,000$) *(continued)*

	Batch size									
strategy	10	20	50	100	200	500	1000	2000	5000	10000

Notes: TS: Thompson sampling with $\hat{\tau}_0$, TS-IPW: Thompson sampling with $\hat{\tau}_{IPW}$, TS-FB: Thompson sampling with $\hat{\tau}_{FB}$, ETC: Explore-then-commit with $\hat{\tau}_0$, LTS-X%: Limited Thompson sampling with X% limitation and $\hat{\tau}_{IPW}$. MSE is calculated as the average of the squared errors in the treatment effect estimate across the simulation runs. Number of simulations = 10,000 for $\sigma < 10$, 20,000 for $10 \geq \sigma < 20$ and 50,000 for $\sigma \geq 20$.

Table A.4: Expected welfare for different strategies ($\sigma = 10$)

	Batch size									
allocation	10	20	50	100	200	500	1000	2000	5000	10,000
$n = 2000$										
TS	1631	1640	1638	1627	1605	1526	1367	1000		
ETC	1120	1196	1268	1363	1469	1550	1442	1000		
LTS-0.5%					1604	1525	1367	1000		
LTS-1%				1625	1603	1524	1366	1000		
LTS-2%			1630	1621	1599	1521	1363	1000		
LTS-5%		1612	1609	1602	1581	1504	1351	1000		
LTS-10%	1558	1566	1565	1557	1536	1466	1323	1000		
LTS-15%	1502	1510	1510	1503	1483	1419	1290	1000		
LTS-20%	1444	1449	1447	1442	1424	1367	1253	1000		
$n = 10000$										
TS	9372	9382	9389	9392	9383	9321	9178	8820	7469	5000
ETC	5626	5840	6375	6946	7533	8496	9004	8896	7498	5000
LTS-0.5%					9371	9308	9163	8800	7451	5000
LTS-1%				9358	9351	9288	9140	8774	7430	5000
LTS-2%			9302	9309	9301	9234	9082	8713	7384	5000
LTS-5%		9096	9106	9111	9097	9026	8871	8508	7241	5000
LTS-10%	8694	8707	8713	8711	8695	8623	8476	8139	6995	5000
LTS-15%	8260	8278	8280	8276	8258	8193	8060	7758	6747	5000
LTS-20%	7819	7829	7829	7826	7809	7750	7634	7370	6498	5000
$n = 20000$										
TS	19,268	19,272	19,288	19,298	19,292	19,243	19,119	18,791	17,465	14,998
ETC	11,181	11,714	12,871	13,743	15,346	17,252	18,468	18,757	17,496	15,000
LTS-0.5%					19,258	19,210	19,079	18,737	17,399	14,949
LTS-1%				19,202	19,204	19,153	19,017	18,666	17,329	14,899

Table A.4: Expected welfare for different strategies ($\sigma = 10$) *(continued)*

	Batch size									
allocation	10	20	50	100	200	500	1000	2000	5000	10,000
LTS-2%			19,063	19,075	19,072	19,014	18,869	18,510	17,183	14,800
LTS-5%		18,583	18,597	18,598	18,588	18,520	18,368	18,009	16,740	14,500
LTS-10%	17,692	17,702	17,712	17,706	17,692	17,622	17,477	17,141	15,995	14,000
LTS-15%	16,765	16,779	16,781	16,773	16,758	16,692	16,561	16,259	15,247	13,500
LTS-20%	15,819	15,827	15,832	15,822	15,810	15,750	15,635	15,371	14,498	13,000
$n = 40000$										
TS	39,152	39,167	39,182	39,210	39,197	39,166	39,058	38,757	37,460	34,998
ETC	22,203	23,374	25,845	28,008	30,444	34,583	37,421	38,571	37,486	35,000
LTS-0.5%					39,104	39,072	38,954	38,625	37,297	34,849
LTS-1%				38,977	38,972	38,932	38,803	38,460	37,127	34,700
LTS-2%			38,655	38,672	38,660	38,607	38,464	38,108	36,782	34,400
LTS-5%		37,582	37,598	37,608	37,586	37,518	37,366	37,008	35,739	33,500
LTS-10%	35,690	35,704	35,715	35,712	35,691	35,619	35,474	35,141	33,994	32,000
LTS-15%	33,763	33,779	33,782	33,776	33,757	33,690	33,559	33,260	32,246	30,500
LTS-20%	31,821	31,829	31,832	31,825	31,807	31,747	31,633	31,372	30,498	29,000

Notes: TS: Thompson sampling, ETC: Explore-then-commit, LTS-X%: Limited Thompson sampling with X% limitation. Expected welfare is calculated as the average of the sum of outcomes ($\sum_{i=1}^{n} Y$) across the simulation runs. Number of simulations = $20,000$ for $n = 10,000$, and $10,000$ otherwise.

Table A.5: Bias for different strategies ($\sigma = 10$)

	Batch size									
strategy	10	20	50	100	200	500	1000	2000	5000	10000
$n = 2000$										
TS	0.527	0.514	0.427	0.347	0.253	0.111	0.033	0.002		
TS-IPW	0.429	0.370	0.242	0.164	0.092	0.025	0.019	0.002		
TS-FB	-0.001	0.076	0.001	0.022	0.002	0.001	0.002	0.002		
ETC	0.010	0.065	0.001	0.025	0.013	0.003	0.003	0.002		
LTS-0.5%					-0.003	-0.008	0.002	0.002		
LTS-1%				-0.008	-0.005	-0.006	-0.002	0.002		
LTS-2%			-0.001	-0.003	0.006	-0.008	0.007	0.002		
LTS-5%		0.008	-0.003	-0.001	0.000	-0.011	-0.001	0.002		
LTS-10%	-0.001	0.007	-0.002	-0.004	0.000	0.000	-0.002	0.002		
LTS-15%	0.000	0.007	-0.005	-0.002	0.001	0.002	-0.003	0.002		
LTS-20%	0.005	0.007	-0.006	0.002	-0.001	0.003	-0.001	0.002		

Table A.5: Bias for different strategies ($\sigma = 10$) *(continued)*

	Batch size									
strategy	10	20	50	100	200	500	1000	2000	5000	10000
$n = 10000$										
TS	0.419	0.394	0.383	0.342	0.279	0.182	0.099	0.035	0.003	0.002
TS-IPW	0.375	0.337	0.306	0.261	0.190	0.140	0.115	0.100	-0.007	0.002
TS-FB	-0.041	-0.043	0.008	0.030	-0.015	0.003	0.004	0.001	-0.001	0.002
ETC	-0.066	-0.023	0.002	0.007	-0.004	0.003	0.003	0.001	0.000	0.002
LTS-0.5%					-0.006	-0.001	-0.004	-0.010	0.001	0.002
LTS-1%				-0.010	-0.003	-0.001	-0.001	-0.005	-0.003	0.002
LTS-2%			-0.010	-0.002	-0.008	0.001	-0.005	-0.006	-0.007	0.002
LTS-5%		-0.004	-0.006	-0.003	0.003	-0.001	-0.005	0.003	-0.005	0.002
LTS-10%	-0.005	-0.004	-0.003	0.002	-0.001	-0.004	-0.001	0.002	-0.004	0.002
LTS-15%	-0.004	-0.005	0.000	0.001	0.001	-0.002	0.000	0.001	-0.003	0.002
LTS-20%	-0.002	-0.004	0.000	0.001	0.001	-0.001	-0.001	0.001	-0.002	0.002
$n = 20000$										
TS	0.378	0.368	0.343	0.312	0.267	0.184	0.108	0.041	0.005	0.001
TS-IPW	0.352	0.341	0.310	0.271	0.207	0.192	0.163	0.135	0.078	0.009
TS-FB	-0.047	0.027	0.015	-0.032	0.022	-0.006	0.001	0.003	0.002	0.001
ETC	0.000	0.023	-0.004	-0.008	0.014	-0.004	0.000	0.001	0.002	0.001
LTS-0.5%					-0.010	0.000	-0.010	-0.009	0.009	0.009
LTS-1%				-0.003	-0.003	-0.002	-0.004	-0.005	0.012	0.007
LTS-2%			-0.005	-0.003	0.001	0.005	0.000	-0.001	0.010	0.001
LTS-5%		0.002	-0.003	-0.002	0.000	0.005	0.000	0.000	0.000	0.000
LTS-10%	0.001	0.002	-0.003	-0.001	0.002	0.001	0.001	-0.001	-0.001	0.001
LTS-15%	0.001	0.001	-0.004	-0.002	0.001	0.001	-0.001	0.000	-0.002	0.001
LTS-20%	0.000	-0.001	-0.001	0.001	0.001	0.002	-0.002	-0.002	0.000	0.000
$n = 40000$										
TS	0.305	0.306	0.300	0.289	0.251	0.189	0.117	0.047	0.002	0.000
TS-IPW	0.291	0.294	0.281	0.285	0.265	0.228	0.205	0.201	0.069	0.011
TS-FB	-0.078	-0.027	0.047	0.036	0.005	0.000	0.003	0.003	-0.005	0.001
ETC	-0.001	-0.005	0.061	0.025	0.010	0.001	-0.001	0.000	-0.002	0.000
LTS-0.5%					0.003	0.003	0.006	0.000	0.002	-0.001
LTS-1%				0.000	0.002	0.001	0.007	-0.003	0.005	0.002
LTS-2%			0.000	0.001	0.001	0.002	0.001	-0.001	-0.002	-0.001
LTS-5%		0.001	0.000	0.004	0.003	0.000	0.001	0.000	-0.003	-0.001
LTS-10%	-0.001	0.000	0.003	0.002	0.003	0.001	0.001	-0.002	-0.002	0.000
LTS-15%	-0.001	0.000	0.002	0.001	0.001	0.001	0.000	-0.002	-0.002	0.001
LTS-20%	-0.001	0.000	0.003	0.002	0.001	0.000	-0.001	-0.002	-0.002	0.001

Table A.5: Bias for different strategies ($\sigma = 10$) *(continued)*

	Batch size									
strategy	10	20	50	100	200	500	1000	2000	5000	10000

Notes: TS: Thompson sampling with $\hat{\tau}_0$, TS-IPW: Thompson sampling with $\hat{\tau}_{IPW}$, TS-FB: Thompson sampling with $\hat{\tau}_{FB}$, ETC: Explore-then-commit with $\hat{\tau}_0$, LTS-X%: Limited Thompson sampling with X% limitation and $\hat{\tau}_{IPW}$. Bias is calculated as the average difference between the estimate and the true treatment effect across the simulation runs. Number of simulations = 20,000 for $n = 10,000$, and 10,000 otherwise.

Table A.6: MSE for different strategies ($\sigma = 10$)

	Batch size									
strategy	10	20	50	100	200	500	1000	2000	5000	10000
$n = 2000$										
TS	3.045	2.203	1.399	0.992	0.616	0.343	0.243	0.200		
TS-IPW	3.210	2.629	2.303	2.401	2.464	2.604	2.227	0.200		
TS-FB	40.238	19.600	7.975	4.059	1.995	0.789	0.409	0.200		
ETC	20.402	10.048	4.107	2.064	1.047	0.462	0.271	0.200		
LTS-0.5%					2.147	1.994	1.560	0.200		
LTS-1%				1.620	1.509	1.375	1.087	0.200		
LTS-2%			1.176	1.120	1.089	0.943	0.769	0.200		
LTS-5%		0.703	0.688	0.681	0.651	0.566	0.471	0.200		
LTS-10%	0.440	0.451	0.443	0.444	0.432	0.386	0.335	0.200		
LTS-15%	0.365	0.348	0.345	0.352	0.338	0.308	0.279	0.200		
LTS-20%	0.289	0.297	0.290	0.298	0.284	0.267	0.253	0.200		
$n = 10000$										
TS	1.355	1.072	0.887	0.656	0.440	0.238	0.141	0.088	0.052	0.040
TS-IPW	1.461	1.294	1.384	1.515	1.722	2.508	3.616	4.936	3.608	0.040
TS-FB	39.354	19.621	8.090	4.034	1.988	0.802	0.401	0.202	0.080	0.040
ETC	19.620	9.940	4.059	2.010	1.001	0.407	0.210	0.111	0.053	0.040
LTS-0.5%					0.989	1.047	1.129	1.199	0.929	0.040
LTS-1%				0.628	0.624	0.646	0.661	0.673	0.492	0.040
LTS-2%			0.380	0.380	0.373	0.382	0.384	0.364	0.268	0.040
LTS-5%		0.186	0.191	0.183	0.182	0.183	0.175	0.166	0.123	0.040
LTS-10%	0.105	0.106	0.105	0.105	0.103	0.100	0.099	0.095	0.075	0.040
LTS-15%	0.083	0.075	0.076	0.077	0.075	0.073	0.073	0.070	0.059	0.040
LTS-20%	0.062	0.061	0.061	0.062	0.060	0.059	0.060	0.058	0.051	0.040
$n = 20000$										

Table A.6: MSE for different strategies ($\sigma = 10$) *(continued)*

	Batch size									
strategy	10	20	50	100	200	500	1000	2000	5000	10000
TS	1.102	1.007	0.721	0.565	0.409	0.230	0.132	0.079	0.043	0.027
TS-IPW	1.187	1.178	1.080	1.181	1.374	2.142	3.360	5.423	4.824	1.119
TS-FB	40.125	20.155	7.858	3.982	1.955	0.795	0.394	0.200	0.078	0.039
ETC	20.110	9.920	3.951	2.013	0.986	0.407	0.205	0.104	0.045	0.027
LTS-0.5%					0.630	0.668	0.685	0.759	0.735	0.526
LTS-1%				0.383	0.368	0.388	0.394	0.425	0.378	0.266
LTS-2%			0.216	0.217	0.216	0.216	0.222	0.220	0.194	0.140
LTS-5%		0.098	0.101	0.100	0.098	0.098	0.095	0.095	0.085	0.065
LTS-10%	0.053	0.053	0.054	0.055	0.054	0.053	0.052	0.052	0.047	0.038
LTS-15%	0.042	0.038	0.039	0.038	0.037	0.037	0.037	0.037	0.035	0.030
LTS-20%	0.031	0.030	0.031	0.031	0.030	0.030	0.030	0.030	0.029	0.026
$n = 40000$										
TS	0.709	0.704	0.570	0.496	0.372	0.222	0.129	0.076	0.040	0.023
TS-IPW	0.762	0.821	0.806	0.996	1.164	1.754	2.879	4.926	5.361	1.328
TS-FB	40.465	19.838	7.935	3.991	2.001	0.806	0.397	0.202	0.080	0.040
ETC	19.978	10.142	4.042	2.023	0.983	0.405	0.208	0.104	0.043	0.023
LTS-0.5%					0.376	0.402	0.400	0.413	0.432	0.383
LTS-1%				0.214	0.217	0.219	0.215	0.218	0.221	0.189
LTS-2%			0.117	0.120	0.118	0.115	0.118	0.115	0.112	0.099
LTS-5%		0.050	0.053	0.051	0.051	0.051	0.050	0.049	0.047	0.042
LTS-10%	0.027	0.027	0.028	0.028	0.028	0.028	0.027	0.027	0.025	0.023
LTS-15%	0.022	0.019	0.020	0.020	0.020	0.020	0.020	0.019	0.018	0.017
LTS-20%	0.016	0.015	0.016	0.016	0.016	0.016	0.016	0.015	0.015	0.014

Notes: TS: Thompson sampling with $\hat{\tau}_0$, TS-IPW: Thompson sampling with $\hat{\tau}_{IPW}$, TS-FB: Thompson sampling with $\hat{\tau}_{FB}$, ETC: Explore-then-commit with $\hat{\tau}_0$, LTS-X%: Limited Thompson sampling with X% limitation and $\hat{\tau}_{IPW}$. MSE is calculated as the average of the squared errors in the treatment effect estimate across the simulation runs. Number of simulations = $20,000$ for $n = 10,000$, and $10,000$ otherwise.