Why Adaptively Collected Data Have Negative Bias and How to Correct for It

| Xinkun Nie | Xiaoying Tian | Jonathan Taylor | James Zou |
| Stanford University | Stanford University | Stanford University | Stanford University |

Abstract

From scientific experiments to online A/B testing, the previously observed data often affects how future experiments are performed, which in turn affects which data will be collected. Such adaptivity introduces complex correlations between the data and the collection procedure. In this paper, we prove that when the data collection procedure satisfies natural conditions, then sample means of the data have systematic *negative* biases. As an example, consider an adaptive clinical trial where additional data points are more likely to be tested for treatments that show initial promise. Our surprising result implies that the average observed treatment effects would underestimate the true effects of each treatment. We quantitatively analyze the magnitude and behavior of this negative bias in a variety of settings. We also propose a novel de-biasing algorithm based on selective inference techniques. In experiments, our method can effectively reduce bias and estimation error.

1 INTRODUCTION

Much of modern data science is driven by data that is collected adaptively. A scientist often starts off testing multiple experimental conditions, and based on the initial results may decide to collect more data points from some conditions and less data from other settings. A sequential clinical trial initially groups the participants into different treatment regimes, and depending on the continuous feedback, may reallocate participants into the more promising treatments. In e-commerce, companies often use online A/B tests to

collect user data from multiple variants of a project, and could adaptively collect more data from a subset of the variants (multi-arm bandit algorithms are often used here to decide which variant to collect data from as a function of the data log history).

The key characteristic of adaptively collected data is that the analyst sequentially collects data from multiple alternatives (e.g. different treatments, products, etc.). The choice of which alternative to gather data from at a particular time depends on the previously observed data from all the options. The collected data could be used in many different ways. In some settings, the analyst simply wants to use it to identify the single best alternative, and may not care about the data beyond this goal (this setting motivates many bandit problems). In many other settings, the data itself could be used to estimate various statistical parameters. In the sequential clinical trial example, many scientists would like to use the data to estimate the effects of each of the treatments. Even if the company sponsoring the trials may care most about identifying the best treatment, other scientist using the data may care about the effect size estimates of other treatments in the data for their own applications.

Our contributions. We study the problem of estimation using adaptively collected data. We prove that when the adaptive data collection procedure satisfies two natural conditions (precisely defined in Sec. 2), then the sample mean of the collected data is negatively biased as an estimator for the true mean. This means that the effect size empirically observed is systematically less than the true effect size for every alternative. We provide intuition for this counter-intuitive result, and compare and analyze the magnitude of this negative bias across different conditions and collection procedures. We then propose a novel randomized algorithm called the conditional Maximum Likelihood Estimator (cMLE) based on selective inference to reduce this ubiquitous bias, and compare it with a simple approach using an independent set of held-out data. We validate the performance of our bias-reduction algorithm in extensive experiments. All the proofs and

additional experiments are in the Appendix.

Related works. Multi-arm bandits and its variations are extensively studied in machine learning. The goal of our work is different from that of the standard bandit setting. In bandits, the data sampled from an arm (i.e. one of the alternatives) is considered a reward and the objective is to *design* adaptive algorithms to pick arms so as to maximize total reward (or minimize regret). Our goal is not to design such algorithms and we are agnostic to the reward. We take the perspective of an analyst who is given such an adaptively collected dataset and wants to estimate statistical parameters.

Xu et al. [2013] empirically observed estimation bias due to selection in specific multi-arm bandit algorithms. They were primarily interested in estimating the values of the top two arms, and used data splitting with a held-out set in their experiments to reduce bias. We are the first one to rigorously prove that such underestimation is a general phenomenon. Our cMLE approach builds upon recent advances in selective inference [Taylor and Tibshirani, 2015, Tian and Taylor, 2015], which derives valid confidence intervals accounting for selection effects of the algorithm. Selective inference has been applied to regression problems (e.g. LASSO, Stepwise regression), and has not been considered for the adaptive data collection setting before. We build upon results from recent developments in this area [Tian and Taylor, 2015, Tian et al., 2016, Harris et al., 2016].

The problem of selection bias has been extensively studied, especially in the context of Winner's Curse in genetic association studies [Ionita-Laza et al., 2009]. There the bias arises from selective reporting rather than adaptivity in data collection. There is a related line of recent work [Dwork et al., 2015, Russo and Zou, 2016] in adaptive data analysis that is complementary to ours. In their work the data is fixed (and is typically i.i.d.) and the adaptivity is in the analyst. In contrast, in our work the data collection itself is adaptive.

2 ADAPTIVE DATA COLLECTION HAS NEGATIVE BIAS

Model of adaptive data collection. We have K unknown distributions that we would like to collect data from. There are T rounds of data collection and at round $t \in [T]$ the distribution $s_t \in [K]$ is selected, and we draw $X^{(s_t)}$, an independent sample, from s_t. The data collection procedure can be modeled by a selection function $s_t = f(\Lambda_t)$, where Λ_t is the history of the observed samples up to time t. More precisely, let $X_i^{(k)}$ denote the i-th sample from distribution k and $N_t^{(k)}$ denote the number of times that distribution k is sampled by round t, which could be a random variable, then

$\Lambda_t = \{\{X_1^{(1)}, ..., X_{N_t^{(1)}}^{(1)}\}, ..., \{X_1^{(K)}, ..., X_{N_t^{(K)}}^{(K)}\}\}$. The history of distribution k up to round t is denoted by $\Lambda_t^{(k)} = \{X_1^{(k)}, ..., X_{N_t^{(k)}}^{(k)}\}$. We use $\Lambda_t^{(-k)}$ to denote the history up to round t of all the distributions except for the k-th one; $\Lambda_t^{(-k)} = \{\{X_1^{(i)}, ..., X_{N_t^{(K)}}^{(i)}\}\}_{i \in [K] \setminus k}$. We allow f to be a randomized function, and will sometimes write $f(\Lambda_t, \omega)$, where $\omega \in \Omega$ is a random seed, to highlight this randomness. Let $\overline{X_t^{(k)}} \equiv \frac{\sum_{i=1}^{N_t^{(k)}} X_i^{(k)}}{N_t^{(k)}}$ denote the sample average of distribution k at round t.

Example. The simplest example of adaptive data collection is the Greedy algorithm. In Greedy, at round t, the selection function chooses to sample the distribution with the highest empirical mean. Then $f(\Lambda_t) = \arg\max_{k \in [K]} \overline{X_t^{(k)}}$. Often in practice, a randomized version of Greedy, called ϵ-Greedy, is also used. In ϵ-Greedy with probability ϵ we uniformly randomly select a distribution and with probability $1 - \epsilon$, we perform Greedy. This corresponds to the selection

$$f(\Lambda_t, \omega) = \begin{cases} \arg\max_{k \in [K]} \overline{X_t^{(k)}}, & \text{if } \omega > \epsilon \\ k, k \in [K], & \text{if } \frac{\epsilon}{K} \cdot (k-1) < \omega < \frac{\epsilon}{K} \cdot k \end{cases}$$

where $\omega \sim \text{Unif}[0, 1]$. All common multi-arm bandit algorithms can be modeled as a selection function f.

Many adaptive data collection procedures correspond to a selection function f that satisfies two natural properties: *Exploit* and *Independence of Irrelevant Option (IIO)*. *Exploit* means that all else being equal, if distribution k is selected in a scenario where it has lower sample average, then k would also be selected in a scenario where it has higher sample average. *IIO* means that if distribution k is not selected then the precise values observed from k does not affect which of the other distributions is selected. We precisely define these two properties next.

Definition 1 (Exploit)**.** *Given any $t \in [T]$, $k \in [K]$, realization $\Lambda_t^{(-k)}$ and random seed ω. Suppose $\Lambda_t^{(k)}$ and $\Lambda_t^{'(k)}$ are two sample histories of distribution k of length n with sample means $\overline{X_t^{(k)}} \leq \overline{X_t^{'(k)}}$. Then $f(\Lambda_t^{(k)} \cup \Lambda_t^{(-k)}, \omega) = k$ implies $f(\Lambda_t^{'(k)} \cup \Lambda_t^{(-k)}, \omega) = k$. In words,* Exploit *states that given the same context specified by $\Lambda_t^{(-k)}$ and ω, if k is selected when it has smaller sample mean then it should also be selected when it has a larger mean.*

Exploit captures the intuition that when we are looking for options that work well, we are more likely to try out the options that show more promise early on. Note that in *Exploit*, we only compare two sample histories $\Lambda_t^{(k)}$ and $\Lambda_t^{'(k)}$ with the same number of observed samples.

This allows f to also account for the number of samples observed so far (e.g. selecting a distribution k with low sample average if it does not have many samples). Therefore confidence interval based bandit algorithms can also be shown to satisfy *Exploit*.

Definition 2 (Independent of Irrelevant Options (IIO)). *Given any $t \in [T]$ and $k \in [K]$. Let $\Lambda_t = \Lambda_t^{(k)} \cup \Lambda_t^{(-k)}$ and $\Lambda_t' = \Lambda_t^{'(k)} \cup \Lambda_t^{(-k)}$, i.e. Λ_t and Λ_t' have the same histories for distributions $i \neq k$ and could have arbitrary histories for distribution k. Then $\forall i \neq k$,*

$$\Pr\left[f\left(\Lambda_t\right) = i | f\left(\Lambda_t\right) \neq k\right] = \Pr\left[f\left(\Lambda_t'\right) = i | f\left(\Lambda_t'\right) \neq k\right].$$

In words, so long as k is not chosen, which other distribution is selected depends only on the history $\Lambda_t^{(-k)}$ of those distributions.

Estimation bias. In this paper, we are interested in the fundamental problem of estimating the true mean, $\mu_k = \mathbb{E}[X^{(k)}]$, of each of the distributions given a sample history dataset, Λ_T, which is collected through an adaptive procedure. This models the adaptive clinical trials example, where the scientist is interested in estimating $\{\mu_k\}_{k \in [K]}$, the true effects of the treatments. Of course, if the scientist can collect her own data, she could just collect a non-adaptive set of samples and obtain unbiased estimates of $\{\mu_k\}_{k \in [K]}$. However, in many settings like the clinical trials, the scientist does not collect the data; rather it is adaptively collected by a pharmaceutical company with a different objective of finding an optimal treatment or demonstrating efficacy. The simplest and most common approach is to use the sample average $\overline{X_T^{(k)}}$ to estimate the true mean μ_k. Our main result shows that in expectation, the sample average underestimates the true mean if f satisfies *Exploit* and *IIO*: $\mathbb{E}\left[\overline{X_T^{(k)}}\right] \leq \mu_k, \forall k \in [K]$.

Theorem 1. *Suppose $X^{(k)}, k \in [K]$ is a sample drawn from a distribution with finite mean $\mu_k = \mathbb{E}[X^{(k)}]$, and the selection function f satisfies* Exploit *and* IIO. *Then $\forall k$ and $\forall T$, $\mathbb{E}\left[\overline{X_T^{(k)}}\right] \leq \mu_k$. Moreover, the equality holds only if the number of times distribution k is selected, $N_T^{(k)}$, does not depend on the observed history $\Lambda_T^{(k)}$ of k.*

Intuition behind the proof. Here we present the high-level insights for the proof. The detailed proof is in Appendix B. For simplicity, we condition on a fixed realization of distributions $2, \ldots, K$. If the bias of distribution 1 is negative for every realization of distributions $2, \ldots, K$, then taking the expectation shows that the total bias is negative.

Consider a particular sample path history $\Lambda_t^{(1)}$ at some round $t < T$, with corresponding empirical average

$\overline{X_t^{(1)}}$. There are two types of scenarios. First, $\Lambda_t^{(1)}$ could be *lucky* and $\overline{X_t^{(1)}} > \mu_1$. Then the *Exploit* property states that with this lucky sample path history, distribution 1 is likely to be sampled more often in the future in rounds $[t+1, \ldots, T]$. Since these future samples have expected values μ_1, the expected average of final value of $\overline{X_T^{(1)}}$ is likely to decrease closer to μ_1. This is similar to the reversion to mean phenomenon. In scenario two, $\Lambda_t^{(1)}$ is *unlucky* and $\overline{X_t^{(1)}} < \mu_1$. The exploitative nature of f makes it less likely to select distribution 1 and this sample path history is likely to be stuck with the negative bias. Therefore we see that the exploitative property of the adaptive collection procedure creates a fundamental asymmetry in the sample path histories such that the positive bias (lucky) paths revert back to mean but the negative bias (unlucky) paths are stuck at negative. The overall bias becomes negative. The *IIO* property allows us to safely condition on the realizations of distributions $2, \ldots, K$ and isolate the effects of distribution 1. □

Many standard multi-arm bandit algorithms can be modeled by a selection function f that satisfies *Exploit* and *IIO*. While Greedy only has sample mean as its input, upper confidence bound (UCB) type algorithms also account for the number of observations and give preference for the less explored distributions. lil' UCB is the state-of-the-art UCB algorithm [Jamieson et al., 2014] and its details are presented in Appendix A.

Proposition 1. *lil' UCB, Greedy, ϵ-Greedy are all equivalent to selection functions $f(\Lambda_t)$ that satisfy* Exploit *and* IIO.

In Appendix H, we extend Proposition 1 to Thompson Sampling [Thompson, 1933, Agrawal and Goyal, 2012]. When $K = 2$, we do not need the *IIO* condition in order for the bias to be non-positive.

Proposition 2. *Suppose $X^{(1)}, X^{(2)}$ are samples drawn from distributions with finite means μ_1, μ_2 and the selection function f satisfies* Exploit. *Then for $k \in \{1, 2\}$ and all T, $\mathbb{E}\left[\overline{X_T^{(k)}}\right] \leq \mu_k$. Moreover the equality holds only if the number of times distribution k is selected, $N_T^{(k)}$, does not depend on observed values $\Lambda_T^{(k)}$ of k.*

3 QUANTITATIVE CHARACTERIZATION OF BIAS

Analytic example with explicit bias. Consider the setting where $K = 2$, $X^{(1)} \sim Bernoulli(\mu_1)$ and $X^{(2)} \sim Bernoulli(\mu_2)$. A greedy data collection procedure is to draw one sample from each distribution in the first two rounds, and at $T = 3$ sample from the distribution with the larger empirical sample mean. In
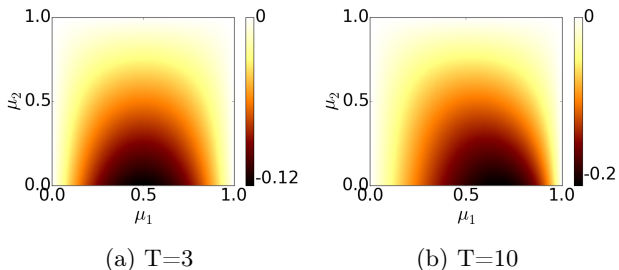
(a) T=3   (b) T=10

Figure 1: The bias of $X^{(1)}$ as a function of μ_1 and μ_2 running the Greedy algorithm, at $T = 3$ and $T = 10$ respectively. Note the difference in scale.

the event of a tie, i.e. both samples are 0 or 1, $X^{(1)}$ is selected by default. We can derive analytic expressions for the bias of the empirical mean of each distribution at $T = 3$.

$$\text{bias}_1 \equiv \mathbb{E}\left[\overline{X_3^{(1)}}\right] - \mu_1 = -\frac{1}{2}\mu_1(1 - \mu_1)\mu_2,$$

$$\text{bias}_2 \equiv \mathbb{E}\left[\overline{X_3^{(2)}}\right] - \mu_2 = -\frac{1}{2}\mu_2(1 - \mu_2)(1 - \mu_1).$$

When $0 < \mu_1, \mu_2 < 1$, both biases are strictly negative. This simple example already demonstrates the interesting phenomenon that *the distribution with the highest mean does not always have the least bias*. Using the above analytical forms, the ratio of the biases is $\frac{\text{bias}_1}{\text{bias}_2} = \frac{\mu_1}{1-\mu_2}$. Therefore bias_2 is worse than bias_1 when μ_1, μ_2 are both close to 1, and bias_1 is worse than bias_2 when μ_1, μ_2 are both close to 0. Figure 1 illustrates the quantitative bias of the first distribution $X^{(1)}$ at times $T = 3$ and $T = 10$. The setup and bias is symmetric for the second distribution $X^{(2)}$.

The insight from our proof of Theorem 1 is that the bias of distribution k at time t should be large if how likely we are to choose k in the future (after t) is sensitive to the value $\overline{X_t^{(k)}}$. This sensitivity increases if there is *consequential competition* for distribution k at time t, i.e. if there are other distribution(s), i, whose empirical average $\overline{X_t^{(i)}}$ is in some *consequential* middle range from the empirical average of distribution k. When they are too far apart, the particular sample values drawn from k are not consequential to the chance of it getting sampled again. If they are too close, having one bad sample value also does not affect the chance of k being drawn as much. We demonstrate the above remarks empirically in the next section.

Experiments quantifying negative bias. We explore the effects on the bias from moving the distribution means apart. We used the lil' UCB algorithm, with algorithm specific parameters $\alpha = 9, \beta = 1, \epsilon = 0.01, \delta = 0.005$, which are the same as in the experiment section of Jamieson et al. [2014]. We ran 1000 independent trials, with horizon $T = 500$. We have

three unknown distributions, all of the form $\mathcal{N}(\mu_i, 1)$, with $\mu_1 = 2, \mu_2 = 1.5, \mu_3 = 1$. In this experiment, we scale the μ_i's by a scaling factor of $1, 2, 3$, and observe the bias of the empirical mean estimates of the three distributions. In Figure 2(a) (b) (c), we plot the bias with the number of rounds.

We first observe all distributions have negatively biased estimates of their true means. Further, these plots illustrate our intuition on the effect of *consequential competition*. The distribution with the second best mean (the magenta curve) has worse bias as we scale up the μ_i's. We hypothesize when the distributions with the second best and the best means are close together, having one bad sample value for the second best one does not affect its chance of being sampled again as much as when their means are farther apart. On the other hand, for the distribution with the lowest true mean (the yellow curve), we observe its bias becomes worse first and then better as we scale up the μ_i's. We hypothesize that the same reason as before explains why the bias becomes worse first. However, as we further scale up the μ_i's, the bad sample values from the distribution with the lowest mean does not affect its future chances of being drawn much more than the good samples values, since its true mean is too far from the distribution with the highest mean.

Next we compare lil' UCB with Greedy, see sub-figure (a) and (d) in Figure 2. We observe that with Greedy in our setting, the empirical mean estimates for the distribution with the lowest mean has the least bias, followed by the distribution with the highest true mean. This is an example in which the distribution with the highest mean might not incur the least bias. With lil' UCB, the bias for the distribution with the highest true mean converges to 0 quickly, but with Greedy it plateaus. In lil' UCB, since it achieves optimal regret, the algorithm finds the distribution the highest true mean in finite number of time steps. The samples we get from that distribution become close to i.i.d. samples as t increases, since the effect of the competition from other distributions is reduced over time. In Greedy it's known that the algorithm can be stuck on drawing from a suboptimal distribution, in which case the empirical average of the particular samples we have drawn from the distribution with the highest true mean must have a negative bias for this to happen. The bias of the best distribution thus doesn't converge to 0. Note that for both lil' UCB and Greedy, the suboptimal distributions can stay negatively biased for large T. However, the negative bias from running lil' UCB is less severe in magnitude, because it uses confidence intervals to better address the issue that the empirical mean may be artificially small purely due to randomness.

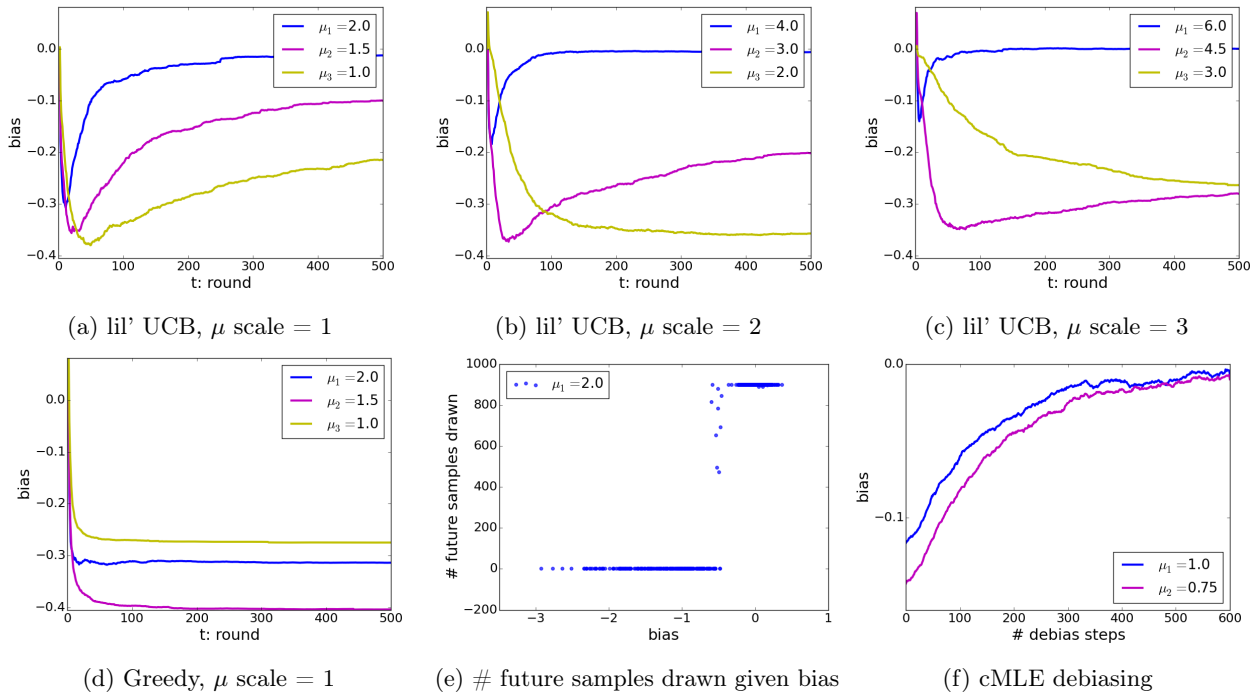Figure 2(e) shows at round step $t = 100$ with horizon

Xinkun Nie, Xiaoying Tian, Jonathan Taylor, James Zou

(a) lil' UCB, μ scale = 1

(b) lil' UCB, μ scale = 2

(c) lil' UCB, μ scale = 3

(d) Greedy, μ scale = 1

(e) # future samples drawn given bias at t=100, with horizon T=1000

(f) cMLE debiasing

Figure 2: In (a-c), we plot the bias of the empirical mean estimates of three unknown distributions running lil' UCB with horizon T=500. Each is distributed according to $\mathcal{N}(\mu_i, 1)$, where μ_i is the mean of the i-th distribution, specified in the legends of the plot. We see that as we scale up μ_i's, so they become more spread out, the bias increases/decreases depending how far the μ_i's are from each other, and what is the order of the distributions. (d) plots the bias of the three unknown distributions running Greedy. (e) plots the number of future samples drawn from distribution 1 given its bias at t=100, running lil' UCB. Here T=1000 with two distributions, $\mathcal{N}(2, 1)$ and $\mathcal{N}(1.5, 1)$. This is a scatter plot over 1000 independent trials. (f) plots the bias as the estimate of the mean converges to the true mean across 600 gradient descent iterations. (a-d) and (f) are all averaged across 1000 independent trials.

$T = 1000$, running lil' UCB with the same hyper-parameters in the same setting as in Figure 2(a), we plot the number of future samples drawn from the distribution with the highest mean (i.e. $\mu = 2.0$) vs. the bias from the empirical average of samples drawn so far from this distribution at time $t = 100$. This confirms our intuition that large negative bias is correlated with fewer future chances of getting sampled.

Our theoretical analysis of Theorem 1 shows that the *marginal* bias of each distribution is negative. The *joint* bias across all the distributions—e.g. how likely is it that all the distributions have negative bias—is also an interesting question. We empirically investigated the frequency at which negative bias occurs across distributions in simulations. In Table 1, we run Greedy, lil' UCB, ϵ-Greedy ($\epsilon = 0.1$), and Thompson Sampling (shown as "TS" in Table 1) on 5 different distributions (see caption of Table 1 for details) for 10,000 trials, and record the fraction of trials at which any m distributions all have negative bias at $T = 100$, where $m = 0, \cdots, 5$. We observe that the results are highly skewed towards large values of m, suggesting that it is much more

Table 1: We run each of the following common bandit algorithms (note "TS" stands for Thompson Sampling) across 10,000 independent trials with 5 distributions, each with $\mathcal{N}(\mu_i, 1)$, where $\mu_1 = 1.0, \mu_2 = 0.75, \mu_3 = 0.5, \mu_4 = 0.38, \mu_5 = 0.25$. In each column, we record the fraction of trials in which m distributions have negative bias at $T = 100$, where $m = 0, \cdots, 5$. In ϵ-Greedy, $\epsilon = 0.1$. In Thompson Sampling (TS), all distributions have prior $\mathcal{N}(0, 25)$.

	m: # of distr. with bias < 0					
	0	1	2	3	4	5
Greedy	0.02	0.09	0.23	0.34	0.24	0.08
lil' UCB	0.01	0.05	0.21	0.36	0.30	0.08
ϵ-Greedy	0.02	0.12	0.27	0.33	0.21	0.05
TS	0.01	0.08	0.24	0.34	0.26	0.07

frequent that more distributions simultaneously have negative bias. We have also included additional results in a variety of settings in Appendix C that confirm this finding. Theoretical analysis of the joint bias across distributions is beyond the scope of this paper and is an interesting direction of future investigation.

4 DEBIASING ALGORITHMS AND EXPERIMENTS

Data splitting. A simple approach to obtain unbiased estimators of μ_k's is to split the data. Data splitting dates back to Cox [1975] and has been discussed by in the context of identifying loci of interest in genetics [Sladek et al., 2007], and online search advertising [Xu et al., 2013]. Wasserman and Roeder [2009] and Meinshausen et al. [2009] discussed data splitting in high-dimensional inference. Fithian et al. [2014] discussed data splitting in post-selective inference. Let k be the distribution the selection function f chooses at time t. Instead of taking one sample from k, we maintain a "held-out" set by taking an additional independent sample from k. We use the first samples as the sample history for f which determines the future selections, and use the "held-out" set composed of the second samples for mean estimation. Since the "held-out" set is composed of i.i.d. samples that are independent of the selection process, its sample average is an unbiased estimate of μ_k. However, if the total number of samples collected is fixed at T rounds, then data splitting suffers from high variance, since half of all the samples are discarded in estimation. Data splitting is a natural baseline and we compare it with more sophisticated debiasing algorithms.

Conditional Maximum Likelihood Estimator (cMLE). Data splitting is a general approach since it is agnostic to the selection function f. If we know the f used to collect the data, then more powerful debiasing could be achieved by explicitly conditioning on the sequence of distributions that are selected by f in a maximum likelihood framework. This conditioning approach is motivated by the recent successes of selective inference, which have been applied, for example, to debias the confidence intervals of the Lasso-selected features by conditioning on the Lasso algorithm [Taylor and Tibshirani, 2015]. To the best of our knowledge, our paper is the first extension of selective inference to adaptive data collection. To illustrate the cMLE approach, we consider the special case where the decision on which distribution to sample at round t is based on comparing the decision statistics of the form,

$$\mathbf{U}_t \triangleq \left(U\left(\overline{X_t^{(1)}}, N_t^{(1)} \right), \ldots, U\left(\overline{X_t^{(K)}}, N_t^{(K)} \right) \right). \quad (1)$$

$\mathbf{U}_t$ depends only on the empirical average $\overline{X_t^{(k)}}$'s and the number of samples $N_t^{(k)}$'s for $k \in [K]$. In other words, the selection function f depends on the history of rewards Λ_t only through $\mathbf{U}_t$. In Greedy, $U\left(\overline{X_t^{(k)}}, N_t^{(k)} \right) = \overline{X_t^{(k)}}$, while in UCB type algorithms, $U_t^{(k)}$ will be the upper confidence bounds that depend

on both $\overline{X_t^{(k)}}$'s and $N_t^{(k)}$'s, where $U_t^{(k)}$ is shorthand for $U\left(\overline{X_t^{(k)}}, N_t^{(k)} \right)$.

Theorem 2. *Let $s_t = f(\Lambda_t)$. Suppose the distributional function for distribution k has density $h_{\theta^{(k)}}$, then the conditional likelihood of the adaptive data collection problem is proportional to*

$$p(\Lambda_T \mid s_t, \ t = 1, \ldots, T) \propto \prod_{k=1}^{K} \prod_{m=1}^{N_T^{(k)}} h_{\theta^{(k)}}(X_m^{(k)})$$

$$\cdot \prod_{t=K}^{T-1} \Pr\left[f(\mathbf{U}_t) = s_{t+1} \mid \mathbf{U}_t \right]. \quad (2)$$

To maximize the conditional likelihood, we need to solve the following optimization problem,

$$\max_{\theta} \ \sum_{k=1}^{K} \sum_{m=1}^{N_T^{(k)}} \log\left[h_{\theta^{(k)}}(X_m^{(k)}) \right]$$

$$+ \sum_{t=K}^{T-1} \log\left[\Pr\left[f(\mathbf{U}_t) = s_{t+1} \mid \mathbf{U}_t \right] \right] \ - \log Z(\theta), \quad (3)$$

where $\theta = (\theta^{(1)}, \ldots, \theta^{(K)})$ are the parameters of interest and $Z(\theta)$ is the partition function in Eqn. (2), that only depends on the parameters θ.

Theorem 2 gives an explicit form for the likelihood function of the adaptive data collection problem (up to a constant). Note that in Eqn 2, Λ_T contains both $X_m^{(k)}$, and $\{\mathbf{U}_t\}_{t=1}^{T}$ (recall that Λ_T is the history of samples up to time T). We give a proof of Theorem 2 in Appendix B.

Adding additional noise to the sample values to improve cMLE optimization. We introduce additional randomization when selecting a distribution. The reasons are two-fold. First, we need exponential-tailed noise in randomization to achieve asymptotically consistent estimates [Tian and Taylor, 2015, Panigrahi et al., 2016]. Second, adding randomization smooths out the the hard boundaries in the sample space in evaluating $\Pr\left[f(\mathbf{U}_t) = s_{t+1} \mid \mathbf{U}_t \right]$. For example, in Greedy,

$$\Pr\left[f(\mathbf{U}_t) = s_{t+1} \mid \mathbf{U}_t \right] = \mathbb{I}\left(\arg\max_k \overline{X_t^{(k)}} = s_{t+1} \right), \quad (4)$$

which means to compute the cMLE, we need to maximize the log-likelihood in a constrained region of the sample space. Optimization on such a region is no easy task. Moreover, since the hard-max function induces discontinuities of likelihood along the boundary of the constrained region, the cMLE will be ill-behaved since the gradient of the log-likelihood can become infinite [Tian and Taylor, 2015, Panigrahi et al., 2016].

We propose adding Gumbel noise to the decision statistics $\mathbf{U}_t$ to smooth out $\Pr[f(\mathbf{U}_t) = s_{t+1} \mid \mathbf{U}_t]$. Note that we could also use other heavy-tailed distributions for the added noise. The Gumbel distribution offers computational convenience, since

$$\Pr[f(\mathbf{U}_t) = k \mid \mathbf{U}_t] = \frac{\exp[U_t^{(k)}/\tau_t]}{\sum_{i=1}^K \exp[U_t^{(i)}/\tau_t]} \qquad (5)$$

has a closed form due to the Gumbel-max trick [Gumbel and Lieblein, 1954] (also see Lemma 1 in Appendix F).

We can now optimize Eqn. 3 using contrastive divergence [Carreira-Perpinan and Hinton, 2005]. Details of the algorithm are in Appendix E. For lil' UCB or Greedy, we can compute $\mathbf{U}_t$ deterministically from $\overline{\mathbf{X}_t}$ and $\mathbf{N}_t$. The selection function after Gumbel randomization is defined as

$$f(\mathbf{U}_t) = \arg\max_k U_t^{(k)} + \epsilon_t^{(k)}, \quad \epsilon_t^{(k)} \overset{\text{iid}}{\sim} G_{\tau_t}, \qquad (6)$$

where G_τ is a Gumbel distribution of mean 0 and scale parameter τ. Similarly, we can also add Gumbel noise to ϵ-Greedy to derive smooth conditional probabilities. We give examples of computing the conditional likelihood functions of common bandit algorithms with added Gumbel noise in Appendix D.

We summarize the debiasing procedure in Algorithm 1. Note that we only compute cMLE with contrastive divergence (see Algorithm 2 in Appendix E) once at time step T when we wish to debias the estimates. Note that with these smooth $\Pr[f(\mathbf{U}_t) = k \mid \mathbf{U}_t]$ in Eqn. 5, we have well-behaved gradients in the parameter updates in computing cMLE.

Debiasing experiments. We empirically show that the cMLE algorithm can reduce bias significantly and reduce the mean squared error (MSE) as well. In Table 2, we see significant bias reduction for the lil' UCB, ϵ-Greedy, and Greedy algorithms using the cMLE debiasing algorithm, in both the $K = 2$ and $K = 5$ cases, where K is the number of distributions. All of our experiments used the same implementation of cMLE with the same hyper-parameters to demonstrate that cMLE can be robustly applied to different distribution settings without fine-tuning. We could still have some residual bias after running cMLE. This is due to the guarantee of asymptotic consistency with added heavy-tailed noise (i.e. the bias tends to 0 as T tends to infinity), and T is finite in our experiments. Table 3 shows the reduction of MSE. The data splitting algorithm achieves consistent estimates, but it incurs high variance since the effective sample size is halved by maintaining a held-out set. Empirically we observe that data splitting suffers from high MSE. In Figure 2(f), we run Greedy with cMLE, with two distributions, $\mathcal{N}(1, 1)$ and $\mathcal{N}(0.75, 1)$.

Algorithm 1 Debiasing algorithm using cMLE. Note that we only compute cMLE with contrastive divergence (see Algorithm 2 in Appendix E) once at time step T when we wish to debias the estimates.

Add Gumbel noise when choosing which distribution to sample from at each time step t. Instead of applying the selection function directly to $\mathbf{U}_t$, we apply it to

$$\left(U_t^{(k)} + \epsilon_t^{(k)}\right), \quad k = 1, \dots, K$$

where $\epsilon_t^{(k)} \overset{\text{iid}}{\sim} G_{\tau_t}$.

Compute conditional likelihood by computing the selection probabilities,

$$\Pr_{\epsilon_t}[f(\mathbf{U}_t) = s_{t+1} \mid \mathbf{U}_t].$$

Note that here f incorporates the randomness of Gumbel randomizations $\{\epsilon_t^{(k)}\}_{k \in [K]}$ as well as the randomness in the original bandit algorithm.

Compute cMLE using approximate gradient descent with contrastive divergence.

We show the convergence of the estimated mean to the true mean as we run gradient descent over 600 iterations. We see that cMLE significantly reduces the bias, while improving the MSE. We also experimented with propensity matching, a commonly used method that weights each observed value of a distribution by one over the probability that this distribution is selected [Austin, 2011]. Propensity matching is unbiased, but has very large variance and thus a much greater MSE by several fold compared to cMLE. We discuss it in more detail in Appendix G.

In both Table 2 and Table 3, we looked at the cases where the horizon T is relatively small. The reasons are two-fold. First, a relatively small T is relevant in many biomedical settings, where the scientist adaptively collects data from several arms corresponding to different experimental conditions, and such collection procedures are expensive. Second, empirically, the magnitude of the bias tends to be the largest when t is small (c.f.Figure 2 (a)-(d)), so we focused on cases where the bias is the largest to demonstrate the effectiveness of cMLE to debias the empirical estimates. For ϵ-Greedy, for example, the biases of all the distributions are essentially 0 when T is large, and cMLE is not needed in this regime. For the Greedy algorithm, the bias of the suboptimal distributions are stuck at quite negative values even for very large T because they are not sampled again after the first few rounds. For completeness, we have also included the debiasing results for the Greedy algorithm when the horizon is large ($T = 1000$) in Table 4. We observe here that cMLE

Table 2: **Bias Reduction.** With $K = 2$, each distribution is drawn from $\mathcal{N}(\mu_i, 1)$. where $\mu_1 = 1.0, \mu_2 = 0.75$. With $K = 5$, each distribution is drawn from $\mathcal{N}(\mu_i, 1)$. where $\mu_1 = 1.0, \mu_2 = 0.75, \mu_3 = 0.5, \mu_4 = 0.38, \mu_5 = 0.25$. In the left columns under each algorithm, we record the bias of the original algorithm at different time steps T. In the right columns, we record the percentage of the original bias that still remains after we run cMLE by adding gumbel noise $\epsilon_g \sim G_\tau$, with scale parameter $\tau = 1.0$, and contrastive divergence with 600 gradient descent iterations. All results are averaged across 1000 independent trials.

	lil' UCB		ϵ-Greedy ($\epsilon = 0.1$)		Greedy	
	orig.	cMLE	orig.	cMLE	orig.	cMLE
T=8,K=2	-0.26	**6.2%**	-0.25	**7.3%**	-0.29	**2.8%**
T=16,K=2	-0.29	**5.2%**	-0.25	**1.6%**	-0.32	**8.3%**
T=20,K=5	-0.32	**14.9%**	-0.31	**9.1%**	-0.35	**18.0%**
T=40,K=5	-0.35	**14.2%**	-0.27	**8.8%**	-0.37	**15.9%**

Table 3: **Mean Squared Error(MSE) reduction.** The experiment setup is the same as in Table 2. The leftmost columns under each algorithm is the MSE of the original algorithm. The second to the left columns are the MSE percentage ratio of the data splitting with a held-out set compared to the MSE of the original algorithm. The right columns are the MSE percentage ratio of the cMLE algorithm after debiasing compared to the MSE of the original algorithm. For ϵ-Greedy, we additionally run propensity matching (prop). Note that both data splitting and prop suffer from high variance despite achieving consistent estimation.

	lil' UCB			$\epsilon-$Greedy$(\epsilon = 0.1)$				Greedy		
	orig.	held	cMLE	orig.	held	prop	cMLE	orig	held	cMLE
T=8,K=2	0.56	108%	**86%**	0.51	123%	295%	**76%**	0.56	108%	**78%**
T=16,K=2	0.50	101%	**40%**	0.38	123%	244%	**52%**	0.53	107%	**45%**
T=20,K=5	0.57	112%	**99%**	0.52	123%	399%	**94%**	0.59	111%	**89%**
T=40,K=5	0.54	104%	**52%**	0.39	135%	290%	**62%**	0.54	107%	**52%**

can almost completely debias the Greedy algorithm. This is expected since cMLE is asymptotically consistent, so as T grows, the bias reduces to 0 significantly. The mean squared error (MSE) has also reduced to a negligible amount for cMLE, but remains huge for the data splitting method. We have also included results running Thompson Sampling in Appendix H for completeness.

Table 4: The bias and mean squared error (MSE) of running the Greedy algorithm with $T = 1000$. The experiment setup is the same as in Table 2, for $K = 2$ and $K = 5$, where K is the number of distributions of interest. The columns under "Bias" record the bias incurred by the original algorithm under "orig.", and the percentage of bias remaining after running the cMLE algorithm under "cMLE". The columns under "MSE" record the mean squared error (MSE) under the original greedy algorithm, the percentage of MSE running the data splitting method (i.e. using a held-out set of samples) in comparison to the original MSE, and the percentage of MSE running cMLE in comparison to the original MSE. We observe the cMLE has superior performance in both bias and MSE reduction.

	Bias		MSE		
	orig.	cMLE	orig.	held-out	cMLE
K=2	-0.2	**0.0%**	0.255	89.8%	**0.4%**
K=5	-0.21	**1.0%**	0.277	94.9%	**1.1%**

5   DISCUSSION

Our main result shows that adaptively collected data are negatively biased when the data collection algorithm f satisfies *Exploit* and *IIO*. This seems counterintuitive because we typically associate optimization (as in exploitative algorithms) with a positive selection bias (i.e.Winner's Curse). For example, if we draw 10 samples from $\mathcal{N}(0, 1)$ and report the max, then we have positive reporting bias. The reason for the discrepancy between these phenomena is that for any sample history of data, the "best" option k's sample mean is likely to be larger than its true mean. However who is the "best" varies in different sample paths, and the bias of each distribution k is negative in expectation.

We explored data splitting and cMLE as two approaches to reduce this bias. Data splitting is unbiased but suffers larger MSE because it ignores half of the samples during estimation. cMLE can reduce bias close to 0 while also reducing MSE. The trade-off is that it requires specific knowledge about f and also requires one to add additional noise to the collected data. Both approaches require modifying the data collection procedure and cannot be generically applied to debias existing adaptively collected data. As adaptively collected data are ubiquitous, developing flexible debiasing approaches to debias observational data is an important direction of future research.

References

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46 (3):399–424, 2011.

M. A. Carreira-Perpinan and G. E. Hinton. On contrastive divergence learning. In *AISTATS*, volume 10, pages 33–40. Citeseer, 2005.

D. Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.

A. Deng, J. Lu, and S. Chen. Continuous monitoring of a/b tests without pain: Optional stopping in bayesian testing. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, pages 243–252. IEEE, 2016.

C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 117–126. ACM, 2015.

E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.

W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

E. J. Gumbel and J. Lieblein. Statistical theory of extreme values and some practical applications: a series of lectures. 1954.

X. T. Harris, S. Panigrahi, J. Markovic, N. Bi, and J. Taylor. Selective sampling after solving a convex problem. *arXiv preprint arXiv:1609.05609*, 2016.

I. Ionita-Laza, A. J. Rogers, C. Lange, B. A. Raby, and C. Lee. Genetic association analysis of copy-number variation (cnv) in human disease pathogenesis. *Genomics*, 93(1):22–26, 2009.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439, 2014.

J. Lu and A. Deng. Demystifying the bias from selective inference: A revisit to dawid's treatment selection problem. *Statistics & Probability Letters*, 118:8–15, 2016.

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

S. Panigrahi, J. Taylor, and A. Weinstein. Bayesian post-selection inference in the linear model. *arXiv preprint arXiv:1605.08824*, 2016.

D. Russo and J. Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2016.

R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130): 881–885, 2007.

J. Taylor and R. J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

X. Tian and J. E. Taylor. Selective inference with a randomized response. *To Appear in the Annals of Statistics*, 2015.

X. Tian, N. Bi, and J. Taylor. Magic: a general, powerful and tractable method for selective inference. *arXiv preprint arXiv:1607.02630*, 2016.

L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

M. Xu, T. Qin, and T.-Y. Liu. Estimation bias in multi-armed bandit algorithms for search advertising. In *Advances in Neural Information Processing Systems*, pages 2400–2408, 2013.